Atmospheres and Oceans on Computers: Fundamentals

Lars Petter Røed†,⋆,‡

†Research and Development Department, Norwegian Meteorological Institute, Postboks 43 Blindern, 0313 Oslo, Norge.
⋆Department of Geosciences, Section Meteorology and Oceanography, University of Oslo, P. O. Box 1022 Blindern, 0315 Oslo, Norway
‡E-mail: larspetter.roed@met.no

Date printed: December 14, 2011
PREFACE

To me numerical methods is one of the most fascinating contemporary tools to solve meteorological and oceanographic problems. Most of the processes in the atmosphere and ocean are highly non-linear, and hence the equations that govern their motions are highly non-linear as well. Only in special cases, and mostly for simplified processes, is it possible to solve the governing equations by analytic means. Hence to solve meteorological and oceanographic problems to their full extent numerical methods are in fact the only tool at our disposal.

It is therefore not surprising that atmospheric scientists and weather forecasters embraced this tool early on. In fact when the first digital computers were developed in the 1940s a whole new science field developed in meteorology now commonly referred to as numerical weather prediction, or simply NWP. Among the first to embrace the new tool were Jules Charney, Ragnar Fjørtoft, and John von Neumann in the late 1940’s. In 1950 (Charney et al., 1950) they published results on the very first attempt of producing a weather forecast using a numerical atmospheric model. Since then NWP has been one of the major science fields to continuously push the computer technology to its very limits.

Another important basis for the rapid development of NWP in the 1950s and 1960s was the deterministic paradigm already stated by Vilhelm Bjerknes in his famous 1904 paper (Bjerknes, 1904) entitled “Das Problem der Wettervorhersage, betrachtet vom Standpunkte der Mechanik und der Physik”. Here he states:

“If it is true, what most scientific persons think, that the atmospheric state at any time can be developed from its earlier state using physical laws, then it follows that the necessary and sufficient condition for a rational solution to the problem of weather forecasting is:

1. A sufficiently accurate knowledge of the present atmospheric state, and
2. A sufficiently accurate knowledge of the equations that govern the development of the atmosphere from one state to the next.”

Although he didn’t acknowledged it at the time the same statement is true regarding forecasts of oceanic “weather”. It i therefore not surprising that as the power and capacity of computers grew it became common also to use computers and numerical methods to solve oceanographic problems. Indeed, the capacity and power of today’s computers are now amenable for oceanographic forecasting purposes on the mesoscale, that is, to forecast eddies, meanders and jets in the ocean, that is, the oceanic weather. Thus since the 1990s a whole new science field has emerged referred to as numerical ocean weather prediction (NOWP).

Moreover, I believe that in the near future the first coupled atmosphere-ocean forecasting models will appear. In fact, the coupling of the model modules and the coupling techniques necessary to do the job are already to a certain extent developed within the climate modeling community. However, much work is still required to adapt these tools for coupled NWP/NOWP systems.

---

1Vilhelm Friman Koren Bjerknes (14 March 1862 - 9 April 1951) was a Norwegian physicist and meteorologist who did much to found the modern practice of weather forecasting.

2The growth in computer power and capacity has been almost exponential since the 1940s.
I therefore strongly believe that anyone aspiring to become a meteorologist, an oceanographer or a climatologist must have a solid knowledge and insight into the fundamental methods used to develop sound numerical oceanographic and meteorological models. Furthermore, I believe that to acquire this knowledge and insight it is necessary to get some “hands-on-experience”. It is with this in mind these lecture notes and the accompanying Computer problems are developed. Consequently, the objective is to give the reader a basis for evaluating the soundness of the various numerical methods implemented in a particular numerical atmosphere and/or ocean model, whether it is a simple model constructed to study a particular process or a full blown NWP and/or NOWP model.

It is a growing concern though that more and more of the research in atmosphere, ocean and climate related fields relies on sometimes large and monstrous computer codes, and that as a rule many of these codes, commonly written in the programming language FORTRAN, are written and amended by scientists who are not necessarily skilled programmers. This concern is further corroborated by a recent statistical survey published by Hannay et al. (2009) which concludes that “the knowledge required to develop and use scientific software is primarily acquired from peers and through self-study, rather than from formal education and training”. The codes may therefore, even though the numerical methods employed are sound, be rather poorly written from a skilled programmers point of view. Only rarely do these codes undergo rigorous testing. Hence the model codes may inadvertently contain errors that may potentially be damaging to the results. The results produced may look reasonable, but in reality they may be totally false, and possibly lead to wrong conclusions. In fact there exists examples in the literature of numerical solutions that are interpreted as a new physical phenomenon that later was shown to be pure artifacts produced by employing an incorrect numerical method.

Mathematically speaking the governing equations we use to solve atmospheric and oceanic problems, and to make forecasts, is a set of partial differential equations. When solving these by numerical means there is a variety of numerical methods available to choose from. It is therefore important to understand why some methods work and some don’t for a specific problem, which is what this Lecture notes is all about. Although these notes are directed towards graduate students in meteorology and oceanography it is my hope that scientists who would like to revisit the fundamental numerical methods used to solve geophysical fluid dynamic problems will find them useful as well.

We should keep in mind that most of the numerical methods used to solve atmospheric problems also works for solving oceanographic problems, and that they were originally developed within the atmospheric community. The reason is that the dynamics of atmosphere and oceans are similar in many respects. It is therefore no need to treat numerical meteorology and oceanography differently. In particular this is true regarding the more fundamental methods. A further rationale is the fact, as already mentioned, that the atmosphere and the ocean is inherently a coupled system. To fully appreciate and understand the coupled system and its modeling one therefore needs to have knowledge of meteorology as well as oceanography, a fact perhaps most evident within climate modeling today.

Though the objective of these notes is to give insight into the fundamental numerical methods to solve problems common to atmosphere and ocean, there are indeed advanced numerical methods and techniques that are unique to each field simply because there are differences between the
two spheres. This is particularly true regarding their thermodynamics. Thus additional courses must be offered to students for whom application of numerical models is essential in their work.

Finally, it is worth emphasizing that the application of numerical methods to solve atmospheric and oceanographic problems is to a large degree a “hands on” exercise. It is of no use to learn the theory behind the numerical methods employed without knowing how to implement them and run them on the computer, not to mention how to visualize them. Thus the reader is not only encouraged to solve the computer problems contained in an accompanying Computer problems, but also to solve the exercises at the end of each chapter.
Acknowledgements

The content below is based on lecture notes I have compiled teaching numerical methods to solve oceanographic and atmospheric problems over the last 23 years or so. I would like to extend my greatest appreciation and sincere thanks to Dr. James J. O’Brien, former director of the Center for Atmosphere-Ocean Predictions (COAPS), and now distinguished professor emeritus at the Florida State University, USA for introducing me to numerical methods to solve oceanographic problems. Some of the material covered is in fact based on notes taken when I followed his lectures in numerical methods to solve oceanographic problems at Florida State University back in 1980/81.

I would also like to thank Dr. Thor Erik Nordeng, Norwegian Meteorological Institute and Adjunct Professor of Meteorology at the University of Oslo for helping me compiling material from his lecture notes on atmospheres on computers. In this respect I also extend my gratitude to Dr. Arne Bratseth former Professor at the University of Oslo now deceased. Some of the material covered in his lecture notes on “Numerical Atmosphere Models” is also used. Finally, my appreciation also goes to the many students who has pointed out misprints and errors in earlier versions of my lecture notes over the years.

This is the 4th version of the lecture notes written for the course GEF4510 at the Meteorology and Oceanography Section, Department of Geosciences, University of Oslo. This particular version is based on the fall 2011 lectures.

Blindern December 5, 2011
Lars Petter Røed (sign.)
Revision history (most recent on top):

All revisions made by the author.

☞ Dec 05, 2011: Spell checked all chapters, and made the final fall 2011 version.

☞ Nov 14, 2011: Moved chapter on “Open Boundary Condition” (7) before Chapter 8.

☞ Nov 08, 2011: Added Section 10.4 in Chapter 10 to include smoothing and filtering techniques. Completed Chapter 6. Note that there may still be misprints present.

☞ Oct 19, 2011: Moved the section on two-dimensional problems from the chapter “Advanced topics” to a chapter in its own right (cf. new Chapter 9). In addition it is reworked with added text as well as numerous changes in comparison with the previous version.

☞ Oct 12, 2011: Corrected equation (5.39) on page 66 and added some explanation on how to arrive at this equation.

☞ Oct 07, 2011: Added some material at the beginning of Chapter 5.

☞ Sep 13, 2011: Misprint in (2.74) in Exercise 3 on page 27 corrected so that the fourth term in the nominator now reads \(4\theta_{j-1}\). Also changed the text of the same Exercise somewhat.

☞ Aug 19, 2011: Moved the section on energy conversions, originally part of Chapter 6, to Chapter 3 and given it the heading “Shallow water equations” (cf. the new Section 3.4 of Chapter 3). Also made a few changes in the wording of the section, and dropped all external forcing terms in the development.

☞ Aug 18, 2011: Changed some of the wording in the introductory Chapter 1, hopefully making it more understandable. Moreover a few misprints in Chapter 1 was detected and corrected.

☞ Jul 07, 2011: This version is based on the fall 2010 Lecture Notes. At this stage the preface is rewritten. No material is added or erased and no major rewriting at this stage.
Contents

Preface ................................................................. iii
Acknowledgements .................................................. vii
Revision history ..................................................... ix

1 INTRODUCTION .................................................. 1
  1.1 The governing equations ..................................... 1
  1.2 Boundary and initial conditions .............................. 3
  1.3 The hydrostatic approximation ............................... 4
  1.4 The Boussinesq approximation ............................... 5
  1.5 The shallow water equations ................................. 7
  1.6 The quasi-geostrophic equations ............................ 9

2 PRELIMINARIES ................................................ 13
  2.1 General PDEs .................................................. 13
  2.2 Elliptic equations ............................................. 14
  2.3 Parabolic equations ........................................... 14
  2.4 Hyperbolic equations ......................................... 15
  2.5 Boundary conditions .......................................... 16
  2.6 Taylor series and expansions ............................... 18
  2.7 Finite difference approximations ......................... 19
  2.8 Truncation errors ............................................. 21
  2.9 Notations ..................................................... 22
  2.10 Orthogonal functions ........................................ 24
  2.11 Fourier series ................................................. 26
  2.12 Fourier transforms .......................................... 27

3 TIME MARCHING PROBLEMS ................................. 29
  3.1 Advection-diffusion .......................................... 30
  3.2 Diffusion ..................................................... 31
  3.3 Advection ..................................................... 33
  3.4 Shallow water equations .................................... 34
4 THE DIFFUSION PROBLEM
4.1 Finite difference form ........................................ 38
4.2 Numerical stability ........................................... 41
4.3 Stability analysis ............................................. 42
4.4 Necessary stability ............................................ 46
4.5 Explicit and implicit schemes ................................. 46
4.6 DuFort-Frankel .................................................. 48
4.7 Crank-Nicholson ............................................... 51
4.8 A direct elliptic solver ....................................... 53

5 THE ADVECTION PROBLEM
5.1 Finite difference form .......................................... 61
5.2 The leapfrog scheme ........................................... 63
5.3 The CFL condition ............................................. 64
5.4 The upstream scheme .......................................... 65
5.5 The diffusive scheme .......................................... 66
5.6 The Lax-Wendroff scheme ................................... 67
5.7 An implicit scheme ............................................ 69
5.8 Solutions to the initial problem of CTCS schemes ......... 70
5.9 Method of characteristics ................................... 71
5.10 Physical interpretation ....................................... 74
5.11 Numerical diffusion .......................................... 75
5.12 Flux correction ................................................ 77
5.13 Numerical dispersion ........................................ 79
5.14 Unphysical solutions ........................................ 82
5.15 The Asselin filter ............................................. 85

6 THE SHALLOW WATER PROBLEM
6.1 Linearization .................................................... 92
6.2 Linear, non-rotating .......................................... 94
6.3 Staggered grids ................................................. 97
6.4 Linear, rotating ................................................ 100
6.5 Non-linear dynamics ......................................... 104
6.6 The semi-implicit method ................................... 106
6.7 Semi-Lagrangian method .................................... 107

7 OPEN BOUNDARY CONDITIONS
7.1 Definition ...................................................... 112
7.2 Radiation conditions .......................................... 113
7.3 Implementation ............................................... 116
7.4 The sponge ..................................................... 119
7.5 Weakly reflective ............................................. 120
7.6 Flow relaxation ................................................ 121
# List of Figures

1.1 The equation of state for the ocean. .................................................. 3

2.1 Displayed is a commonly used grid when employing numerical methods to solve PDEs. The points in the \( x, y \) directions are incremented by \( \Delta x, \Delta y \), respectively, so that there are a total of \( J \) points along the \( x \)-axis and \( K \) points along the \( y \)-axis. The points are counted by using the dummy counters \( j, k \). ........................................... 23

4.1 Displayed is the employed grid we use to solve (4.1) by numerical means. The grid points in the \( x, t \) directions are incremented by \( \Delta x, \Delta t \), respectively. There is a total of \( J + 1 \) points along the \( x \)-axis and \( N + 1 \) points along the \( t \)-axis, counted by using the dummy indices \( j, n \). The coordinates of the grid points are \( x_j = (j - 1)\Delta x \) and \( t^n = n\Delta t \), respectively. .................................................. 39

4.2 Displayed are solutions of the diffusion equation using the scheme (4.4) for respectively \( K = \kappa\Delta t/\Delta x^2 = 0.45 \) (left panel) and \( K = 0.55 \) (right panel) for \( x \in (0, 1) \). The dependent variable \( \theta \) is held fixed at the two boundaries \( x = 0, 1 \) and the initial condition is \( \theta = \sin \pi x \). The solutions are shown for the time levels \( n = 0, n = 50 \) and \( n = 90 \). Note the saw tooth like pattern in the right panel for \( n = 90 \) not present in the left panel. This indicates that the stability condition (4.34) is violated for \( K = 0.55 \), but not for \( K = 0.45 \). ........................................... 45

5.1 Displayed is the grid layout for the Lax-Wendroff scheme. The solid lines denote the grid through the circled \( x_j, t^n \) points. The dashed lines denote the grid through the \( x_{j+\frac{1}{2}}, t^{n+\frac{1}{2}} \)-points which are marked with a cross. ........................................... 67

5.2 Comparison of the numerical solution to the advection equation (5.1) using the leapfrog, the upwind and the Lax-Wendroff schemes with a Courant number of one half (\( C = 0.5 \)). The solution is shown after 10 cycles (periodic boundary condition). The true solution is a bell function as shown by the black solid curve. We note that both the leapfrog and the Lax-Wendroff schemes give rise to numerical dispersion, while the upwind scheme gives rise to numerical diffusion. 70

5.3 Sketch of the characteristics in the \( x, t \) plane. For \( u = u_0 = \text{constant} > 0 \) the characteristics are the straight lines sloping to the the right in \( x, t \) space as given by (5.55). If \( x = L \) marks the end of the computational domain, then all information about the initial condition is lost for times \( t > t_c \). ........................................... 72
5.4 Sketch of the method of characteristics. The distance between the grid points are \( \Delta t \) in the vertical and \( \Delta x \) in the horizontal direction. The sloping solid line is the characteristic through the point \( j, n + 1 \). It is derived from (5.55) and the slope is given by \( 1/u (u > 0) \). The point labeled \( Q \) is therefore a distance \( u \Delta t \) to the left of \( x_j \). As long as \( u \Delta t < \Delta x \) then \( Q \) is located between \( x_{j-1} \) and \( x_j \). If however \( u \Delta t > \Delta x \) then the point \( Q \) is located to the left of \( x_{j-1} \).

5.5 Displayed is an example of the diffusion inherent in the upwind scheme. The solid curve shows the initial distribution at time level \( n = 0 \), while the dashed curve (red) shows the distribution at time level \( n = 200 \). The Courant number is \( C = 0.5 \). Cyclic boundary conditions are used at the boundaries of the computational domain.

5.6 Solutions to the advection equation using the MPDATA scheme suggested by Smolarkiewicz (1983). Left panel corresponds to a scaling factor of 1.0 (no scaling), while the right-hand panel employs a scaling factor of 1.3. The Courant number is 0.5 in both cases. Solid, black lines show the initial value (time step \( n = 0 \)), while the red dotted lines show the solution after 200 time steps. The green dashed lines are after 400 time steps while the blue, dash-dot lines are after 800 time steps. Cyclic boundary conditions are employed. (cf. Computer problem No. 6 in the Computer Problem notes).

5.7 Numerical dispersion for the leapfrog scheme. The curves depicts the numerical phase speed as a function of the wavenumber based on (5.87) for various values of the Courant number \( C = u \Delta t/\Delta x \). The vertical axis indicates the phase speed \( c \) normalized by the advection speed \( u \). The horizontal axis indicates the wavenumber normalized by \( \pi/\Delta x \) where \( \Delta x \) is the space increment or the grid size. The analytic dispersion curve is just a straight line corresponding to the phase speed \( c = u \), that is \( c/u = 1 \). Note that as the wavenumber increases (that is the wavelength decreases) the numerical phase speed deviates more and more from the correct analytic phase speed for all values of the Courant number. For wave numbers which gives \( \alpha \Delta x/\pi > 0.5 \), that is for waves of wavelengths \( \lambda < 4\Delta x \) the slope of the curves indicates that the group velocity is negative. Thus for waves with wavelengths shorter than \( 4\Delta x \) the energy is propagating in the opposite direction of the waves.

6.1 Displayed is the cell structure of a staggered grid in one dimension. The circles are associated with \( \phi \)-points, while the horizontal bar is associated with \( u \)-points within the same cells. The sketched staggering is such that the distance between adjacent \( \phi \)-points and \( u \)-points are one half grid distance apart.

6.2 Upper panel shows the Earth’s surface covered by a 2 degree mesh. Lower panel shows a similar mesh of 30 degrees mesh size. The figure conveniently illustrates how a 2 degree mesh in the ocean would look like in the atmosphere scaled by the Rossby radius of deformation.
7.2 Sketch of the mesh in the $t, x$ plane close to the right-hand open boundary. The computational domain is then to the left of $x = L$. The letters $J$, $J - 1$, and $J - 2$ denote grid points respectively at the open boundary, the first and second points inside the computational domain, while $n$, $n - 1$, and $n + 1$ denote the time levels.

7.3 Sketch of the FRS zone, the interior domain and the computational domain. Also shown are the appropriate indices.

10.1 The diagram illustrates the region of stability for the combined advection-diffusion equation approximated in (10.24). This corresponds to the area inside of the parabola (hatched area). The area inside the rectangular is where both the advection and the diffusion are stable individually. We notice that we obtain a more stringent stability condition to the advection equation when we are adding diffusion.

10.2 Displayed are the two waves of wavelength $4\Delta x$ (solid curve) and $\frac{4}{3}\Delta x$ (dashed curve), in a grid of grid size $\Delta x$. Note that our grid cannot distinguish between the unresolved wave of wavelength $\frac{4}{3}\Delta x$ and the resolved wave of wavelength $4\Delta x$. Thus the energy contained in the unresolved wave will be folded into the low wavenumber space represented by the $4\Delta x$ wave.
Chapter 1

INTRODUCTION

We concern ourselves below with the fundamental tools needed to understand how we put oceans and atmospheres on computers. Specifically we concern ourselves with developing methods whereby some important balance equations in oceanography and meteorology, namely the advection-diffusion equation and a simplified form of the shallow water equations on a rotating earth can be solved by numerical means. To this end we make use of finite difference methods.

We assume that the reader has little or no prior knowledge of or experience in solving differential equations numerically. We therefore explain the methods in some detail. The advection-diffusion equation and the shallow water equations belongs to a class of equations known as partial differential equations. Consequently we include in the preliminary chapter (Chapter 2) a rather detailed account on how various types of partial differential equations relates to the advection-diffusion equations and the shallow water equations.

To motivate the reader, and for later reference purposes, we first show how the advection-diffusion equation and the shallow water equations relates to the full equations governing the motion of the atmosphere and ocean. This necessitates a recapitulation of the governing equations, the boundary conditions and the basic approximations commonly made in meteorology and oceanography. We therefore continue this introductory chapter by deriving the shallow water equations from the full governing equations, highlighting the necessary assumptions and approximations needed to derive them. This also conveniently introduces the notation used throughout the text.

1.1 The governing equations

In the atmosphere and ocean the most prominent dependent variables are the three components $u, v,$ and $w$ of the three-dimensional velocity $\mathbf{v}$, pressure $p$, density $\rho$, and (potential) temperature $\theta^1, 2$. For the atmosphere also humidity $q$ and cloud liquid water content $q_{L}$ must be included, while the salinity $s$ must be included among the prominent variables in the ocean. To determine

---

1Velocity is normally referred to as wind in the atmosphere and current in the ocean.

2In the following bold upright fonts, e.g., $\mathbf{u}$, $\mathbf{v}$, are used to denote a vector, while bold special italic fonts, e.g., $\mathbf{U}$, $\mathbf{V}$, are used to denote tensors.
1.1 The governing equations

INTRODUCTION

these unknowns we need an equal number of equations. These equations are normally referred to as the governing equations since they govern the motion of the two spheres atmosphere and ocean.

Of the variables above only the velocity is a vector. The remaining variables, commonly referred to as the state variables, are all scalars. The state variables, except density and pressure, are all examples of what is referred to as tracers. Other examples of tracers are any dissolved chemical component or substance. Since the salinity, temperature and humidity influence the motion via the pressure forcing they are commonly referred to as active tracers. Tracers that do not influence the motion, like for instance dissolved chemical components, are referred to as passive tracers.

As is common when making a mathematical formulation of a physical problem, the governing equations are developed based on conservation principles, in our case the conservation of mass, momentum, internal energy and tracer content. For the atmosphere and ocean the governing equations in their non-Boussinesq form are

\[ \frac{\partial \rho}{\partial t} + \nabla \cdot (\rho \mathbf{v}) = 0, \]  
\[ \frac{\partial (\rho \mathbf{v})}{\partial t} + \nabla \cdot (\rho \mathbf{v} \mathbf{v}) = -2\rho \Omega \times \mathbf{v} - \nabla p + \rho g - \nabla \cdot (\rho \mathbf{F}_M), \]  
\[ \frac{\partial (\rho C_i)}{\partial t} + \nabla \cdot (\rho C_i \mathbf{v}) = -\nabla \cdot (\rho \mathbf{F}_i) + \rho S_i \quad i = 1, 2, \ldots, \]  
\[ \rho = \rho(p, C_1, C_2, \ldots). \]

Here we use \( \partial_t, \partial_x, \partial_y, \text{ and } \partial_z \) to denote partial differential with respect to the respective subscript. Thus \( \partial_t \rho \) is the time derivative (or time rate of change) of the density. \( C_i \) represents the concentration of any tracer including potential temperature and humidity (atmosphere only) and salinity (ocean only), the tensor \( \mathbf{F}_M \) and vector \( \mathbf{F}_i \) represents fluxes due to turbulent mixing of momentum and tracers, respectively, \( \Omega \) is the Earth’s rotation rate, \( g \) is the gravitational acceleration and \( S_i \) is the tracer source, if any. Finally, we use \( \nabla \) to denote the three-dimensional del-operator defined by

\[ \nabla = i \partial_x + j \partial_y + k \partial_z, \]

Of the above equations (1.1) is the conservation of mass, while (1.2) constitutes the conservation of momentum. Furthermore (1.3) is the tracer conservation equation, while (1.4) is the equation of state. Thus the tracers figuring on the right-hand side of (1.4) are limited to active tracers only. Hence (1.4) relates density and pressure to the active tracers.

It should be noted that in the atmosphere the equation of state is linear and follows the ideal gas law, that is,

\[ p = \rho R \theta \]

where \( R \) is the gas constant\(^4\). In contrast the equation of state for the ocean is highly non-linear and cannot be expressed in a formal, closed form. We may visualize the equation of state for the ocean in a so called T-S (temperature-salinity) diagram where the salinity \( s \) is drawn along the horizontal axis and the (potential) temperature \( \theta \) is drawn along the vertical axis. Since also

\(^3\)See for example Gill (1982) or Griffies (2004)

\(^4\)R = 287.04 Jkg\(^{-1}\)K\(^{-1}\)
1.2 Boundary and initial conditions

We observe that to solve (1.1) - (1.4) we need to specify conditions at the spatial boundaries of the domain. Such conditions are referred to as boundary conditions. Furthermore we also need to specify the state of the ocean and/or atmosphere at some given time (cf. the Bjerknes (1904) statement quoted on page iii). The latter is commonly referred to as the initial conditions.

The boundary conditions are of two major types, namely the dynamic boundary conditions and the kinematic boundary conditions. Normally the bounding surface of the volume containing the ocean or the atmosphere is a material surface. We recall that a material surface is a surface that pressure enter the equation of state a T-S diagram can only be constructed using a reference density. A typical example for \( p = 0 \) is displayed in Figure 1.1.

Figure 1.1: The equation of state for the ocean. Reference pressure is \( p = 0 \). Dotted curves show isolines of density as a function of salinity (horizontal axis) and potential temperature (vertical axis) for a fixed pressure (= 0 dbars). Numbers on curves indicate density in \( \sigma_t \) units where \( \sigma_t = \rho - 1000 \text{kg/m}^3 \). Dashed line denotes the freezing point of sea water. Note that for low temperatures (temperatures close to the freezing point of sea water) the density is close to being a function of salinity alone, while the importance of temperature increases with increasing temperature. Due to the non-linear nature of the equation of state for sea water two parcels of water having different temperatures and salinities may still have the same density as for instance the two square points marked A and B along \( \sigma_t = 20.6 \text{kg/m}^3 \).
1.3 The hydrostatic approximation

In the atmosphere and ocean the horizontal scales of the dominant motions are large compared to the vertical scale. As a consequence the vertical acceleration, $Dw/dt$, is small in comparison to, e.g., the gravitational acceleration $\rho g$. Consequently we replace the vertical momentum equation by the hydrostatic equation in which the gravitational acceleration is balanced by the vertical pressure gradient. When one solves this reduced system the model is said to be hydrostatic, and the motion satisfies the hydrostatic approximation.

To illustrate the hydrostatic approximation we first write the vertical component of the momentum equation using (1.2) in full, that is,

$$\partial_t(\rho w) + \nabla \cdot (\rho vw) = -\partial_z p - \rho g - \nabla \cdot (\rho F^V_M),$$

(1.10)

where $F^V_M$ is the vertical vector component of the mixing tensor $F_M$. The hydrostatic assumption implies that the terms on the left-hand side of (1.10) are small compared to the gravitational

---

5We note that in a Cartesian coordinate system fixed to the Earth’s surface the vertical component of the Coriolis force is small compared to the gravitational pull. The former is therefore dropped in 1.10.
acceleration and hence can safely be neglected\(^6\). Furthermore, since the vertical motion is small compared to the horizontal motion also the friction term may be neglected. Under these circumstances the vertical momentum equation reduces to
\[
\frac{\partial}{\partial z} p = -\rho g, \tag{1.11}
\]
which is the hydrostatic equation\(^7\). When the hydrostatic approximation is valid we normally split the momentum equations into its vertical and horizontal components. The vertical component is then the hydrostatic equation (1.11). The two horizontal components are (in vector form)
\[
\partial_t (\rho u) + \nabla_H \cdot (\rho uu) + \partial_z (\rho w u) + \rho f k \times u = -\nabla_H p + \partial_z \tau - \nabla_H \cdot (\rho F_M^H), \tag{1.12}
\]
where \(f = 2 \Omega \sin \phi\) is the Coriolis parameter where \(\phi\) is the latitude and \(\Omega\) is the Earth’s rotation rate. \(F_M^H\) and \(\tau\) are, respectively, the horizontal and vertical component of the three-dimensional flux tensor \(F_M\) due to turbulent mixing. \(\tau\) is also commonly referred to as the vertical shear stress. Note that we in (1.12) have singled out the horizontal convective acceleration due to the vertical velocity and the vertical flux term due to turbulent mixing.

The tracer equation is left unchanged, but as in the momentum equation we may separate the turbulent mixing into one term associated with vertical mixing and one term associated horizontal mixing. Hence it may be written
\[
\partial_t (\rho C_i) + \nabla_H \cdot (\rho C_i u) + \partial_z (\rho C_i w) = -\partial_z (\rho F^V) - \nabla_H \cdot (\rho F_i^H) + \rho S_i \quad i = 1, 2, \cdots, \tag{1.13}
\]
where \(F^V\) and \(F_i^H\) are respectively the vertical and horizontal components of the turbulent mixing.

### 1.4 The Boussinesq approximation

Another common approximation employed, particularly in ocean models, is the Boussinesq approximation. The fundamental basis for this approximation is that in many cases the dynamics of the atmosphere and in particular the oceans is independent of the fact that the atmospheres and oceans are compressible. Under these circumstance we can treat the motion as if the sphere is incompressible. This implies that any parcel of fluid conserves its volume, and that this is true even if the parcel is heated. Thus the Boussinesq approximation is only true as long as the change in density for any parcel of fluid is small with respect to the density itself, that is,
\[
\frac{1}{\rho} \frac{D\rho}{dt} = \frac{D\ln \rho}{dt} \approx 0, \tag{1.14}
\]
\(^6\)As noted this approximation relies on the fact that in most cases the dominant part of the motion, that is, the energetic part, lies in the long wavenumber band, and hence the horizontal scale is significantly longer than the vertical scale (consists of long waves in shallow water). Consequently, both the vertical velocity and its acceleration is small compared to the gravitational acceleration. The exceptions are cases that include steep topography and/or strong convection, in which cases one has to revert to non-hydrostatic equations.
\(^7\)The name is used since a fluid at rest in the gravitational field satisfies exactly this equation. This is often referred to as a static fluid, hence the name hydrostatic.
where the operator $\frac{D}{dt}$ is the material derivative, defined by

$$\frac{D}{dt} = \partial_t + \mathbf{v} \cdot \nabla.$$  \hspace{1cm} (1.15)

Under the Boussinesq approximation the approximation (1.14) is taken as an equality. The mass conservation (1.1) then reduces to

$$\nabla \cdot \mathbf{v} = 0.$$  \hspace{1cm} (1.16)

Use of an ocean model employing the Boussinesq approximation, a Boussinesq ocean, has one major disadvantage. One particularly pertinent example is the expected change in sea level, or ocean volume, under global warming. When uniformly heating the ocean the equation of state implies that the density decreases. For a non-Boussinesq ocean, which is mass conserving, the response to the decrease in density is to expand its volume. Hence the sea level rises. In contrast a Boussinesq ocean, which conserves volume, responds to heating by decreasing the density, that is, by loosing mass. Obviously the latter is highly unrealistic.

The reason why the Boussinesq approximation is still widely used in the ocean modeling community, despite the Boussinesq fluid’s inability to expand due to heating, is the fact that it effectively filters out the acoustic waves while allowing us to retain pressure changes in response to density changes. To filter out the acoustic waves is advantageous in numerical perspective since. A will be shown below (Chapter 5) the time step is then not restricted by these very fast waves (see Section 4.2), dramatically decreasing the wall clock time (or CPU time) spent to perform even relatively short time integrations.

In summary the density under the Boussinesq approximation is treated as a constant except when it appears together with the gravitational acceleration. Under these circumstances the horizontal component of the momentum equation (1.12) becomes

$$\partial_t \mathbf{u} + \nabla \cdot (\mathbf{v} \mathbf{u}) + f \mathbf{k} \times \mathbf{u} = -\rho_0^{-1} \nabla H p + \rho_0^{-1} \partial_z \tau - \nabla H \cdot (\mathbf{F}_M^H),$$  \hspace{1cm} (1.17)

where $\rho_0$ is a reference density. Similarly the tracer conservation equation (1.3) reduces to

$$\partial_t C_i + \nabla \cdot (C_i \mathbf{v}) = -\nabla \cdot \mathbf{F}_i + S_i$$  \hspace{1cm} (1.18)

for $i = 1, 2, \cdots$.

We notice that it is quite common to combine the Boussinesq and the hydrostatic equations in meteorology and oceanography. Under these circumstances the governing equations reduce to

$$\nabla_H \cdot \mathbf{u} + \partial_z w = 0,$$  \hspace{1cm} (1.19)

$$\partial_t \mathbf{u} + \nabla_H \cdot (\mathbf{u} \mathbf{u}) + \partial_z (w \mathbf{u}) + f \mathbf{k} \times \mathbf{u} = -\rho_0^{-1} \nabla_H p + \rho_0^{-1} \partial_z \tau - \nabla_H \cdot (\mathbf{F}_M^H),$$  \hspace{1cm} (1.20)

$$\partial_z p = -\rho g,$$  \hspace{1cm} (1.21)

$$\partial_t C_i + \nabla_H \cdot (C_i \mathbf{u}) + \partial_z (C_i w) = -\partial_z P^V - \nabla_H \cdot \mathbf{F}_i^H + S_i ; \hspace{0.5cm} i = 1, 2, \cdots,$$  \hspace{1cm} (1.22)

\footnote{Also referred to as the individual derivative}
together with the equation of state (1.4). We note that when applying the hydrostatic and Boussinesq approximation the vertical velocity component and the density are reduced to diagnostic variables just as pressure. This is in contrast to the horizontal velocity components \(u\) and the tracers \(C_i\), e.g., potential temperature, which are prognostic variables in the sense that they are governed by prognostic equations, that is, equations containing a time rate of change term of the variable in question.

Finally we note that the introduction of more and more simplifications sometimes complicates the numerical problem. For instance the fairly popular rigid lid approximation implies that the equations must be solved globally rather than locally since the solution at one point not only depends on its nearest neighboring points, but in fact depends on all the points within the computational domain. This requires us to solve an elliptic problem for each time step, although the problem in itself, as a time marching problem is hyperbolic\(^9\). We will return to elliptic solvers in Section 4.8 on page 53 when solving an elliptic problem by a direct method.

### 1.5 The shallow water equations

A very common reduced set of equations in meteorology and oceanography is the so called shallow water equations. We may derive these equations from the full governing equation (1.1) - (1.4). We first assume that the hydrostatic and Boussinesq approximations are valid. Hence the starting point is mass conservation in the form (1.19), the momentum equations in the form (1.21) and (1.20), the tracer equation in the form (1.22) together with the equation of state (1.4). The additional assumption made is that the density is assumed to be uniform in time and space, i.e., \(\rho = \rho_0\) where \(\rho_0\) is a constant. We note that this makes the tracer equations (1.22) for the active tracers as well as the equation of state (1.4) obsolete. The resulting governing equations then reduces to

\[
\nabla_H \cdot u + \partial_z w = 0 \tag{1.23}
\]
\[
\partial_z p = -\rho_0 g, \tag{1.24}
\]
\[
\partial_t u + \nabla_H \cdot (uu) + \partial_z (wu) = -f k \times u - \rho_0^{-1} \nabla_H p + \rho_0^{-1} \partial_x \tau - \nabla_H \cdot \mathcal{F}_M \tag{1.25}
\]

We note the assumption of a uniform density allow us to integrate (1.24) from any arbitrary height/depth \(z\) to a reference surface \(z = \eta(x,y,t)\), viz.,

\[
p = p_s + g \rho_0 (\eta - z) \tag{1.26}
\]

where \(p_s\) is the pressure at the reference surface. In the ocean the reference surface is commonly the surface of the ocean in which case \(\eta\) is the deviation of the sea surface from its equilibrium level \(z = 0\). In the atmosphere it is common to let the reference surface be the surface of the Earth, e.g., \(\eta = -H\), where \(H\) is measured as the distance from some fixed level (commonly set to \(z = 0\)).

\(^9\)For definitions of elliptic and hyperbolic problems see Sections 2.2 and 2.4 of Chapter 2
1.5 The shallow water equations

Integrating (1.23) and (1.25) from the bottom \( z = -H(x, y) \) to the top \( z = \eta(x, y, t) \), and using the kinematic boundary conditions (1.7) and (1.9) and the dynamic boundary condition \( p = 0 \) at \( z = \eta \) we get,

\[
\begin{align*}
\partial_t U + \nabla_H \cdot \left( \frac{UU}{h} \right) + f k \times U &= -gh \nabla_H (h - H) + \rho_0^{-1}(\tau_s - \tau_b) + X, \\
\partial_t h + \nabla_H \cdot U &= 0,
\end{align*}
\]

where \( U = \int_{-H}^{\eta} u \, dz \) is the volume flux of fluid through the fluid column of height/depth \( h = \eta + H \), \( \tau_s \) and \( \tau_b \) are, respectively, the turbulent vertical momentum fluxes at the top and bottom of the fluid column, and \( X \) is what is left of the horizontal momentum fluxes when integrated vertically from bottom to top. To arrive at (1.27) we have also integrated (1.24) from some arbitrary height/depth \( z \) to the top \( z = \eta \). Furthermore we have used the fact that \( H \) is independent of time to replace, e.g., \( \partial_t \eta \) by \( \partial_t h \). Finally we have absorbed the term arising from the approximation

\[
\int_{-H}^{\eta} \nabla_H \cdot (u u) \, dz \approx \nabla_H \cdot \left( \frac{UU}{h} \right)
\]

into the last term \( X \) on the right-hand side of (1.27). We commonly refer to (1.27) and (1.28) as the shallow water equations. Written in this form the shallow water equations are said to be written in flux form. We note that \( U \) is the total volume flux of fluid through the fluid column of height/depth \( h \). Thus the mean depth average velocity is \( \hat{u} = U/h \). Replacing \( U \) by \( \hat{u} \) the shallow water equations become

\[
\begin{align*}
\partial_t (h \hat{u}) + \nabla_H \cdot (h \hat{u} \hat{u}) + f k \times h \hat{u} &= -gh \nabla_H (h - H) + \rho_0^{-1}(\tau_s - \tau_b) + X, \\
\partial_t h + \nabla_H \cdot (h \hat{u}) &= 0,
\end{align*}
\]

For later reference purposes we note that the acceleration terms \( \partial_t (h \hat{u}) + \nabla_H \cdot (h \hat{u} \hat{u}) \) in (1.30) can be written

\[
\begin{align*}
\partial_t (h \hat{u}) + \nabla_H \cdot (h \hat{u} \hat{u}) &= h (\partial_t \hat{u} + \hat{u} \cdot \nabla_H \hat{u}) + \hat{u} [\partial_t h + \nabla_H \cdot (h \hat{u})] \\
&= h (\partial_t \hat{u} + \hat{u} \cdot \nabla_H \hat{u}),
\end{align*}
\]

where we have used (1.31) to arrive at the last equal sign. Thus (1.30) and (1.31) is written

\[
\begin{align*}
\partial_t \hat{u} + \hat{u} \cdot \nabla_H \hat{u} + f k \times \hat{u} &= -g \nabla_H \eta + \frac{\tau_s - \tau_b}{\rho_0(H + \eta)} + \frac{X}{(H + \eta)}, \\
\partial_t \eta + \nabla_H \cdot [(H + \eta) \hat{u}] &= 0,
\end{align*}
\]

We note that when the shallow water equations are written in their non-flux form, as displayed in (1.33) and (1.34), the mass conservation equation (1.34) becomes non-linear as well. This is in contrast to the mass conservation equation in flux form, that is, (1.28), which is linear.
1.6 The quasi-geostrophic equations

Another common set of reduced equations are based on quasi-geostrophic theory as for instance detailed in Pedlosky (1979) or Stern (1975). Here we essentially follow Stern (1975).

We first note that the starting point for the quasi-geostrophic equations are the governing equations employing the hydrostatic and Boussinesq approximations. Without loss of generality we may therefore start with the shallow water equations (1.33) and (1.34). If we neglect the forcing terms on the right-hand side of (1.33) we get

\[
\frac{D_H u}{dt} + f k \times u = -g \nabla h, \tag{1.35}
\]

\[
\frac{1}{h} \frac{D_H h}{dt} + \nabla_h \cdot u = 0, \tag{1.36}
\]

where we have dropped the circumflex for clarity. The notation \( \frac{D_H}{dt} \) is used to denote the two-dimensional version of the operator (1.15), that is,

\[
\frac{D_H}{dt} = \partial_t + u \cdot \nabla h, \tag{1.37}
\]

If we furthermore assume that the acceleration \( \frac{D_H u}{dt} \) is small compared to the Coriolis acceleration, that is, assume that the Rossby number

\[
R \equiv \frac{|\frac{D_H u}{dt}|}{|f k \times u|} \ll 1, \tag{1.38}
\]

the momentum equation (1.35) reduces to

\[
f k \times u = -g \nabla h. \tag{1.39}
\]

We note that (1.39) is linear and describes a balance between the Coriolis term and the pressure term. Hence (1.39) is referred to as the geostrophic or thermal wind equation and the balance is called the geostrophic balance. Solving for the relative velocity we get

\[
u = \frac{g}{f} k \times \nabla h. \tag{1.40}
\]

The introduction of the Rossby number tells us that we may view the thermal wind equation as a first approximation to an expansion in terms of the Rossby number in which terms of \( \mathcal{O}(R) \) or higher are neglected, that is,

\[
u = \frac{g}{f} k \times \nabla h + \mathcal{O}(R). \tag{1.41}
\]

We note that (1.39) obviously provides no information about the space-time variations in either the velocity field or the pressure field. To obtain information on those dynamics to order \( \mathcal{O}(R) \) we have to look elsewhere. For instance we may obtain it from the relevant asymptotic form of the vorticity equation.
To derive the vorticity equation we start by defining the relative vorticity

\[ \zeta = k \cdot \nabla_H \times u. \]  

(1.42)

Then operating \( k \cdot \nabla_H \times \) on (1.35) and then substituting for \( \nabla_H \cdot u \) from (1.36) we get

\[ \frac{D_H}{dt} \left( \frac{\zeta + f}{h} \right) = 0. \]  

(1.43)

Here \( \zeta + f \) is the absolute vorticity, while \( (\zeta + f)/h \) is the potential vorticity for a barotropic fluid\(^{10} \). If we let \( L \) be a typical lateral (horizontal) scale of \( u \), so that \(|u \cdot \nabla_H u| \sim |u|^2/L \), then the necessary condition for \( R \ll 1 \), which is commonly referred to as the quasi geostrophy condition (cf. eq. 1.38), to be satisfied is

\[ \frac{|u|}{fL} \ll 1. \]  

(1.44)

The condition is however not sufficient since the remaining acceleration term in \( D_H/dt \) is the local time rate of change \( \partial_t u \) which might be comparable to the Coriolis term \( f k \times u \). Consequently we must additionally require that the initial condition is in geostrophic balance, that is, satisfies (1.39). The smallness of \( \partial_t u \) compared to the Coriolis acceleration then depends on the smallness of \( u \cdot \nabla_H u \). Under these circumstances we may safely regard (1.44) as being the same as requiring \( R = |u|/fL \ll 1 \). We may then compute the temporal evolution of the geostrophic field from the asymptotic vorticity equation.

To derive the asymptotic vorticity equation we first observe, by use of (1.42), that

\[ \frac{|\zeta|}{f} = \frac{|u|}{fL}. \]  

(1.45)

Hence (1.44) requires that the relative vorticity is small compared to \( f \) by a factor of \( R \). We also note that the variation in layer thickness \( h \), obtained from (1.39) is

\[ h - H_m \sim \frac{fL}{g} |u| \quad \text{or} \quad \frac{h - H_m}{H_m} \sim RF^2, \]  

(1.46)

where \( H_m \) is the mean layer thickness and

\[ F = \left( \frac{L_D^2}{L_D} \right)^{\frac{1}{2}}, \]  

(1.47)

where \( L_D = gH_m/f^2 \) is the Rossby radius of deformation. If we now assume \( F \sim \mathcal{O}(1) \) or less, which is tantamount to assuming that \( L \) is not large compared to Rossby’s deformation radius, then the layer thickness variation in (1.46) is small to the same order as the ratio of the relative vorticity \( \zeta \) to the planetary vorticity \( f \), that is, \( \zeta/f \).

\(^{10}\)Recall that we have assumed that the density is constant. The fluid is therefore barotropic. The potential vorticity may also be derived for a baroclinic fluid in a similar fashion, but has then a different mathematical expression.
Under these circumstances we first rewrite the potential vorticity equation (1.43) to get
\[
\frac{D}{dt} \left( \frac{\zeta + f}{h} \right) = \frac{f}{H_m} \left( \partial_t + \mathbf{u} \cdot \nabla \right) \left( \frac{1 + \frac{\zeta}{f}}{1 + \frac{h - H_m}{H_m}} \right) = 0. \tag{1.48}
\]
We are now in a position to expand this expression in terms of \( R \), and thus we get
\[
\left( \partial_t + \mathbf{u} \cdot \nabla \right) \left[ 1 + \frac{\zeta}{f} - \frac{h - H_m}{H_m} + \mathcal{O}(R^2) \right] = \left( \partial_t + \mathbf{u} \cdot \nabla \right) \left( \frac{\zeta}{f} - \frac{h}{H_m} \right) + \mathcal{O}(R^3) = 0. \tag{1.49}
\]
The leading terms in (1.49) are \( \mathcal{O}(R^2) \) since the (non-dimensional) magnitude of the acceleration terms \( \partial_t \mathbf{u} \) and \( \mathbf{u} \cdot \nabla \mathbf{u} \) are \( \mathcal{O}(R) \). Thus the fractional error in the asymptotic vorticity equation
\[
\left( \partial_t + \mathbf{u} \cdot \nabla \right) \left( \frac{\zeta}{f} - \frac{h}{H_m} \right) = 0 \tag{1.50}
\]
and in the asymptotic momentum equation (1.39) are both of \( \mathcal{O}(R) \). Hence substitution of \( \mathbf{u} \) from (1.40) wherever the latter appears in (1.50) the resulting differential equation for the layer thickness (or pressure) \( h \) is also asymptotic when \( R \ll 1 \). It is thus permissible to evaluate the velocity and the relative vorticity in (1.50) using the geostrophic equation (1.39) or (1.40). In fact if we substitute the expression (1.40) for \( \mathbf{u} \) into (1.42) and then into (1.50) we first get
\[
\zeta = \frac{g}{f} \nabla_H^2 h, \tag{1.51}
\]
and then
\[
\left[ \partial_t + \frac{g}{f} (\mathbf{k} \times \nabla_H h) \cdot \nabla_H \right] \left( \nabla_H^2 h - L_D^{-1} h \right) = 0. \tag{1.52}
\]
Equation (1.51) and (1.52) together with the geostrophic equation (1.39) are commonly referred to as the quasi-geostrophic equations (QG equations). Thus we may use the quasi-geostrophic vorticity equation (1.52) to compute the pressure or layer thickness \( h \) at an arbitrary time \( t > 0 \) from any initial distribution at time \( t = 0 \). The resulting solution is then almost geostrophic, but not quite. Hence we use the name quasi-geostrophic. We emphasize that it is only under very stringent conditions, as explained, that these equations are valid.

We finally remark that, although each step in the hierarchy of the approximations, that is the Boussinesq approximation, the hydrostatic approximation, the shallow water equations, and finally the quasi-geostrophic approximation, removes or filters out a certain class of phenomena, the advantage of such procedures is that they allow us to isolate effects having different space-time scales. In a numerical contexts they are also very useful in establishing solutions against which our numerical solutions may be tested or verified.
1.6 The quasi-geostrophic equations

INTRODUCTION
Chapter 2

PRELIMINARIES

The equations that govern the motion of the atmosphere and the ocean, as well as the hierarchy of equations that follows employing the various approximations as outlined in the introductory chapter (Chapter 1), belongs to a class of equations called partial differential equations (henceforth PDEs). They differ from ordinary differential equations in that there are more than one independent variable, and sometimes several dependent variables.

In this chapter we learn more about PDEs and reveal that they have different characters depending on the physics they describe. We also introduce some basic mathematics underlying two of the most important numerical methods used to solve atmospheric and oceanic problems, namely finite difference methods and spectral methods. These mathematics include knowledge about Taylor series expansions, orthogonal functions, Fourier series and Fourier transforms. Finally, we include some notations that conveniently helps us to solve PDEs using numerical methods.

2.1 General PDEs

In general a PDE is written

\[ \hat{a} \partial_{x'}^2 \theta + 2 \hat{b} \partial_{x'} \partial_{y'} \theta + \hat{c} \partial_{y'}^2 \theta + 2 \hat{d} \partial_{x'} \theta + 2 \hat{e} \partial_{y'} \theta + \hat{f} \theta = \hat{g}. \] (2.1)

Here \( \partial_{x'}, \partial_{y'} \) denotes differentiation with respect to the independent variables \( x', y' \), while \( \theta = \theta(x', y') \) denotes the dependent variable. The coefficients \( \hat{a}, \hat{b}, ..., \hat{g} \) are in general functions of the independent variables, that is, \( \hat{a} = \hat{a}(x', y') \), etc. Note that \( x' \) and \( y' \) represents any independent variable, for instance time or one of the spatial variables, while \( \theta \) represents any dependent variable, e.g., velocity, pressure, density, salinity, or humidity.

We illustrate this by using the shallow water equations (1.33) and (1.34) as an example. To simplify the illustration we first neglect all forcing terms. We then linearize them by assuming that the deviation of the height \( h \) of a fluid column is small compared to its equilibrium depth
2.2 Elliptic equations

If \( \hat{b}^2 - \hat{a}\hat{c} < 0 \) then the roots of (2.1) are imaginary, distinct, and complex conjugated. The corresponding PDE is then elliptic. The classic example is Poisson’s equation,

\[
\nabla_{\tilde{H}}^2 \phi = \partial_x^2 \phi + \partial_y^2 \phi = g(x, y),
\]

where again \( \nabla_{\tilde{H}} \) is the two-dimensional part of the three-dimensional del operator. We arrive at this equation by letting \( \theta = \phi, x' = x, y' = y, \hat{a} = \hat{c} = 1, \hat{\epsilon} = \hat{\tilde{g}} = g \) and \( \hat{b} = \hat{d} = \hat{\epsilon} = \hat{\tilde{f}} = 0 \) in (2.1). Other examples are the Helmholtz equation

\[
\nabla_{\tilde{H}}^2 \phi + f(x, y)\phi = g(x, y),
\]

and the Laplace equation

\[
\nabla_{\tilde{H}}^2 \phi = 0.
\]

2.3 Parabolic equations

If \( \hat{b}^2 - \hat{a}\hat{c} = 0 \) then the corresponding PDE is parabolic. The classic example is the diffusion equation or the heat conduction equation,

\[
\partial_t \theta = \kappa \partial_x^2 \theta,
\]

where \( \kappa \) is the diffusion coefficient (heat capacity). To arrive at (2.9) from (2.1) we let \( \theta = \theta, x' = x, y' = t, \hat{a} = 1, \hat{b} = \hat{c} = \hat{d} = \hat{\tilde{f}} = \hat{\tilde{g}} = 0, \) and \( \hat{\epsilon} = 1/2 \). We observe that (2.9) is
a simplified, one-dimensional version of the full three-dimensional tracer equation (1.3), where the advection term as well as the source and sink terms are neglected. In fact under the latter circumstances the three-dimensional tracer equation (1.18) for a Boussinesq fluid may be written

$$\frac{\partial}{\partial t}C_i = \nabla \cdot (\mathbf{K} \cdot \nabla C_i),$$

(2.10)

where the diffusive tracer flux $F_i$ is parameterized as $F_i = -\mathbf{K} \cdot \nabla C_i$ where in turn $\mathbf{K}$ is a matrix (dyade) describing the conductive efficiency of the medium with regard to the tracer $C_i$ (cf. Section 3.1). Thus $\mathbf{K} = \kappa_{mn} \mathbf{I}_m \mathbf{I}_n$, $m, n = 1, 2, 3$. To retrieve (2.9) we simply let $\kappa_{11} = \kappa$ and $\kappa_{mn} = 0$ for $m \neq 1$ and $n \neq 1$ and assume that $\kappa$ is constant.

Let us for a moment assume that the atmosphere/ocean is at rest ($v = 0$) and that there are no sources or sinks for the tracer $C_i$ ($S_i = 0$). Then (1.3) reduces to (2.10) implying that the diffusion balance is of fundamental importance when solving atmospheric and oceanographic problems.

### 2.4 Hyperbolic equations

If $\hat{b}^2 - \hat{a}\hat{c} > 0$ then the roots of (2.1) are real and distinct. The corresponding PDE is then hyperbolic. The classic example is the wave equation,

$$\frac{\partial^2}{\partial t^2} \phi - c_0^2 \frac{\partial^2}{\partial x^2} \phi = 0.$$  

(2.11)

To derive (2.11) from (2.1) we let $\theta = \phi, x' = t, y' = x, \hat{a} = 1, \hat{b} = 0, \hat{c} = -c_0^2$, and $\hat{d} = \hat{e} = \hat{f} = \hat{g} = 0$. Then $\hat{b}^2 - \hat{a}\hat{c} = -(c_0^2)^2 = c_0^2$ which is indeed positive.

We note that by defining

$$\Phi = \partial_t \phi - c_0 \partial_x \phi$$

(2.12)

we may rewrite the wave equation (2.11) to get

$$\partial_t \Phi + c_0 \partial_x \Phi = 0.$$  

(2.13)

Since $c_0$ is a constant (2.13) may be written

$$\partial_t \Phi + \partial_x (c_0 \Phi) = 0.$$  

(2.14)

We observe that (2.14), commonly referred to as the advection equation, is a one-dimensional version of (1.1) with $\rho$ replaced by $\Phi$ and $v$ replaced by $c_0 \hat{v}$. It is also a one-dimensional version of (1.3) with suitable replacements. Thus the advection equation is of fundamental importance in the modeling of atmospheric and oceanographic motions. It also indicates that the equations governing atmospheric and oceanographic motions viewed as time marching problems are inherently hyperbolic.

We also notice that the two-dimensional version of the shallow water equations (1.27) and (1.28) is inherently a hyperbolic problem. To illustrate this we start with the two-dimensional, linearized version of the shallow water equation, that is, (2.2), (2.3) and (2.4). We start by
manipulating (2.2) and (2.3) to find $u, v$ as functions of $h$. To this end we first differentiate (2.2) with respect to time, and then add (2.3) multiplied by $f$. This results in an equation containing $h$ and $u$ only, that is,

$$ \left( \partial_t^2 + f^2 \right) u = -gf \partial_y h - g \partial_t \partial_x h. \quad (2.15) $$

Similarly by first differentiate (2.3) with respect to time and then adding (2.2) multiplied by $-f$ gives an expression relating $h$ and $v$, that is,

$$ \left( \partial_t^2 + f^2 \right) v = gf \partial_x h - g \partial_t \partial_y h. \quad (2.16) $$

Next we substitute the results into (2.4) to get

$$ \left( \partial_t^2 + f^2 - gHm \nabla^2_H \right) \partial_t h = 0. \quad (2.17) $$

Let $h = H_m + h'$ and let $h' = 0$ at time $t = 0$. Integration in time $t$ then yields

$$ \left( \partial_t^2 + f^2 - gHm \nabla^2_H \right) h' = 0. \quad (2.18) $$

If we in addition assume that the motion is independent of one of the dependent variables, say $y$, we get

$$ \left( \partial_t^2 + f^2 - gHm \nabla^2_H \right) h'' = 0. \quad (2.19) $$

We note that (2.19) is hyperbolic in $t$ and $x$. Similarly we observe that (2.18) is elliptic in $x$ and $y$. Thus, we note that although the steady state solution to (2.19) is elliptic, the time marching problem is inherently hyperbolic.

The governing equations describing the time evolution of atmospheric and oceanographic motions are fundamentally hyperbolic. It is important to keep this in mind when developing numerical methods to solve atmosphere-ocean problems.

We will return to the shallow water equations in Section 6.1 on page 92. There we use them as an example problem to show how multiple variable problems are solved using numerical methods.

2.5 Boundary conditions

To solve for the dependent variables we have to integrate the governing PDE in time and space. Thus the solution inherently contains integration constants. The number of integration constants is determined by the order of the PDE. For instance upon integration the linearized shallow water equations (2.2) - (2.4) in time $t$ gives three integration constants, while integration in space ($x, y$) gives another four integration constants (two in $x$ and two in $y$), a total of seven. Thus we need seven conditions to determine these constants. These conditions are commonly referred to as boundary conditions.

We emphasize that the number of boundary conditions needed must be exactly the same as the number of integration constants, no more, no less. If we specify too many boundary
conditions the system is over-specified, and if we specify too few we end up with an under-
specified system. It is therefore imperative that we adhere to this fundamental principle when
we make use of numerical methods to solve our governing equations. We emphasize that the
computer always produce numbers. If we over- or under-specify our system, the computer will
still produce numbers. These numbers may even look realistic or correct, but are nevertheless
incorrect. The reason is that the only way to ensure that our solution exists and is unique is
to have an equal number of boundary conditions and integration constants. Furthermore, as a
corollary, the solution to our problem is equally dependent on the boundary conditions as on any
other forcing.

To determine for instance the solution to the elliptic Poisson equation (2.6) we need four
boundary conditions, two in \( x \) and two in \( y \). To determine the solution to the diffusion equation
(2.9) we need three boundary conditions, two in \( x \) and one in time \( t \). Finally, to determine the
solutions to the wave equation we need a total of four conditions to determine the four integration
constants, namely two in \( t \) and two in \( x \). As we increase the dimensions of the equation we
note that the number of integration constants increases and thus also the number of boundary
conditions needed.

There are essentially two types of boundary conditions belonging to the class of \textit{natural}
boundary conditions\(^1\), namely

- Dirichlet conditions,

in which case the variable is known at the boundary, and

- Neuman conditions,

in which case the derivative normal to the boundary is specified. Most other boundary conditions
are just combinations of these.

In Section 1.2 we mentioned that there cannot be any flow through an impermeable wall, that
is, no throughflow across a solid wall, and formulated this condition as

\[
\mathbf{n} \cdot \mathbf{v} = 0 \quad (2.20)
\]

at the wall surface. Here \( \mathbf{n} \) denotes the unit vector perpendicular to the wall. In fact this is a
classic example of the Dirichlet type boundary condition in that specifying the condition
\( \mathbf{n} \cdot \mathbf{v} = 0 \) at the surface constituting the wall is tantamount to specifying the variable itself at the surface.
Thus the condition is of the Dirichlet type.

Next we may derive a classic example of a Neuman type condition by the condition prevailing
at an insulated wall. The natural condition dictated by the physics is that for the wall to be
insulated there cannot be any heat exchange across the boundary. Thus the diffusive flux of heat
through the boundary must be zero. In mathematical terms this is tantamount to

\[
\mathbf{n} \cdot \mathbf{F}_\theta = 0, \quad (2.21)
\]

\(^1\)A natural boundary condition is one in which the condition is dictated by the physics. This is in contrast to open
boundaries treated in Chapter 7.
2.6 Taylor series and expansions

where $\mathbf{F}_{\theta} = -\kappa \nabla \theta$ is the diffusive heat flux vector. Thus (2.21) is the same as specifying the gradient (in this case a zero gradient) at the boundary. Thus the (2.21) is of the Neuman type.

As alluded to the two conditions may be combined to give other natural boundary conditions. One is the so called Cauchy condition or “slip” condition. For instance consider a flat bottom or surface at $z = -H$ (or $z = 0$) at which we give the following condition

$$\nu \partial_z u = C_D u \quad ; \quad z = -H,$$  \hfill (2.22)

where $\nu$ is the vertical eddy viscosity, $u$ is the horizontal component of the current (or wind), and $C_D$ is a drag coefficient (more often than not the latter is a constant).

Other common boundary conditions are cyclic or periodic boundary conditions. A periodic boundary condition is one in which the solution is specified to be periodic in space, that is, that the solution repeats itself beyond a certain distance. Thus a periodic boundary condition in $x$ for a given tracer concentration $C(x)$ would be

$$C(x, t) = C(x + L, t),$$  \hfill (2.23)

where $L$ is the distance over which the solution repeats itself, for instance the wavelength in a monochromatic wave. Such conditions are commonly in use when solving problems where the atmosphere or ocean is considered to be contained in a zonal channel bounded to the south and north by a zonal wall. In the longitudinal direction the solution is then dictated by physics to naturally repeat itself every 360 degrees.

2.6 Taylor series and expansions

The basis for all numerical finite difference methods is that all “good” functions can be expanded in terms of a Taylor series. A good function is simply one for which the function itself and all its derivatives exist and are continuous\(^2\). One characteristic of a good function is that it can always be expanded in a so called Taylor series. Another is that it can be represented by an infinite sum of orthogonal functions such as for instance trigonometric function (Sections 2.10 and 2.11).

Consider the function $\theta(x, t)$ to be a good function. Then we know $\theta(x, t)$ and all its derivatives at a particular point in space, say $x = x_0$. We may then use a Taylor series expansion to find the values of $\theta$ at the neighboring point $x_0 + \Delta x$, where $\Delta x > 0$ denotes a positive increment in space. Thus,

$$\theta(x_0 + \Delta x, t) = \theta(x_0, t) + \partial_x \theta(x_0, t) \Delta x + \frac{1}{2} \partial_x^2 \theta(x_0, t) \Delta x^2 + \frac{1}{6} \partial_x^3 \theta(x_0, t) \Delta x^3 + \mathcal{O}(\Delta x^4).$$  \hfill (2.24)

We may repeat this for any point in space, thus in general we have

$$\theta(x + \Delta x, t) = \theta(x, t) + \partial_x \theta(x, t) \Delta x + \frac{1}{2} \partial_x^2 \theta(x, t) \Delta x^2 + \frac{1}{6} \partial_x^3 \theta(x, t) \Delta x^3 + \mathcal{O}(\Delta x^4).$$  \hfill (2.25)

\(^2\)This definition is somewhat different from the one offered in the little known but enlightening book by M. J. Lighthill entitled “Good functions” (Lighthill, 1970)
If we solve (2.25) with respect to the first derivative we get
\[ \partial_x \theta(x, t) = \frac{\theta(x + \Delta x, t) - \theta(x, t)}{\Delta x} + \mathcal{O}(\Delta x). \]  
(2.26)

We may repeat this procedure by using a Taylor series to find the value of the function \( \theta(x, t) \) at the point \( x - \Delta x \) to get
\[ \theta(x - \Delta x, t) = \theta(x, t) - \partial_x \theta(x, t) \Delta x + \frac{1}{2} \partial_x^2 \theta(x, t) \Delta x^2 - \frac{1}{6} \partial_x^3 \theta(x, t) \Delta x^3 + \mathcal{O}(\Delta x^4). \]  
(2.27)

If we again solve with respect to the first derivative we get
\[ \partial_x \theta(x, t) = \frac{\theta(x, t) - \theta(x - \Delta x, t)}{\Delta x} + \mathcal{O}(\Delta x). \]  
(2.28)

Moreover, by subtracting (2.27) from (2.25), and solving for the first derivative of \( \theta \) at the point \( (x, t) \) we get
\[ \partial_x \theta(x, t) = \frac{\theta(x + \Delta x, t) - \theta(x, t) + \theta(x - \Delta x, t)}{2 \Delta x} + \mathcal{O}(\Delta x^2). \]  
(2.29)

We emphasize that the choice \( \Delta x \) is arbitrary, we may equally well choose \( 2\Delta x \), in which case the Taylor series (2.25) and (2.27) becomes
\[ \theta(x + 2\Delta x, t) = \theta(x, t) + 2 \partial_x \theta(x, t) \Delta x + 2 \partial_x^2 \theta(x, t) \Delta x^2 + \frac{4}{3} \partial_x^3 \theta(x, t) \Delta x^3 + \mathcal{O}(\Delta x^4), \]  
(2.30)

\[ \theta(x - 2\Delta x, t) = \theta(x, t) - 2 \partial_x \theta(x, t) \Delta x + 2 \partial_x^2 \theta(x, t) \Delta x^2 - \frac{4}{3} \partial_x^3 \theta(x, t) \Delta x^3 + \mathcal{O}(\Delta x^4). \]  
(2.31)

Again by subtracting the two and solving for \( \partial_x \theta \) we get
\[ \partial_x \theta(x, t) = \frac{\theta(x + 2\Delta x, t) - \theta(x - 2\Delta x, t)}{4 \Delta x} + \mathcal{O}(\Delta x^2). \]  
(2.32)

### 2.7 Finite difference approximations

To derive possible finite difference approximations (FDAs) to the various derivatives of our PDEs we actually utilize the Taylor series above. For instance to derive possible FDAs to the first derivative \( \partial_x \theta \), that is, the first derivative of \( \theta \) with respect to \( x \), we utilize the expressions (2.29), (2.26) and (2.28) above. For instance using (2.29) we simply neglect the higher order terms, in this case terms of \( \mathcal{O}(\Delta x^2) \), to obtain
\[ [\partial_x \theta]^t_x = \frac{\theta(x + \Delta x, t) - \theta(x, t)}{2 \Delta x}. \]  
(2.33)

\(^3\)Henceforth an FDA is denoted by brackets, that is, the FDA approximation to \( \partial_x \theta \) at the point \( x, t \) is denoted \([\partial_x \theta]^t_x\). We will also henceforth make use of the notation \( \partial_x \theta(x, t) = \partial_x \theta^t_x \) and \( \partial_x^2 \theta(x, t) = \partial_x^2 \theta^t_x \).
We emphasize that this approximation is valid to $O(\Delta x^2)$. Similarly we derive other possible FDAs of $\partial_x \theta$ by neglecting terms of $O(\Delta x)$ in (2.26) and (2.28), respectively, that is,

$$
[\partial_x \theta]^t_x = \frac{\theta(x + \Delta x, t) - \theta(x, t)}{\Delta x},
$$

and

$$
[\partial_x \theta]^t_x = \frac{\theta(x, t) - \theta(x - \Delta x, t)}{\Delta x}.
$$

We note that while (2.33) is centered on the spatial point $x$, (2.34) and (2.35) are one-sided. The approximation (2.33) is therefore denoted a centered approximation, while (2.34) and (2.35) are denoted a forward, one-sided approximation and a backward, one-sided approximation respectively, or simply forward and backward approximations. Another difference between the centered and the forward and/or backward FDAs is the order of the neglected terms. While the centered difference is valid to second order, that is, $O(\Delta x^2)$, the forward and backward FDAs are only valid to first order, that is, $O(\Delta x)$.

We may perform exactly the same calculations based on Taylor series expansion to derive a FDA to the derivatives in time $t$. For instance by expanding $\theta$ in time we get

$$
\theta|^t_{x+\Delta t} = \theta|^t_x + \partial_t \theta|^t_x \Delta t + \frac{1}{2} \partial^2_t \theta|^t_x \Delta t^2 + \frac{1}{6} \partial^3_t \theta|^t_x \Delta t^3 + O(\Delta t^4),
$$

and

$$
\theta|^t_{x-\Delta t} = \theta|^t_x - \partial_t \theta|^t_x \Delta t + \frac{1}{2} \partial^2_t \theta|^t_x \Delta t^2 - \frac{1}{6} \partial^3_t \theta|^t_x \Delta t^3 + O(\Delta t^4).
$$

To construct a centered FDA to the time rate of change of $\theta$ we simply subtract (2.37) from (2.36) and solve with respect to $\partial_t \theta$ to obtain

$$
\partial_t \theta|^t_x = \frac{\theta|^t_{x+\Delta t} - \theta|^t_{x-\Delta t}}{2 \Delta t} + O(\Delta t^2),
$$

which when dropping terms of $O(\Delta t^2)$ gives the FDA

$$
[\partial_t \theta]^t_x = \frac{\theta|^t_{x+\Delta t} - \theta|^t_{x-\Delta t}}{2 \Delta t}.
$$

Thus we observe that the centered in time FDA (2.39) is valid to second order.

Similarly we may construct approximations to higher order derivatives. For instance to find a centered FDA to the second order derivative of $\theta$ with respect to $x$ we first simply add the two Taylor expansion (2.25) and (2.27) and solve with respect to $\partial^2_x \theta$ to give

$$
\partial^2_x \theta|^t_x = \frac{\theta|^t_{x+\Delta x} - 2 \theta|^t_x + \theta|^t_{x-\Delta x}}{\Delta x^2} + O(\Delta x^2).
$$

Then by neglecting terms of $O(\Delta x^2)$ in (2.40) a FDA to the second order derivative is

$$
[\partial^2_x \theta]^t_x = \frac{\theta|^t_{x+\Delta x} - 2 \theta|^t_x + \theta|^t_{x-\Delta x}}{\Delta x^2}.
$$
Since this expression gives equal weight to the points $x + \Delta x$ and $x - \Delta x$, that is, to the points on either side of $x$, the approximation is centered. Like in (2.33) we note that the neglected terms are of $O(\Delta x^2)$. This is in contrast to the forward and backward approximations in which the neglected terms were of $O(\Delta x)$. Thus the centered approximations appear to share the fact that the neglected terms are of higher order than the one-sided approximations.

As exemplified in (2.41) we may formulate FDAs to any higher order derivative with respect to $t$, $x$ and other spatial independent variables. For instance to derive a centered in space FDA for $\partial^3_x \theta$ we combine (2.25) and (2.27) with (2.30) and (2.31) to first obtain

$$\partial^3_x \theta|_x^{t} = \frac{\theta|_x^{t+2\Delta x} - 2\theta|_x^{t+\Delta x} + 2\theta|_x^{t-\Delta x} - \theta|_x^{t-2\Delta x}}{2\Delta x^3} + O(\Delta x^2), \quad (2.42)$$

and hence that

$$[\partial^3_x \theta]_x^{t} = \frac{\theta|_x^{t+2\Delta x} - 2\theta|_x^{t+\Delta x} + 2\theta|_x^{t-\Delta x} - \theta|_x^{t-2\Delta x}}{2\Delta x^3}, \quad (2.43)$$

represents a second order FDA of $\partial^3_x \theta$. Since the FDA (2.43) is centered it comes at no surprise that (2.43) is valid to second order. Please consult exercise 3 on page 27 to derive a second order FDA to $\partial^4_x \theta$ using Taylor series expansions.

### 2.8 Truncation errors

As alluded to the main difference between the one-sided and centered FDAs is the order of the terms neglected when making the approximation from the Taylor series expansion. While we neglected terms of $O(\Delta x^2)$ when using the centered FDA, the terms we neglected when using the one-sided approximation was $O(\Delta x)$. Thus the centered FDA is more accurate than the one-sided FDA. While the centered FDA has an error of second order, the one-sided FDA has an error of first order. Since the error is a direct consequence of truncating the Taylor series expansion, we often refer to this error as the truncation error in that the series is truncated when making the FDA. The order of the truncation error is therefore a measure of the accuracy of the scheme we have constructed.

As shown in Section 10.1 we may also use the Taylor series expansion to construct FDAs that are truncated to even higher orders, e.g., to $O(\Delta x^n)$ where $n \geq 3$. Such FDAs are thus even more accurate and are therefore referred to as higher order schemes or higher order FDAs. We note from (10.8) that when constructing such approximations we have to include points that are distances $2\Delta x$ away from the point $x$. Incidentally we also did this when deriving a centered FDA for $\partial^3_x \theta$ in (2.43). Although we desire our approximations to be as accurate as possible we emphasize that higher order schemes have other potential complications associated with troubles at boundaries, higher order computational modes in space, and a more stringent instability criteria.

Finally, we underscore that it is good practice to ensure that all the FDAs we make use of to approximate the various terms in our governing equations have the same truncation error in space and/or time, but not necessarily to the same order in both time and space. For instance consider a one dimensional wave propagating in a direction forming an angle to the $x$ and $y$ directions.
The only way to ensure that the numerical solution then has the same accuracy regardless of the propagation direction of the wave is to use FDAs that have the same accuracy along all spatial directions or axes.

2.9 Notations

When solving a PDE using numerical methods, and in particular finite difference methods, it is common to define a grid or mesh which covers the domain over which the solution is to be found. As an example let us consider a two-dimensional spatial problem for which we seek a solution to the Laplace equation (2.8) within a quadratic domain where \( x, y \) both starts at 0 and ends at \( L \). We start by covering the domain by a quadratic mesh as displayed in Figure 2.1. We keep track of the grid points in the mesh by counting along the \( x \)-axis and the \( y \)-axis, respectively. Let us furthermore assume that there are \( J \) points along the \( x \)-axis and \( K \) points along the \( y \)-axis. To count the points we use dummy indices for instance \( j \) along the \( x \)-axis and \( k \) along the \( y \)-axis. The point \( x = 0 \) along the \( x \)-axis is then associated with \( j = 1 \), while the point \( x = L \) along the \( x \)-axis is associated with \( j = J \). Similarly we associate the point \( y = 0 \) with \( k = 1 \) and the point \( y = L \) with \( k = K \). The \( j \)th point along the \( x \)-axis is then \( x = x_j \) where the subscript refers to the value for \( x \) at the \( j \)th point along the \( x \)-axis. Similarly we let \( y = y_k \) be associated with the \( k \)th point along the \( y \)-axis. The coordinates of the grid junctions are then given by \( x_j, y_k \).

Let us denote the distance between two adjacent points along the \( x \)-axis by \( \Delta x \) and the distance between two adjacent points along the \( y \)-axis by \( \Delta y \). Then the \( j \)th point along the \( x \)-axis is denoted

\[
x_j = (j - 1)\Delta x,
\]

while the \( k \)th point along the \( y \)-axis is denoted

\[
y_k = (k - 1)\Delta y.
\]

We note in particular that \( x_1 = y_1 = 0 \) and that \( x_J = y_K = L \). We also notice for later convenience that the latter gives

\[
\Delta x = L/(J - 1), \quad \Delta y = L/(K - 1),
\]

respectively\(^4\). It is also common to use the notation \( \theta_{jk} \) to denote the value of the variable \( \theta(x, y) \) at the grid point \( x_j, y_k \). Thus

\[
\theta_{jk} = \theta(x_j, y_k) = \theta[(j - 1)\Delta x, (k - 1)\Delta y].
\]

Furthermore follows that

\[
\theta_{jk} = \theta(x_j, y_k) = \theta[(j - 1)\Delta x, (k - 1)\Delta y]
\]

\(^4\)FORTRAN 90/95 allows us to use \( j = 0 \) and \( k = 0 \) as dummy counters. Under these circumstances \( x_j = j\Delta x \) and \( y_k = k\Delta y \). Thus \( x_0 = y_0 = 0 \) while \( x_J = y_K = L \) as before. Under this circumstance \( \Delta x = L/J \), and \( \Delta y = L/K \)
Figure 2.1: Displayed is a commonly used grid when employing numerical methods to solve PDEs. The points in the $x, y$ directions are incremented by $\Delta x, \Delta y$, respectively, so that there are a total of $J$ points along the $x$-axis and $K$ points along the $y$-axis. The points are counted by using the dummy counters $j, k$.

$$\theta_{j-1k} = \theta(x_{j-1}, y_k) = \theta(j-2)\Delta x, (k-1)\Delta y]$$

(2.49)

$$\theta_{jk+1} = \theta(x_j, y_{k+1}) = \theta[(j-1)\Delta x, k\Delta y]$$

(2.50)

To discriminate between spatial and temporal variables we hereafter use a superscript when counting time, which is common practice. Thus

$$t^n = n\Delta t, \quad n = 0, 1, 2, \cdots$$

(2.51)

---

The apparent inconsistency in starting the time counter at $n = 0$ and the space counter at $j = 1$ is historical.
when counting time, where $\Delta t$ is the time step and $n$ is the time counter. Thus the variable $\theta(x, t)$ at the point $x_j, t^n$ in space and time is written

$$\theta^n_j = \theta^n_{x_j} = \theta(x_j, t^n) = \theta[(j - 1)\Delta x, n\Delta t] \quad (2.52)$$

We note that when the variable is four-dimensional the notation we use is

$$\theta^n_{jkl} = \theta^n(x_j, y_k, z_l, t^n) \quad (2.53)$$

where $z_l = (l - 1)\Delta z$.

As an example let us consider the Taylor series expansions (2.25) and (2.27). Using the preceding notation we find

$$\theta^n_{j+1} = \theta^n_j + \partial_x \theta^n_j \Delta x + \frac{1}{2} \partial_x^2 \theta^n_j \Delta x^2 + \frac{1}{6} \partial_x^3 \theta^n_j \Delta x^3 + O(\Delta x^4) \quad (2.54)$$

$$\theta^n_{j-1} = \theta^n_j - \partial_x \theta^n_j \Delta x + \frac{1}{2} \partial_x^2 \theta^n_j \Delta x^2 - \frac{1}{6} \partial_x^3 \theta^n_j \Delta x^3 + O(\Delta x^4) \quad (2.55)$$

and hence that the forward in space FDA to the first derivative is written

$$[\partial_x \theta]^n_j = \frac{\theta^n_{j+1} - \theta^n_j}{\Delta x}, \quad (2.56)$$

while the second order, centered approximation is written

$$[\partial_x \theta]^n_j = \frac{\theta^n_{j+1} - 2\theta^n_j + \theta^n_{j-1}}{2\Delta x}. \quad (2.57)$$

Similarly follows that the second order, centered FDA to the second derivative with the above notation is written

$$[\partial_x^2 \theta]^n_j = \frac{\theta^n_{j+1} - 2\theta^n_j + \theta^n_{j-1}}{\Delta x^2}. \quad (2.58)$$

Finally we remark that the increments $\Delta x, \Delta y, \Delta z$ and $\Delta t$ do not have to be constant, but may be allowed to vary in space and even in time. If the increments vary in space only we refer to the grid as an unstructured mesh. If the increments vary in both time and space we refer to the grid as an adaptive unstructured mesh.

### 2.10 Orthogonal functions

Note that when using finite difference techniques for time dependent or evolutionary problems, we only consider grid-point values of the dependent variables; no assumption is made about how the variables behave between grid points. An alternative approach is to expand the dependent

---

To save space on the computer we never store all time levels. Hence we never make use of do loops when stepping forward in time. How many time steps we need to store depend on the time stepping scheme we use. If for instance the scheme is a two time level scheme we store only two time levels (and sometimes only one level).
variables in terms of a finite series of smooth orthogonal functions. The problem is then reduced to solving a set of ordinary differential equations which determine the behavior in time of the expansion coefficients.

As an example consider the general linear one-dimensional time dependent problem

\[ \partial_t \phi = \mathcal{H}[\phi] \quad \text{for} \quad x \in (-L, L) \quad \text{and} \quad t > 0 \]  

(2.59)

where \( \phi = \phi(x,t) \) is a good function as defined in Section 2.6 and \( \mathcal{H} \) is a linear differential operator in \( x \). Note that to solve (2.59) we have to specify suitable boundary conditions at \( x = \pm L \) and initial condition at \( t = 0 \). Here we will simply assume that the condition at \( x = \pm L \) is that \( \phi \) is cyclic and that the initial value is known. Since \( \phi \) is a good function it may be expanded in terms of an infinite set of orthogonal functions \( e_n(x) \), where \( n = 1, 2, 3, \ldots \). Thus

\[ \phi = \sum_{n=-\infty}^{\infty} \varphi_n(t) e_n(x), \]  

(2.60)

where \( \varphi_n(t) \) are the time dependent expansion coefficients\(^7\). Without loss of generality we may assume that the expansion functions \( e_n(x) \) are orthonormal so that

\[ \int_{-L}^{L} e_n(x)e_m^*(x)dx = \begin{cases} 1 & ; \quad n = m \\ 0 & ; \quad n \neq m \end{cases}. \]  

(2.61)

where \( e_n^*(x) \) is the complex conjugate of \( e_n(x) \). Consider that we know the expansion functions \( e_n(x) \). It is then the expansion coefficients \( \varphi_n(t) \) whose behavior we want to determine. To this end we first multiply (2.59) by \( e_m^* \), and then integrate over all possible \( x \)-values, to give

\[ \int_{-L}^{L} \partial_t \phi(x,t)e_m^*(x)dx = \int_{-L}^{L} \mathcal{H}[\phi]e_m^*(x)dx. \]  

(2.62)

The left-hand side is further developed by use of (2.60) and (2.61) to give

\[ \int_{-L}^{L} \partial_t \phi e_m dx = \int_{-L}^{L} \left( \sum_n \partial_t \varphi_n e_n \right) e_m dx = \sum_n \partial_t \varphi_n \int_{-L}^{L} e_n e_m^* dx = \sum_n \partial_t \varphi_n. \]  

(2.63)

Since the operator \( \mathcal{H} \) only operates on \( x \) follows in addition that

\[ \mathcal{H}[\phi] = \sum_m \varphi_m \mathcal{H}[e_m]. \]  

(2.64)

Using these results we get

\[ \partial_t \varphi_n = \sum_m \varphi_m \int_{-L}^{L} \mathcal{H}[e_m]e_n^* dx \quad ; \quad \forall m. \]  

(2.65)

\(^6\)Note that the expansion functions \( e_n(x) \) are in general complex functions, e.g., \( e_n(x) = e^{i\alpha_n x} \) where \( \alpha_n \) is the wavenumber associated with the \( n^{th} \) eigenvalue.

\(^7\)In fact this is a general method commonly used to separate variables when analytically solving differential equations involving more than one independent variable.
We thus have a set of coupled, ordinary differential equations for the time rate of change for the expansion coefficients $\varphi_n$.

It is now interesting to consider how our choice of expansion functions can greatly simplify the problem.

1. If the expansion functions are eigenfunctions of $\mathcal{H}$, we have $\mathcal{H}[e_m] = \lambda_m e_m$ where $\lambda_m$ are the eigenvalues. Equation (2.65) then becomes

$$\partial_t \varphi_m = \lambda_m \varphi_m \quad \forall m$$

and becomes decoupled.

2. If the original equation is

$$\mathcal{G} [\partial_t \varphi] = \mathcal{H}[\varphi]$$

where $\mathcal{G}$ is a linear operator, then our problem is simplified by using expansion functions that are eigenfunctions of $\mathcal{G}$ with eigenvalues $\lambda_n$. We then have,

$$\lambda_n \partial_t \varphi_n = \sum_{m} \varphi_m \int_{-L}^{L} \mathcal{H}[e_m] e_n^* dx. \quad (2.68)$$

### 2.11 Fourier series

A much used orthogonal set of expansion functions are the trigonometric functions $e^{i\alpha_n x}$ where $\alpha_n$ are an infinite number of discrete wavenumbers\(^8\). Thus any good function $\phi(x, t)$ may be written

$$\phi(x, t) = \sum_{n=-\infty}^{\infty} \varphi_n(t; \alpha_n) e^{i\alpha_n x}. \quad (2.69)$$

The series (2.69) is called a **Fourier series** and the expression

$$\varphi_n(t) e^{i\alpha_n x} \quad (2.70)$$

is called a **Fourier component**. We note that the complex conjugate to the expansion functions are $e^{-i\alpha_n x}$, and hence the Fourier series may be written

$$\phi(x, t) = \phi_0 + \sum_{n=1}^{\infty} \varphi_n(t; \alpha_n) e^{i\alpha_n x}. \quad (2.71)$$

It is important to realize that the subscript $n$ attached to the expansion coefficients implies that they are different for each wavenumber, and hence depends on the wavenumber $\alpha_n$ as well as time.

---

\(^8\)In the above problem with cyclic boundary conditions at $x = \pm L$ the wavenumbers are $\alpha_n = n\pi/L$. 

26
2.12 Fourier transforms

Finally, let us assume that the function $\phi$ depends on $x$ only, and is a good function. Under these circumstances we may define a function $\tilde{\phi}$ such that

$$\tilde{\phi}(\alpha) = \int_{-\infty}^{\infty} \phi(x) e^{-i\alpha x} \, dx. \quad (2.72)$$

We observe that $\tilde{\phi}$ is a complex function consisting of a real as well as an imaginary part. As is common we refer to $\tilde{\phi}$ as the Fourier transform of the real function $\phi$. Furthermore we notice that the $\tilde{\phi}$ is a continuous function of the wavenumber $\alpha \in [-\infty, +\infty]$. Hence if we know the Fourier transform the original real function $\phi$ is retrieved from the real part of the inverse Fourier transform defined by

$$\phi(x) = \frac{1}{2\pi} \int_{-\infty}^{\infty} \tilde{\phi}(\alpha) e^{i\alpha x} \, d\alpha. \quad (2.73)$$

We observe that the “expansion coefficient” $\tilde{\phi}$ now is a continuous function of the wavenumber $\alpha \in [-\infty, +\infty]$, and that the summation in (2.69) is replaced by an integral. We may also plot the Fourier transform $\tilde{\phi}$ as a function of $\alpha$. In that case the space spanned by $\tilde{\phi}$ and $\alpha$ is called the Fourier space and its distribution the Fourier spectrum.

As revealed by (2.73) the Fourier transform (2.72) is simply the amplitude associated with the wave of wavenumber (or wavelength) $\alpha$. The amplitude in a sense reveals how much “energy” is associated with each wavelength. Thus if we plot the Fourier transform in Fourier space the distribution reveals how much energy is contained in the various wavelengths. The waves with wavelengths having the highest amplitudes are also the wavelengths that contain the highest energy content. Knowing the Fourier transform thus reveals information about the wavelengths that dominates the motion.

We use the information in Fourier spectrum to construct the grid, particularly the size of the spatial increments to choose (cf. Figure 2.1). If we intend to resolve the dominant portion of the motion we must choose the increments so that we have enough points per wavelength to resolve it. Ideally we should have 10 points per wavelength. As a minimum we must require that the size of the increments are such that we have 4 points per wavelength. Finally we emphasize that our solutions are real functions. Hence, if we know the Fourier transform we find the solution to our problem by first finding the inverse Fourier transform (2.73) and then extracting its real part.

Exercises

1. Show that both the Helmholtz and the Laplace equations are elliptic in $x$ and $y$.

2. Show that the diffusion equation is parabolic in $t, x$ and $t, y$, but elliptic in $x, y$.

3. Show by use of Taylor series expansions that a possible centered FDA of $\partial^4_x \theta(x)$ is

$$[\partial^4_x \theta]_j = \frac{\theta_{j+2} - 4\theta_{j+1} + 6\theta_j - 4\theta_{j-1} + \theta_{j-2}}{\Delta x^4}, \quad (2.74)$$
and that the truncation error is $O(\Delta x^2)$. Note that we have to use points that are distances $2\Delta x$ away from the point $x_j$ itself. This is common when deriving centered FDAs to higher order derivatives (cf. 2.43).

4. Assume that $\theta(x, t)$ and all of its derivatives tend to zero as $x \to \pm \infty$. Show that under these conditions the Fourier transform of $\partial_x \theta(x, t)$ and $\partial_x^2 \theta(x, t)$ are

$$\tilde{\partial_x \theta} = i\alpha \tilde{\theta} \quad \text{and} \quad \tilde{\partial_x^2 \theta} = -\alpha^2 \tilde{\theta},$$

respectively, where the notation $\tilde{\psi}$ denotes the Fourier transform of $\psi$.

5. Show by making use of the results in Exercise 4 that a formal analytic solution to the diffusion equation

$$\partial_t \theta = \kappa \partial_x^2 \theta,$$

where $\theta = \theta(x, t)$, $\kappa$ is a constant, and the boundary conditions are

$$\theta = \begin{cases} 0 & x \to +\infty, -\infty \large; \\ \theta_0 e^{-\alpha^2 (x/a)^2} & t = 0 \end{cases},$$

is

$$\theta = \frac{a \theta_0}{2\sqrt{\pi}} \int_{-\infty}^{\infty} e^{-\alpha^2 (a^2 + \sqrt{2})^2} e^{i\alpha x} d\alpha$$

(2.78)
Chapter 3

TIME MARCHING PROBLEMS

Most of the problems in the atmospheric and oceanographic sciences involve solving a time marching problem. Typically, we know the state of the atmosphere or the ocean at one specific time and want to know what the state is at a later time. Our task in numerical weather prediction (NWP) and numerical ocean weather prediction (NOWP) is then to use the governing equations of Section 1.1 on page 1 to find the state of the sphere in question at some later time, just as postulated by Bjerknes (1904) (cf. his statement quoted on page iii). Such problems are known as initial value problems in mathematics.

A particular balance inherent in our governing equations, as displayed in (1.1) - (1.4), and particularly evident in the tracer equation (1.3), is a balance between the time rate of change of a variable in response to advective and diffusive fluxes. As the name indicates the problem is a combination of two different physical processes. The first is associated with advection, in which case the underlying PDE is hyperbolic (Section 2.4). The second is associated with diffusion, in which case the underlying PDE is parabolic (Section 2.3).

Another important balance is the geostrophic balance (1.39) on page 9. This possibility of a balance between the pressure force and the Coriolis acceleration makes geophysical fluid dynamics stand out compared to non-rotating fluid dynamics. The importance of this balance in atmosphere and ocean dynamics is illustrated through the shallow water equations (1.33) and (1.34). Any deviation from this balance manifests itself through non-zero acceleration terms, e.g., the local time rate of change of the velocity, commonly referred to as ageostrophic terms.

The advection and diffusion problems (and their combination) and the shallow water equations are of fundamental importance in meteorology and oceanography. Knowledge on how to solve these simple equations by numerical means is a “must” for everyone who aspires to become a meteorologist and/or oceanographer. In the three Chapters to follow, that is, Chapters 4 through 6, we therefore give a detailed account of how to solve respectively the diffusion problem, the advection problem and the shallow water equations by use of numerical methods. In Chapter 10 (Section 10.2 on page 146) we also give insight into how to solve the combined advection-diffusion problem.

We maintain that it is of fundamental importance to obtain knowledge on how to treat the various terms in these three fundamental problems numerically correct. At the same time these relatively simple problems conveniently serves the purpose of introducing some of the basic con-
cepts needed to solve atmospheric and oceanographic problems employing numerical methods. Moreover, they serve the purpose of illustrating some of the pitfalls.

Before venturing into details we first highlight some physical properties peculiar to each of the three problems. The motivation is that these important properties must be retained in any numerical solutions, or else the solution must be discarded as being false or incorrect. Thus to check the behavior of the solutions against these fundamental properties is part of what is often referred to as model verification which is the first step in a chain of activities commonly referred to as model quality assurance or model evaluation procedures (Anon, 1991; Lynch and Davies, 1995; Hackett et al., 1995). When coding errors are thus found we refer to the process as debugging which simply means to weed out all errors in the program code.

3.1 The advection-diffusion equation

We now focus on the tracer conservation equation for a Boussinesq fluid. We start by recalling that the tracer equation for a Boussinesq fluid is given by (1.18). Neglecting possible tracer sources ($S_i = 0$) we get

$$\frac{\partial \theta}{\partial t} + \nabla \cdot F = 0,$$

(3.1)

where $\theta$ is any dependent variable (or tracer), for instance potential temperature, and $F$ is a flux vector that includes both the advective flux and the flux due to turbulent mixing. If $\theta$ is the potential temperature then (3.1) is the conservation equation for internal energy or heat content neglecting any source terms.\(^1\)

Note that the flux vector represents physical processes that transfer properties from one location to the next. In the atmosphere and oceans this is basically caused by two distinct and different physical processes. One is advective processes transporting or propagating properties from one place to the next via the motion. The second is turbulent mixing, that is, small scale, inherently chaotic processes that cause properties to be exchanged between two locations without invoking any mean motion. It is therefore useful to separate the flux vector $F$ into two parts, as we did in (1.18), that is,

$$F = F_A + F_D$$

(3.2)

where $F_A$ represents the flux due to advective processes, commonly referred to as the advective flux vector, and $F_D$ represents the fluxes due to turbulent mixing, commonly referred to as the diffusive flux vector.

Since the advective and diffusive flux vectors represent two very contrasting physical processes, they naturally have very different mathematical formulations or parameterizations. The advective flux vector $F_A$ depends on the motion only. The advective flux of the property $\theta$ therefore follows the path of the individual fluid parcels. Thus its parameterization, or mathematical formulation, becomes

$$F_A = v \theta.$$

(3.3)

\(^1\)The total energy of a system consists of the internal energy and the mechanical energy. The internal energy is concerned with the heat content and is an important part of the thermodynamics. In contrast the mechanical energy concerns the motion of the fluid and is thus part of the fluid dynamics.
In contrast the mathematical formulation of the diffusive flux vector is somewhat more complex simply because the turbulent mixing is a complex process. In many respects the turbulent mixing acts to even out disturbances in the atmospheric and oceanic tracer field, and hence have many characteristics similar to processes like diffusion and conduction. In fact this is the reason why we refer to this flux as the diffusive flux vector. Accordingly the most common parameterization of the diffusive flux vector is Fickian diffusion, that is,

\[ F_D = -\kappa \nabla \theta, \quad (3.4) \]

where \( \kappa \) is the diffusion coefficient or conductive capacity\(^2\). Equation (3.4) expresses that the larger the gradient (or difference) the larger the diffusive flux and hence the more effective diffusion is to decrease any differences in the tracer \( \theta \) over small distances.

### 3.2 Diffusion

If we for a moment neglect the advective part of the flux vector the time rate of change of the tracer concentration is balanced by the diffusive flux only, that is.

\[ \partial_t \theta = -\nabla \cdot F_D. \quad (3.5) \]

Using the parameterization (3.4) for the diffusive flux, we get

\[ \partial_t \theta = \kappa \nabla^2 \theta. \quad (3.6) \]

which is a typical parabolic problem (cf. eq. 2.9 on page 14). The resulting equation is called the diffusion equation, and solving it is referred to as solving the diffusion problem.

Recall that one of the important properties of the turbulent mixing is to even out small scale differences in the tracer fields. We thus have to ensure that our parameterization of the diffusive flux vector indeed have this property\(^3\). The noisiness of a field is commonly measured by its variance. A measure of this variance is \( \theta^2 \), that is, the square of the tracer concentration. We may thus investigate whether the noise increases or decreases by analyzing the time rate of change of \( \theta^2 \).

To arrive at an equation for the time rate of change of the variance we thus first multiply (3.5) by the tracer concentration \( \theta \) itself to get

\[ \partial_t \theta^2 = -2\theta \nabla \cdot F_D. \quad (3.7) \]

The left hand side of (3.7) is the time rate of change of the variance. Let us assume that (3.5) and by implication (3.7) are valid within a volume \( V \) bounded by a surface \( \Omega \). Then we get the time rate of change of the aggregated variance by integrating (3.7) over the total volume \( V \), or

\[ \partial_t \left( \int_V \theta^2 dV \right) = -2 \int_{\Omega} \theta F_D \cdot \delta \sigma + 2 \int_V F_D \cdot \nabla \theta dV. \quad (3.8) \]

\(^2\)Its original formulation is due to a Dr. Adolf Eugen Fick who in 1855 formulated the parameterization (3.4).

\(^3\)In passing we note that if this is true the parameterization also acts to even out any noise created by our choice of numerical methods, if any, when solving the equation numerically.
Here the vector $\delta \sigma = n \delta \sigma$ where $n$ is a unit vector directed along the outward normal to the surface $\Omega$ and $\delta \sigma$ is an infinitesimal surface element. To derive (3.8) we also used the Gauss theorem. At the boundary $\Omega$ we must specify a boundary condition. We simply assume that the condition is either a Dirichlet or a Neuman condition. In the former case we let $\theta = 0$ at $\Omega$, while in the latter case we let $n \cdot F_D = 0$ at the surface $\Omega$. In either case we observe that the first term on the right-hand side of (3.8) is zero. Hence (3.8) reduces to

$$\partial_t \int_V \theta^2 dV = 2 \int_V F_D \cdot \nabla \theta dV.$$  \hspace{1cm} (3.9)

If

$$F_D \cdot \nabla \theta \leq 0$$  \hspace{1cm} (3.10)

the right-hand side of (3.9) is negative. Hence we get

$$\partial_t \int_V \theta^2 dV \leq 0,$$  \hspace{1cm} (3.11)

which shows that as long as (3.10) is satisfied then the diffusion term acts to even out any noise in the $\theta$ field. We notice that (3.10) is always satisfied as long as the diffusive flux vector $F_D$ is directed opposite to $\nabla \theta$. Under these circumstances we refer to the parameterization of the diffusive flux vector as being *down the gradient*. We recall from (3.4) that in the case of Fickian diffusion $F_D = -\kappa \nabla \theta$. Hence

$$F_D \cdot \nabla \theta = -\kappa (\nabla \theta)^2 \leq 0,$$  \hspace{1cm} (3.12)

which reveals that Fickian diffusion is indeed down the gradient and thus always tends to even out any noise in our solution.

Thus we conclude that the diffusive flux vector, when properly parameterized, always acts to even out the variance in any tracer field. In this regard it is worthwhile to underscore that most problems in oceanography and meteorology are non-linear. While there is no transfer of energy from one wavelength to the next in a linear system, this is not true for a non-linear system. In such systems energy input on long wavelengths (small wave numbers) is always in the end transferred to progressively shorter wavelengths (high wave numbers). This fact was described elegantly in the following rhyme credited to G. I. Taylor\(^4\):

> “Big whirls have smaller whirls that feed on their velocity, and little whirls have lesser whirls, and so on to viscosity .... in the molecular sense.”

However, when making the finite difference approximations to our PDEs the wavelengths that we resolve is limited by the specified spatial increments, say $2\Delta x$, often referred to as

\(^4\)Geoffrey Ingram Taylor (1886 - 1975) made fundamental contributions to turbulence, championing the need for developing a statistical theory, and performing the first measurements of the effective diffusivity and viscosity of the atmosphere. He is commonly remembered as the namesake for several basic fluid flow instabilities (Taylor - Couette, Rayleigh - Taylor, and Saffman - Taylor).
the Nyquist wavelength (or frequency in the time domain). Thus as the energy is cascading downwards toward shorter wavelengths we must, in our numerical solutions, mimic this process across the Nyquist wavelength to wavelengths which are not resolved by our grid. Since diffusion has the property of damping differences it is one tool at hand that may prove useful to handle what is known as non-linear instability (cf. Section 10.3 on page 149).

### 3.3 Advection

Next let us for a moment neglect the diffusive part of the flux vector. Under this circumstances the flux vector reduces to the advective flux vector only, and hence any time rate of change in the tracer $\theta$ is caused by advection, that is,

$$\partial_t \theta + \nabla \cdot \mathbf{F}_A = \partial_t \theta + \nabla \cdot (\mathbf{v} \theta) = 0.$$  \hspace{1cm} (3.13)

We recognize this balance as the “wave equation” (cf. eq. 2.13 on page 15). Solving this equation is consequently often referred to as solving the advection problem.

As we alluded to in Chapter 3.1 the advection problem is a balance between the first and second term on the left hand-side of (3.1), that is,

$$\partial_t \theta + \nabla \cdot \mathbf{F}_A = 0,$$  \hspace{1cm} (3.14)

where $\theta$ is any tracer variable and $\mathbf{F}_A$ is the advective flux vector. As for the diffusion problem we are looking for solutions within a limited volume $V$ in space bounded by the surface $\Omega$, and for all times $t \in [0, \infty]$. On the surface $\Omega$ the equations are replaced by the boundary conditions, while the initial condition replaces the equations at time $t = 0$.

Let the advective flux be parameterized by the common parameterization $\mathbf{F}_A = \mathbf{v} \theta$, and let the boundary condition at the surface $\Omega$ be such that $\mathbf{F}_A \cdot \delta \sigma = 0$. Then by performing the same operation as in Section 3.2 we find that the total variance becomes

$$\partial_t \int_V \theta^2 dV = 2 \int_V \mathbf{F}_A \cdot \nabla \theta dV = \int_V \mathbf{v} \cdot \nabla \theta^2 dV = - \int_V \theta^2 \nabla \cdot \mathbf{v} dV.$$  \hspace{1cm} (3.15)

Thus the total variance may increase or decrease depending on the sign of the velocity divergence. If the sum of the divergence is positive then the variance will decrease, while if it is negative then the variance will increase. The case $\nabla \cdot \mathbf{v} = 0$ is special. In this case the right hand-side of (3.15) is zero and hence any disturbances creating a variance in $\theta$ will just prevail, that is, the total variance is conserved.

As mentioned in Section 1.4 the Boussinesq ocean is to a good approximation divergence free due to its incompressibility (see also Gill, 1982, side 85). Thus in the ocean the advection process does not lead to any decrease or increase in the property being advected. Hence any disturbance generated in a limited domain may be advected to other locations undisturbed. This

\[5\] In practice we have to limit the computation to a finite time span

\[6\] This is achieved by assuming $\mathbf{v} = 0$ or $\mathbf{v} \cdot \delta \sigma = 0$, that is, no flow across the boundary.
is not true for the atmosphere since the atmosphere is highly compressible. Thus in limited areas where the divergence is positive ($\nabla \cdot \mathbf{v} > 0$), that is, the individual fluid parcels are drawn apart, any disturbances in the total tracer variance are smoothed. In contrast the disturbances tend to increase in areas where $\nabla \cdot \mathbf{v} < 0$.

Finally we emphasize that the properties outlined above regarding the advection are important to retain when solving the advective problem by numerical means. In particular we stress that when the fluid is divergence free, like the ocean, then the total variance should be conserved. We also note that this is in stark contrast to the diffusion problem where all down the gradient diffusive fluxes give a decrease in the total tracer variance.

### 3.4 Shallow water equations

As alluded to the third and final fundamental balance equation important in atmosphere and ocean dynamics are the shallow water equations as displayed in (1.23) through (1.25) on page 7. Neglecting the forcing terms on the right-hand side of (1.25) we get

$$\nabla_H \cdot \mathbf{u} + \partial_z w = 0 \quad (3.16)$$

$$\partial_t \mathbf{u} + \nabla_H \cdot (\mathbf{u} \mathbf{u}) + \partial_z (w \mathbf{u}) + f \mathbf{k} \times \mathbf{u} = -g \nabla_H \eta \quad (3.17)$$

where we have used (1.26) to substitute for the pressure\(^7\).

Again we will, as we did for the advection and diffusion equations, investigate the properties of the time rate of change of the variance of the motion integrated over a fixed volume $V$, that is,

$$\partial_t \left( \int_V \mathbf{u}^2 dV \right) \quad (3.18)$$

We note that $e_K = \frac{1}{2} \mathbf{u}^2$ is the kinetic energy per unit mass. Thus

$$\int_V \mathbf{u}^2 dV = \int_V 2e_K dV = 2E_K \quad (3.19)$$

where $E_K$ is the total kinetic energy\(^8\). The total variance of the motion is therefore twice the total kinetic energy. We note that the kinetic energy is a positive definite quantity, that is, $E_K \geq 0$.

To find an equation for the time rate of change of the kinetic energy we start by multiplying (3.17) by $\mathbf{u}$ and get

$$\partial_t e_K + \nabla_H \cdot (e_K \mathbf{u}) + \partial_z (e_K w) = -g \mathbf{u} \cdot \nabla_H \eta \quad (3.20)$$

where we have made use of the fact that

$$\mathbf{u} \cdot [\nabla_H \cdot (\mathbf{u} \mathbf{u}) + \partial_z (w \mathbf{u})] = \nabla_H \cdot (e_K \mathbf{u}) + \partial_z (e_K w), \quad (3.21)$$

\(^7\)Note that we use $\mathbf{u}$ to denote the horizontal component of the velocity, that is, $\mathbf{v} = \mathbf{u} + w \mathbf{k}$

\(^8\)To be precise $E_K$ is the total kinetic energy per unit density, but since we have assumed a uniform density, that is, $\rho_0 = \text{constant}$, it is common to refer to $E_K$ as the total kinetic energy.
which follows by use of the continuity equation (3.16). Finally we note that the contribution from the Coriolis term vanishes since \( \mathbf{u} \cdot (\mathbf{k} \times \mathbf{u}) = \mathbf{k} \cdot (\mathbf{u} \times \mathbf{u}) = 0 \). Next we integrate (3.20) over the fixed volume \( V \) to get

\[
\partial_t E_K = C, \tag{3.22}
\]

where

\[
C = -\int_V g\mathbf{u} \cdot \nabla H \eta dV. \tag{3.23}
\]

We observe that under the assumptions that there is no forcing terms\(^9\), the time rate of change of the variance of the motion, or the kinetic energy per unit mass is proportional to \( C \), a quantity yet to be interpreted.

To interpret \( C \) we first split the volume integral in (3.23) into an area integral and a vertical integration to get

\[
C = -g \int_A \left( \int_{-H}^{H} \mathbf{u} \cdot \nabla H \eta \eta dV \right) dA, \tag{3.24}
\]

where \( A \) is the projected area of the volume \( V \) onto a horizontal surface. We now expand the inner vertical integral. Noting that \( \eta \) and \( \nabla H \eta \) are independent of depth/height we first get

\[
\int_{-H}^{H} \mathbf{u} \cdot \nabla H \eta \eta dV = -\nabla H \eta \cdot \left( \int_{-H}^{H} \eta \mathbf{u} dV \right) = \nabla H \cdot \left( \int_{-H}^{H} \eta \mathbf{u} dV \right) \tag{3.25}
\]

Integrating the continuity equation (3.16) from bottom to top yields

\[
\nabla H \cdot \left( \int_{-H}^{H} \eta \mathbf{u} dV \right) = -\partial_t \eta, \tag{3.26}
\]

where we have used the kinematic boundary condition (1.7) and (1.9) in Section 1.2 on page 4. Thus (3.25) becomes

\[
\int_{-H}^{H} \mathbf{u} \cdot \nabla H \eta \eta dV = \eta \partial_t \eta + \nabla H \cdot \left( \int_{-H}^{H} \eta \mathbf{u} dV \right). \tag{3.27}
\]

Substituting (3.27) into (3.24), noting that the second term on the right-hand side of (3.27) is a flux term that vanishes upon integration over the area \( A \), we get

\[
C = -\partial_t E_\Phi \quad \text{or} \quad \partial_t E_\Phi = -C, \tag{3.28}
\]

where \( E_\Phi \) is

\[
E_\Phi = \int_V \phi dV - \int_{V_0} \phi dV = \int_A \left( \int_{-H}^{H} \phi dV - \int_{-H}^{H} \phi dV \right) dA = \int_A \frac{1}{2} g \eta^2 dA, \tag{3.29}
\]

\(^9\)The forcing terms leads only to external source or sink terms, that is, irreversible energy conversion terms irrelevant for the present presentation.
that is, $E_\Phi$ is the potential energy (per unit density) from which is subtracted the initial\textsuperscript{10} potential energy\textsuperscript{11}. We note that $E_\Phi$ as $E_K$ is a positive definite quantity.

Thus $C$ is a term that is able to convert potential energy into kinetic energy or vice versa. Thus it constitutes a \textit{reversible} transfer of energy between the two mechanical energy forms. That it is a reversible transfer of energy is perhaps best illustrated if we substitute (3.28) into (3.22) to get

$$\partial_t (E_K + E_\Phi) = 0.$$ \hfill (3.30)

Since $E_K + E_\Phi$ is the total (mechanical) energy, (3.30) shows that the total energy is conserved. Thus if the total kinetic energy $E_K$ experience an increase there is a similar decrease in the available gravitational potential energy $E_\Phi$. We also notice that this gain in kinetic energy may later be lost to potential energy through the term $C$.

We therefore conclude that under the assumptions of no external exchange of energy (a consequence of neglecting the forcing terms in eq. 3.17), the time rate of change of the variance of the motion is proportional to the conversion of kinetic energy to potential energy. For any numerical scheme to be trustworthy this latter property should be retained when constructing our numerical schemes (Arakawa and Lamb, 1977).

\textsuperscript{10}The initial state is defined as one at rest, and in static equilibrium, that is, $u(x, y, z, 0) = 0$ and $\eta(x, y, 0) = 0$.

\textsuperscript{11}$E_\Phi$ is commonly referred to as the available gravitational potential energy per unit mass first introduced by Lorenz (1955) (see also Røed, 1997, 1999). The gravitational potential energy per unit mass is $\int_V gzdV$. However, as noted already by Lorenz (1955) the gravitational potential energy contains an enormous amount of energy not available to release into kinetic energy. Hence he introduced the concept of \textit{available potential energy}. Here we use the notation \textit{available gravitational potential energy} since we only concern ourselves with the potential energy in the gravity field.
Chapter 4

THE DIFFUSION PROBLEM

In this chapter we present finite difference methods whereby the diffusion equation can be solved by numerical means. We first consider the diffusion equation in its simplest form. Hence we will study a one-dimensional diffusion process in space. Additionally we assume that the diffusive flux can be parameterized as a down the gradient diffusion and that the diffusion (mixing) coefficient is uniform in time and space. Under these circumstances the diffusion equation (3.5) in its continuous form reads,

\[ \frac{\partial \theta}{\partial t} = \kappa \frac{\partial^2 \theta}{\partial x^2}, \tag{4.1} \]

where \( \theta \) can be any variable, e.g., potential temperature, density, velocity components, and \( \kappa \) is the diffusion coefficient.

As alluded to in Section 2.3 on page 14, we note that the partial differential equation (PDE) (4.1) is parabolic in nature. Thus the physical characteristic of the problem is that properties are transferred from one location to the next by conduction. We emphasize that this is very different from hyperbolic type problems, e.g., advection problems. The diffusion process acts simply to even out differences without dissipation. If we for instance start with a very narrow tracer distribution (cf. Figure 5.5 on page 75) diffusion acts to transfer these high values to adjacent locations at the expense of the peak value as time is marching on. Thus as time passes the peak is diminished while the values at adjacent locations increases. If we allow the diffusion process to go on forever within an infinite domain the tracer values becomes infinitely small, but will cover an infinite domain. In summary, the diffusion process transfers properties from one location to the next by conduction that acts to diminish differences in a distribution so that the end result is a much smoother field.

An obvious example of a diffusion process is heat conduction (or turbulent mixing) in both atmospheres and oceans. Then \( \theta \) is the potential temperature and \( x \) any of the independent variables in space. Another classic atmosphere-ocean example of a diffusion problem is the so called Ekman problem, which in the atmosphere explains how the velocity is reduced in the planetary boundary layer due to friction at the surface. In the ocean the Ekman problem explains how the momentum due to surface traction is transferred downwards in the water column.
4.1 Finite difference equation

Our concern is to find a numerical solution to (4.1) for all times for a given computational domain or area, e.g., \( x \in [0, L] \). At the two space boundaries and for the initial time \( t = 0 \) (4.1) is replaced by the respective boundary conditions.

Let us for instance consider lateral heat conduction in the atmosphere and ocean. The coordinate axis \( x \) is then one of the horizontal (or lateral) coordinates, while the dependent variable \( \theta \) for instance describes the deviation of the potential temperature away from a given mean temperature profile. Let us furthermore assume that we know the anomalous distribution in space at time \( t = 0 \). Thus \( \theta(x, 0) = \theta_0(x) \) where \( \theta_0(x) \) is a known function of the lateral coordinate \( x \). Our task is to find, by numerically solving (4.1), how the anomaly evolves in time between the two positions \( x = 0, L \). By considering that \( \theta = 0 \) at \( x = 0 \) and \( x = L \) we imply that the boundary condition is a Dirichlet condition, that is, that the value of the dependent variable is fixed for all times. We also assume that the initial anomaly is different from the trivial solution \( \theta_0(x) = 0; \forall x \), that is, there exists at least one position in space where \( \theta_0(x) \neq 0 \).

To find a numerical solution to (4.1) we follow the notation in Section 2.9. Thus we first divide the intervals \( x \in (0, L) \) and \( t \in (0, T) \), where \( T \) is some finite time, into respectively \( J \) and \( N \) sections of width \( \Delta x \) and \( \Delta t \), respectively. They then form a grid whose grid points are located at \((x_j, t^n)\) where \( x_j = (j - 1)\Delta x \) and \( t^n = n\Delta t \). Here \( j \) and \( n \) are counters, counting the number of steps needed to reach the grid point \((x_j, t^n)\). Thus \( j \in [1, J + 1] \) and \( n \in [0, N] \) where \( x_{j+1} = L \) and \( t^N = T \) (see Figure 4.1 on page 39).

Next we must define a finite difference approximation to the derivatives \( \partial_t \theta \) and \( \partial_x^2 \theta \) at the grid points. Using a forward in time approximation to express \( \partial_t \theta \bigg|_j^n \) and a centered in space approximation to express \( \partial_x^2 \theta \bigg|_j^n \) it follows from Section 2.6 that

\[
[\partial_t \theta]_j^n = \frac{\theta_{j+1}^n - \theta_j^n}{\Delta t}, \quad [\partial_x^2 \theta]_j^n = \frac{\theta_{j+1}^n - 2\theta_j^n + \theta_{j-1}^n}{\Delta x^2}.
\] (4.2)

By substituting the expressions (4.2) into (4.1) we get

\[
\frac{\theta_{j+1}^n - \theta_j^n}{\Delta t} = \kappa \frac{\theta_{j-1}^n - 2\theta_j^n + \theta_{j+1}^n}{\Delta x^2} \quad ; \quad \left\{ \begin{array}{ll}
\quad j = 2(1)J \\
\quad n = 0(1)N
\end{array} \right.
\] (4.3)

Solving with respect to \( \theta_{j+1}^n \) we finally get

\[
\theta_{j+1}^n = \theta_j^n + \frac{\kappa \Delta t}{\Delta x^2} \left( \theta_{j-1}^n - 2\theta_j^n + \theta_{j+1}^n \right) \quad ; \quad \left\{ \begin{array}{ll}
\quad j = 2(1)J \\
\quad n = 0(1)N
\end{array} \right.
\] (4.4)

Note that (4.3) and (4.4) are valid for \( j = 2(1)J \) and for \( n = 0(1)N \) only. At the boundaries \( j = 1 \) \((x = x_1 = 0)\) and \( j = J + 1 \) \((x = x_{J+1} = L)\) and for \( n = 0 \) \((t = t^0 = 0)\) the boundary and initial conditions prevail. In numerical language they are

\[
\theta_j^0 = \theta_{0j} \quad ; \quad \forall j \quad \text{and} \quad \theta_1^n = \theta_{J+1}^n = 0 \quad ; \quad \forall n,
\] (4.5)
4.1 Finite difference form

\[ \theta_j^n = \frac{\kappa \Delta t}{\Delta x^2} (\theta_j^{n-1} - 2\theta_j^n + \theta_j^{n+1}) \quad ; \quad j = 2(1)J. \]  

(4.6)

Thus for the first “wet” point \( j = 2 \)

\[ \theta_2^1 = \theta_2^0 + \frac{\kappa \Delta t}{\Delta x^2} (\theta_1^0 - 2\theta_2^0 + \theta_3^0), \]  

(4.7)

in which all the \( \theta_j^0 \)'s on the right-hand side of (4.7) are known from the boundary and/or initial conditions (4.5). We may then proceed to evaluate \( \theta_3^1, \theta_4^1, \cdots \) up to and including \( \theta_J^1 \). For the last wet point \( j = J \) we get in particular

\[ \theta_J^1 = \theta_J^0 + \frac{\kappa \Delta t}{\Delta x^2} (\theta_{J-1}^0 - 2\theta_J^0 + \theta_{J+1}^0). \]  

(4.8)

We note that again all the \( \theta_j^0 \)'s on the right-hand side are known from the initial and/or boundary conditions (4.5). This procedure thus provides values for the potential temperature anomaly at
all the interior grid points for time level \( n \) (or at time \( t = \Delta t \)), that is, \( \theta^n_j \). Note that at the boundaries \( j = 1 \) and \( J + 1 \) \( \theta \) is known from the boundary condition (4.5). This reflects the well known property of differential equations, whether they are PDEs of ordinary (ODEs), namely that they are valid only in the interior of a domain. At the boundaries (whether in time or space) the equations are replaced by the boundary condition. Thus (4.6) together with the boundary conditions gives us \( \theta \) for all \( j = 1(1)J \) at time level \( n = 1 \).

We may then proceed to compute \( \theta \) at time level \( n = 2 \). We do this by substitution of \( n = 1 \) into (4.4). We then get

\[
\theta^2_j = \theta^1_j + \frac{\kappa \Delta t}{\Delta x^2} (\theta^1_{j-1} - 2\theta^1_j + \theta^1_{j+1}) ; \quad j = 2(1)J. \tag{4.9}
\]

Having thus found \( \theta \) at time level \( n = 2 \) we may proceed to time level \( n = 3 \) and so on for all time levels \( n \) up to and including \( n = N \).

We note that since \( x_{J+1} = L = J \Delta x \) we cannot choose \( J, L \) and \( \Delta x \) independently. Once two of them are chosen the third is given by the formula

\[
J = \frac{L}{\Delta x}. \tag{4.10}
\]

Likewise follows that

\[
N = \frac{T}{\Delta t} \tag{4.11}
\]

showing that \( N, \Delta t \) and \( T \) also depend on each other.

Since we know \( \theta \) at time \( t = 0 \) (or time level \( n = 0 \)), the application of a forward, one-sided finite difference approximation in time, as for instance employed in (4.4), is the obvious choice. The accuracy of this scheme is thus \( O(\Delta t) \), while the application of a centered finite difference approximation in space makes the spatial accuracy higher (\( O(\Delta x^2) \)). To increase the accuracy to the same level in time we may for instance employ a centered in time scheme for the time rate of change as well, that is, let

\[
[\partial_t \theta]^n_j = \frac{\theta^{n+1}_j - \theta^{n-1}_j}{2\Delta t}. \tag{4.12}
\]

Substitution of (4.12) into (4.1) then gives is

\[
\frac{\theta^{n+1}_j - \theta^{n-1}_j}{2\Delta t} = \kappa \frac{\theta^n_{j-1} - 2\theta^n_j + \theta^n_{j+1}}{\Delta x^2} ; \quad \begin{cases} j = 2(1)J \\ n = 0(1)N \end{cases}, \tag{4.13}
\]

or

\[
\theta^{n+1}_j = \theta^{n-1}_j + \kappa \frac{2\Delta t}{\Delta x^2} (\theta^n_{j-1} - 2\theta^n_j + \theta^n_{j+1}) ; \quad \begin{cases} j = 2(1)J \\ n = 0(1)N \end{cases}. \tag{4.14}
\]

To obtain the solution at the first time level \( n = 1 \), that is, to obtain \( \theta^1_j \) we again substitute \( n = 0 \) into (4.14) which gives

\[
\theta^1_j = \theta^1_{j-1} + \kappa \frac{2\Delta t}{\Delta x^2} (\theta^0_{j-1} - 2\theta^0_j + \theta^0_{j+1}) ; \quad j = 2(1)J. \tag{4.15}
\]
We must then require knowledge of $\theta_j^{-1}$. This corresponds to knowing the potential temperature anomaly at a time $t < 0$, in this case at one time level prior to the first time level. By using the one-sided forward scheme we avoid this problem, but sacrifice accuracy. As shown in the next Sections 4.2 - 4.5 there are, however, more pressing needs that makes us shy away from using a centered in time, centered in space finite difference approximation to solve the diffusion equation numerically.

### 4.2 Numerical stability

Another major problem with the scheme (4.14) is that it is numerically unstable. This means that the numerical solution, instead of following the continuous solution, steadily deviates from it. Commonly this happens explosively just like an analytic instability (think of baroclinic and barotropic instabilities in the atmosphere and ocean). We therefore call this behavior numerical instability to distinguish it from the physical barotropic and baroclinic instabilities that we would actually like to simulate using our numerical model. For our numerical solution to have any legitimacy we must require that it is numerically stable. This is an absolute requirement and is formulated as follows:

A numerical scheme is stable if and only if the numerical solution is limited within any given finite time span.

As a prelude to how we analyze the numerical scheme with respect to its numerical stability, let us first consider the analytic solution to (4.1). To this end we note that any good function $\theta$ may be written as a sum of cosines and sines or even more compact as a sum of exponentials (see for instance Section 2.11 on page 26 or Lighthill, 1970, page 3)

$$\theta(x, t) = \sum_{m=-\infty}^{\infty} \Theta_m(\omega_m, t) e^{i\omega_m x} = \sum_{m=-\infty}^{\infty} \theta_m$$

(4.16)

where $\omega_m$ is the wavenumber of the $m$’t Fourier component. Each component in (4.16), that is,

$$\theta_m = \Theta_m e^{i\omega_m x},$$

(4.17)

is called a Fourier component. Here $\Theta_m(t)$ is the time dependent amplitude of the $m$’t component. Substituting (4.16) into (4.1) we obtain

$$\partial_t \Theta_m = -\kappa \omega_m^2 \Theta_m.$$  

(4.18)

Note that we have dropped the summation, that is, we analyze each Fourier component separately. We observe that (4.18) is an ordinary differential equation (ODE). Solving it with respect to $\Theta$ gives

$$\Theta_m = \Theta_m^0 e^{-\kappa \omega_m^2 t}.$$  

(4.19)
Here $\Theta_m^0$ is the initial amplitude of mode $m$, that is, the value of $\Theta_m$ at $t = 0$. We find these initial amplitudes by expanding the initial distribution of $\theta$ into a Fourier series that is,

$$\theta(x, 0) = \sum_{m = -\infty}^{\infty} \Theta_m^0 e^{i\alpha_m x} \quad (4.20)$$

Thus substituting (4.19) into (4.16) we get

$$\theta(x, t) = \sum_{m = -\infty}^{\infty} \Theta_m^0 e^{-\kappa_m^2 t} e^{i\alpha_m x}, \quad (4.21)$$

which is then the analytic solution to (4.1). We note by looking at (4.19) that the amplitude $\Theta_m$ of each individual Fourier component decreases monotonically and exponentially as time increases. Furthermore, we observe that the shortest waves (highest wave numbers) decrease faster than the amplitude of the longer waves (low wave numbers). This is in accord with Section 3.2 where we concluded, based on (3.11), that diffusion acts to smooth out disturbances. Moreover we learn from (4.21) that this smoothing is not the same for all wavelengths. In fact it is selective in the sense that small scale disturbances are smoothed fast while the longer waves are less prone to damping in the same time period. Thus diffusion acts like a filter efficiently smoothing the small scale noise, if any, without significantly damping the longer period motion.

As is obvious we would like the numerical solution in general to behave accordingly. In particular we expect the numerical solution to the diffusion equation to decrease monotonically in time. Thus if the numerical solution increases in time it is obviously wrong and possibly unstable. Note that this instability has nothing to do with the accuracy of the chosen scheme. Yet it is the initial truncation error inherent in our scheme that is allowed to grow uncontrolled when the solution is unstable. We will return to this in Section 4.3 below.

To be able to analyze whether our chosen scheme is stable or not we need a proper mathematical definition. The requirement of numerical stability is commonly formulated by stating that for any finite time $T$, that is, for $0 < T < \infty$, there must exist a finite number, say $B$, such that

$$\left| \frac{\theta}{\theta_0} \right| \leq B, \quad (4.22)$$

where $\theta_0$ is the initial value of $\theta$, that is, $\theta_0 = \theta(x, 0)$. For linear systems, and to certain degree also non-linear systems, it is possible to analyze the stability of the chosen scheme analytically. Note that we always perform such an analysis before implementing the chosen scheme on the computer.

### 4.3 Stability analysis: von Neumann’s method

One such method is the so called von Neumann’s method. To analyze the stability von Neumann suggested to use a method somewhat similar to solving the equations analytically. The first step is to define a discrete Fourier component similar to the analytic one given in (4.17), that is,

$$\theta_j^n = \Theta_j e^{i\alpha_j \Delta x}, \quad (4.23)$$

...
where $\Theta_n$ is the discrete amplitude at time level $n$ and $\alpha$ is the wavenumber of that particular discrete mode\(^1\). We now define a growth factor $G$ by

$$
G \equiv \frac{\Theta_{n+1}}{\Theta_n} \Rightarrow \Theta_{n+1} = G\Theta_n \quad \text{and} \quad \Theta_{n-1} = G^{-1}\Theta_n.
$$

(4.24)

Thus $G$ is the amplification of the amplitude $\Theta$ as we proceed from one time level to the next. We observe that (4.24) is formally similar to (4.22), except that the growth factor $G$ is defined as the ratio between the next and the former time level, that is, between time level $n + 1$ and time level $n$, while (4.22) is the ratio between the value at a random time level and the initial value. Letting $n = 0$ in (4.24) then gives

$$
\Theta_1 = G\Theta_0
$$

(4.25)

where $\Theta_0$ is the initial amplitude. By letting $n = 1$ in (4.24) and making use of (4.25) we obtain

$$
\Theta_2 = G\Theta_1 = G^2\Theta_0
$$

(4.26)

Continuing by letting $n = 3, 4, \ldots$ up to a random number $n = l$ then gives

$$
\Theta_l = G^l\Theta_0
$$

(4.27)

Thus $G^l$ is the ratio between the amplitude at the random time level $n = l$ or random time $t = l\Delta t$ and the initial amplitude. Thus (4.22) is satisfied if

$$
|G| \leq 1
$$

(4.28)

since then $G^l$ decreases as the time level or time increases\(^2\). The criterion (4.28) is called von Neumann’s condition for stability. Note that it is a sufficient condition, not a necessary condition. We return to this in Section 4.4 below.

As our first example we use von Neumann’s method to analyze the forward in time, centered in space scheme for the diffusion as given by (4.4). Thus we first substitute (4.23) into (4.4) to obtain

$$
\Theta_{n+1} = \Theta_n + \frac{\kappa\Delta t}{\Delta x^2} \left(e^{i\alpha\Delta x} - 2 + e^{-i\alpha\Delta x}\right) \Theta_n
$$

(4.29)

where the common factor $e^{i\alpha j\Delta x}$ is removed. Noting that $e^{i\alpha\Delta x} + e^{-i\alpha\Delta x} = 2\cos \alpha\Delta x$ we get

$$
\Theta_{n+1} = \left[1 - 2\frac{\kappa\Delta t}{\Delta x^2}(1 - \cos \alpha\Delta x)\right] \Theta_n.
$$

(4.30)

We thus find the growth factor by simply dividing (4.30) by $\Theta_n$,

$$
G = 1 - 2\frac{\kappa\Delta t}{\Delta x^2}(1 - \cos \alpha\Delta x).
$$

(4.31)

---

1 Note that we have dropped the subscript $m$ on $\Theta$ and $\alpha$ for clarity.

2 Confer Computer Problem No. 1
To satisfy (4.28) we observe that

\[-1 \leq 1 - 2 \frac{\kappa \Delta t}{\Delta x^2} (1 - \cos \alpha \Delta x) \leq 1. \tag{4.32}\]

Since \(0 \leq (1 - \cos \alpha \Delta x) \leq 2\) the right-hand side inequality is satisfied for all \(\Delta t\) and \(\Delta x\). The inequality on the left-hand side, however, is satisfied if and only if

\[\frac{\kappa \Delta t}{\Delta x^2} (1 - \cos \alpha \Delta x) \leq 1. \tag{4.33}\]

Recall that \(0 \leq (1 - \cos \alpha \Delta x) \leq 2\), and hence (4.33) is satisfied for all wavenumbers \(\alpha\) if

\[\frac{\kappa \Delta t}{\Delta x^2} \leq \frac{1}{2} \text{ or } \Delta t \leq \frac{\Delta x^2}{2\kappa}. \tag{4.34}\]

This condition also ensures that (4.32) is satisfied, and hence that von Neumann’s condition (4.28) is satisfied as well. Furthermore (4.34) tells us that we cannot choose \(\Delta x\) and \(\Delta t\) independently. Once \(\Delta x\) is chosen the time step \(\Delta t\) must be chosen in accord with (4.34). We therefore say that the forward in time, centered in space scheme (4.4) is conditionally stable under the condition (4.34).

We also observe from (4.33) that the waves that first violate the inequality are waves with wavenumbers given by

\[\cos \alpha \Delta x = -1 \tag{4.35}\]

corresponding to those waves that maximizes \(1 - \cos \alpha \Delta x\). The solutions to (4.35) are wavenumbers \(\alpha_m\) given by

\[\alpha_m \Delta x = (2m - 1)\pi; \quad m = 1, 2, \ldots. \tag{4.36}\]

with corresponding wavelengths

\[\lambda_m = \frac{2\pi}{\alpha_m} = \frac{2\Delta x}{2m - 1}. \tag{4.37}\]

The most dominant of these waves is the wave corresponding to \(m = 1\). Thus the most unstable wave has wavelength

\[\lambda_1 = 2\Delta x. \tag{4.38}\]

This implies that the numerical instability will appear as “2\(\Delta x\)” noise, that is, noise of wavelength 2\(\Delta x\), and commonly as a saw tooth pattern as displayed in Figure 4.2.

We mentioned that the forward in time, centered in space scheme (4.4) for the diffusion equation is a conditionally stable scheme. If no such condition can be found that makes the chosen FD approximation satisfies the von Neumann’s condition (4.28), then we say that the scheme is unconditionally unstable. If von Neumann’s condition is satisfied regardless of our choice of spatial and temporal increments (e.g., \(\Delta x\) and \(\Delta t\)), then we say that the scheme is unconditionally stable. If the special case \(|G| = 1\) is true then we in addition say that the scheme is neutrally stable.
Figure 4.2: Displayed are solutions of the diffusion equation using the scheme (4.4) for respectively $K = \kappa \Delta t / \Delta x^2 = 0.45$ (left panel) and $K = 0.55$ (right panel) for $x \in (0, 1)$. The dependent variable $\theta$ is held fixed at the two boundaries $x = 0, 1$ and the initial condition is $\theta = \sin \pi x)$. The solutions are shown for the time levels $n = 0, n = 50$ and $n = 90$. Note the saw tooth like pattern in the right panel for $n = 90$ not present in the left panel. This indicates that the stability condition (4.34) is violated for $K = 0.55$, but not for $K = 0.45$.

It is worthwhile mentioning that when $|G| < 1$ it follows from (4.24) that $|\Theta_{n+1}| < |\Theta_n|$. Thus, inherent for all schemes for which $|G| < 1$ is that they include artificial numerical energy dissipation$^3$. We emphasize that even if the physical problem does not exhibit energy dissipation the numerical solution may exhibit such a decrease. We therefore refer to this artificial energy dissipation as numerical dissipation. We note that this dissipation depends on the absolute value of the growth factor and hence by implication on our choice of scheme and spatial and temporal increments. It is therefore of importance to ensure that the numerical dissipation is as small as possible by making choices so that the absolute value of the growth factor is as close to one as possible.

For problems that include natural energy dissipation it is therefore important to ensure that the numerical energy dissipation is small compared to the physical dissipation. We therefore always favor neutral schemes ($|G| = 1$), since such schemes are energy conserving, a highly desirably property. If this is not possible we recommend to choose the time step and the space increments so as to minimize the numerical energy dissipation. This is the same as requiring $|G|$ to be as close to one as possible. Regarding the forward in time, centered in space scheme this implies that we have to choose a time step $\Delta t$ that is small enough to satisfy (4.34), but at the same time is large enough to make $\Delta t \sim \Delta x^2 / 2 \kappa$.

$^3$In this context energy dissipation means that the amplitude of the solution decreases in time
4.4 The necessary stability condition

We mentioned above that von Neumann’s condition is a sufficient condition. This implies that if (4.28) is satisfied then the scheme is definitively stable. The question is if its too strict, that is, if it is also the necessary condition?

To determine this we return to the original requirement as formulated in (4.22). Substituting the discrete Fourier component (4.23) into (4.22) then gives

\[ \left| \frac{\Theta_{n+1}}{\Theta_0} \right| \leq B \implies |G|^n \leq B. \] (4.39)

Taking the natural logarithmic on both sides then gives

\[ n \ln |G| \leq \ln B \equiv B’. \] (4.40)

Even if von Neumann’s condition is too strict \(|G|\) cannot be very much larger than one. Thus we may write \(|G| = 1 + \epsilon\) where \(\epsilon\) is a small (\(\epsilon \ll 1\)) positive number. Hence it follows that \(\ln |G| = \ln(1 + \epsilon) \approx \epsilon\). Furthermore we note that \(t^n = n \Delta t\) or that at \(n = t^n / \Delta t\). When we substitute these expressions into (4.40) we obtain

\[ \epsilon \leq \frac{B’ \Delta t}{t^n} = O(\Delta t). \] (4.41)

Thus the necessary condition that satisfies the numerical stability requirement is

\[ |G| \leq 1 + O(\Delta t). \] (4.42)

This shows that von Neumann’s condition (4.28) is indeed too strict. However, most physical problems, even those containing instabilities, always involves some physical energy dissipation. Thus for all practical purposes we may apply the sufficient condition \(|G| \leq 1\) when analyzing the numerical stability of our schemes, in particular if \(|G| \ll 1\).

Finally, we remark that the growth factor \(G\) associated with the one-dimensional diffusion equation, as displayed by (4.31), is a scalar. For multi-variable and multi-dimensional problems the growth factor will commonly be a tensor or matrix, say \(G\). The sufficient condition is then that its spectral radius is less than or equal to one. This is tantamount to requiring that the largest eigenvalue of \(G\) is less than or equal to one.

4.5 Explicit and implicit schemes

The schemes (4.4) and (4.14) both contains the variable at time level \(n + 1\) on the left-hand side. In addition all the terms on the right-hand side are evaluated at the time level \(n\) or earlier \((n-1, n-2, \ldots)\). We refer to such schemes as being explicit. In contrast, if the right-hand side includes variables evaluated at the new time level \(n + 1\) we refer to the scheme as being implicit. Furthermore if we treat a multi-variable problem, e.g., the shallow water equations, where some
of the terms are treated as being explicit and some implicit we commonly refer the scheme to as being semi-implicit.

Explicit schemes, as exemplified by (4.4) and (4.14), are always relatively simple to solve. Once the unknowns are known for one time level at all grid points, the computation of the next time level is straightforward, we just proceed from one grid point to the next. This is in most cases not true for implicit and semi-implicit schemes.

Whether a scheme is explicit or implicit also impacts the stability of the scheme. Let us for instance consider the centered in time, centered in space scheme applied to the diffusion equation (4.1). If we use an explicit scheme we then evaluate all the terms on the right-hand-side at the time level \( n \). Accordingly the scheme becomes the one displayed in (4.14). If we choose to make the scheme implicit we would then evaluate all the terms on the right-hand side at the new time level \( n + 1 \). The scheme then becomes

\[
\theta^{n+1}_j = \theta^{n-1}_j + \kappa \frac{2\Delta t}{\Delta x^2} \left( \theta^{n+1}_{j-1} - 2\theta^{n+1}_j + \theta^{n+1}_{j+1} \right) ; \quad \{ j = 2(1)J \} \quad n = 0(1) \ldots .
\]

(4.43)

Let us first analyze the stability of the explicit scheme using von Neumann’s method. Substitution of the Fourier component (4.23) into (4.14), removing the common factor \( e^{i\alpha j \Delta x} \) and using (4.24) then gives

\[
G = G^{-1} - \kappa \frac{4\Delta t}{\Delta x^2} (1 - \cos \alpha \Delta x).
\]

(4.44)

Multiplying by \( G \) and rearranging terms yields the equation

\[
G^2 + 2\lambda G - 1 = 0,
\]

(4.45)

where

\[
\lambda = \kappa \frac{2\Delta t}{\Delta x^2} (1 - \cos \alpha \Delta x) \geq 0,
\]

(4.46)

to determine the growth factor. Solving (4.45) with respect to \( G \) we get the two solutions

\[
G_{1,2} = -\lambda \pm \sqrt{1 + \lambda^2}.
\]

(4.47)

We recall that in order to be numerically stable both solutions must satisfy von Neumann’s condition. We observe that

\[
|G_2| = \lambda + \sqrt{1 + \lambda^2} \geq 1,
\]

(4.48)

and hence that the centered in time, centered in space explicit scheme for the diffusion equation is unconditionally unstable. Thus:

\textit{Never use a centered in time, centered in space scheme for the diffusion problem. It is always unconditionally unstable.}

Then what about the implicit scheme (4.43)? To analyze the stability of the scheme (4.43) we first rearrange the terms to obtain

\[
\theta^{n+1}_{j-1} - \left( 2 + \frac{\Delta x^2}{2\kappa \Delta t} \right) \theta^{n+1}_j + \theta^{n+1}_{j+1} = -\frac{\Delta x^2}{2\kappa \Delta t} \theta^{n-1}_j ; \quad \{ j = 2(1)J \} \quad n = 0(1) \ldots .
\]

(4.49)
4.6 DuFort-Frankel

We note in passing that the implicit formulation require us to solve for $\theta$ at time level $n+1$ at the three grid points $j-1$, $j$, and $j+1$ simultaneously. We solve these kind of equations efficiently by employing an elliptic solver as for instance shown in Section 4.8. Curiously enough the implicit formulation of the parabolic diffusion equation turns it into an elliptic numerical equation.

To determine the growth factor we employ von Neumann’s method. Thus substituting the Fourier component into (4.49) we get

$$G \left[ 2 \cos \alpha \Delta x - \left( 2 + \frac{\Delta x^2}{2\kappa \Delta t} \right) \right] = -\frac{\Delta x^2}{2\kappa \Delta t} G^{-1}. \tag{4.50}$$

Multiplying by $\frac{2\kappa \Delta t}{\Delta x^2} G$ and solving with respect to the growth factor $G$ then gives

$$|G_1| = |G_2| = \frac{1}{\sqrt{1 + \frac{4\kappa \Delta t}{\Delta x^2} (1 - \cos \alpha \Delta x)}}. \tag{4.51}$$

We observe that $|G_{1,2}|$ is in fact less than or equal to one for all $\Delta t$ and $\Delta x$. Hence the implicit formulation of the CTCS scheme is unconditionally stable. This property is shared by all implicit and semi-implicit schemes. They are in fact all of them unconditionally stable. In contrast we recall that the explicit formulation gave an unconditionally unstable scheme. Furthermore, we notice that the implicit formulation (4.43) does not contain any constraint on the time step $\Delta t$, and hence we may choose $\Delta t$ to be as long as we wish.

Nevertheless it is commonly wise to stay away from implicit schemes, in particular for the terms forming the dominant balance in our governing equations. The rationale is, as exemplified by (4.51), that the growth factor $|G|$ is less than one for all wavelengths longer than the shortest resolved waves of wavelength $2\Delta x$. Moreover, $|G|$ decreases with increasing time step. Thus the implicit formulation always contains numerical dissipation which increases with increasing time step. Again this is a property shared by all implicit schemes. Although implicit schemes are unconditionally stable, we strongly advice to choose a time step $\Delta t$ so as to bring the growth factor as close to one as possible. The rationale is that we must control the numerical dissipation inherent in all implicit schemes. Thus although implicit schemes have no restrictions on the length of the time step its length is nevertheless constrained by the need to keep the numerical dissipation as small as possible.

4.6 Consistency and inconsistency: DuFort-Frankel

One important aspect about the schemes we use to solve our governing equations numerically are that they should mimic their continuous counterpart. We must therefore require that the employed scheme converges toward the continuous equations in the limit when $\Delta t \to 0$ and the space increments $\Delta x, \Delta y, \Delta z \to 0$ independently. Note that this requirement is independent of how they go to zero and independent of how fast each of them goes to zero. If this requirement is met we say that the scheme is consistent. If not we say that the scheme is inconsistent. Together with the absolute requirement of numerical stability it forms the two fundamental properties that our schemes should obey.
As is obvious all schemes where the finite difference approximations are based on Taylor series expansions as outlined in Section 2.6 satisfies the consistency requirement. Since both of the schemes (4.4) and (4.14) are based on a Taylor series expansion, they are both prime examples of consistent schemes.

We may, however, quite easily construct numerical schemes without having to resort to Taylor series expansions. It is particular in cases when the scheme is not constructed based on Taylor series expansions that we need to analyze its consistency. If the scheme turns out to be inconsistent we must require as a minimum that the dominant physical processes that the governing equations simulate are consistently represented in the finite difference analogue. One example of such a scheme elaborated below is the DuFort-Frankel scheme.

To construct the DuFort-Frankel scheme we start with the consistent centered in time, centered in space scheme (4.14). As shown above (Section 4.5) this scheme is unconditionally unstable when applied to the diffusion equation. However, in many cases the diffusion term is added just to even out small scale noise, that is, to dissipate energy that accumulates on the smaller scales. In these cases the term does not represent the dominant physics and we may relax on our consistency requirement. To construct the scheme we observe that the value of $\theta_j^n$ at the grid point $x_j, t^n$ in space and time can be thought of as a linear interpolation in time of the two adjacent grid points $x_j, t^{n+1}$ and $j, n - 1$, or

$$
\theta_j^n = \frac{1}{2} \left( \theta_j^{n+1} + \theta_j^{n-1} \right).
$$

(4.52)

Using this as a substitute for $\theta_j^n$ on the right-hand side of (4.14) we obtain

$$
\theta_j^{n+1} = \theta_j^{n-1} + 2\frac{\kappa \Delta t}{\Delta x^2} \left( \theta_{j-1}^{n-1} - \theta_j^{n-1} + \theta_{j+1}^{n+1} - \theta_j^{n+1} + \theta_j^n \right); \quad \left\{ \begin{array}{l} j = 2(1)J \\ n = 0(1)N - 1 \end{array} \right.
$$

(4.53)

We first note that the introduction of the term $\theta_j^{n+1}$ on the right-hand side makes the new scheme implicit. As all implicit schemes we therefore expect it to be stable. Curiously enough, by simply adding some implicity to the centered in time, centered in space scheme we have turned it into an unconditionally stable scheme (see Exercise 4 on page 57). We note that in contrast to the implicit scheme (4.43), the implicity is now limited to the single term $\theta_j^{n+1}$ involving only the space grid point $x_j$. We may move this term from the right-hand side of (4.53) to its left-hand side. After some rewriting we get

$$
\theta_j^{n+1} = \left[ \theta_j^{n-1} + \chi (\theta_{j-1}^n - \theta_j^{n-1} + \theta_{j+1}^n) \right] (1 + \chi)^{-1}.
$$

(4.54)

where

$$
\chi = \frac{2\kappa \Delta t}{\Delta x^2}.
$$

(4.55)

Thus the numerical algorithm associated with the implicit DuFort-Frankel scheme (4.53) is solved explicitly. This is one reason why it has become so popular, in particular in oceanography, (e.g., Adamec and O’Brien, 1978). A second reason, mentioned in the previous paragraph, is that the scheme is implicit and thus, like all implicit schemes, is unconditionally stable.
4.6 DuFort-Frankel

THE DIFFUSION PROBLEM

It remains to analyze the consistency of the scheme. To this end we employ the Taylor series expansions of Section 2.6,

\[ \theta^n_j = \theta^n_{j+1} = \theta^n_{j-1} + \partial \theta^n_j \Delta x + \frac{1}{2} \partial^2 \theta^n_j \Delta x^2 + \frac{1}{6} \partial^3 \theta^n_j \Delta x^3 + O(\Delta x^4), \]  
(4.56)

and

\[ \theta^{n+1}_j = \theta^n_{j+1} = \theta^n_{j-1} + \partial \theta^n_j \Delta t + \frac{1}{2} \partial^2 \theta^n_j \Delta t^2 + \frac{1}{6} \partial^3 \theta^n_j \Delta t^3 + O(\Delta t^4). \]  
(4.57)

Substitution of these series in (4.53) then gives

\[ \partial \theta^n_j - \kappa \partial^2 \theta^n_j = -\kappa \partial^2 \theta^n_j + O(\Delta t^2) + O(\Delta x^2). \]  
(4.58)

where

\[ \kappa = \kappa \left( \frac{\Delta t}{\Delta x} \right)^2. \]  
(4.59)

To be consistent all the terms on the right-hand side of (4.58) must converge to zero in the limit \( \Delta x \to 0 \) and \( \Delta t \to 0 \). This is, however, not the case for the first term on the right-hand side which tends to infinity if \( \Delta x \) tends to zero faster than \( \Delta t \). We therefore note that (4.58) only converges to the continuous equation if \( \kappa \to 0 \) when \( \Delta x \to 0 \) and \( \Delta t \to 0 \). This implies that the scheme is consistent if and only if \( \kappa \Delta t \) tends to zero faster than \( \Delta x \). Thus there is a condition associated with the consistency of the scheme, and hence, in line with the formulation used for the stability condition, we refer to the DuFort-Frankel scheme as a conditionally consistent scheme under the condition that \( \Delta x \to 0 \) slower than \( \Delta t \to 0 \).

As already mentioned at the end of Section 3.2 the diffusion term is often used as a numerical method or trick to dissipate energy contained on the smaller scales in atmospheric and oceanographic models. Commonly this “noise” is created due to the presence of non-linear terms in the governing equations. This leads to non-linear interaction among the various wavelengths which in turn is responsible for a more or less continuous cascade of energy towards progressively smaller and smaller scales. If we neglect to dissipate the energy contained in the tail of the energy spectrum we get an accumulation of energy at the \( 2\Delta x - 4\Delta x \) scales. In turn this accumulation at some time or another into the integration leads to a violation of the linear, numerical stability criterion and the numerical model blows up.

When the diffusion term is used for this purpose it does not represent any of the physical processes that we want to resolve. Rather it is introduced as a numerical method to avoid our model to blow up. Nevertheless it does represent low energy physics acting on scales smaller than those resolved by our grid.\(^5\) It is therefore a parameterization of these processes. Since this parameterization and/or the parameters it contains may change in accord with the models resolution we refer to it as subgrid scale (SGS) parameterization. Note that the SGS processes act on scales smaller than our grid resolution.

\(^4\)The formulation “if and only if” is often written “iff”.

\(^5\)Note that for a given grid size \( \Delta x \) the resolution equals the Nüquist wavelength \( 2\Delta x \).
4.7 The Crank-Nicholson scheme

We will now consider another popular scheme called the *Crank-Nicholson* scheme. Like the Dufort-Frankel scheme it is also implicit. Its popularity is due to two facts. First, like all implicit schemes, it is unconditionally stable. Second it has a second order accuracy in time and space.

We start by defining the convenient notation
\[
\delta^2_x \psi^n_j = \psi^n_{j-1} - 2\psi^n_j + \psi^n_{j+1},
\]
which permit us to write the explicit, forward in time, centered in space scheme (4.4) as
\[
\theta^{n+1}_j = \theta^n_j + \frac{\kappa \Delta t}{\Delta x^2} \delta^2_x \theta^n_j.
\]

Recall that this scheme is conditionally stable under the condition
\[
\chi \equiv \frac{2\kappa \Delta t}{\Delta x^2} \leq 1.
\]

We may now construct an implicit scheme by changing the second term on the right-hand side of (4.61) from time level \(n\) to time level \(n+1\). The result is
\[
\theta^{n+1}_j = \theta^n_j + \frac{1}{2} \chi \delta^2_x \theta^{n+1}_j.
\]

We recall from Section 4.5 that implicit schemes are always stable. Thus (4.63) is an unconditionally stable scheme. Note that it is also consistent since it is based on Taylor series expansions. We also note that (4.63) is similar to (4.43) except that (4.63) is forward in time while (4.43) is centered in time.

We now combine the two schemes (4.61) and (4.63) to obtain
\[
\theta^{n+1}_j = \theta^n_j + \frac{1}{2} \chi \left[ \gamma \delta^2_x \theta^{n+1}_j + (1 - \gamma) \delta^2_x \theta^n_j \right],
\]
where \(\gamma\) is a number so that \(0 \leq \gamma \leq 1\). If \(\gamma = 0\) then (4.64) reduces to the explicit scheme (4.61). If \(\gamma = 1\) then (4.64) reduces to the implicit scheme (4.63). If \(\gamma\) is between 0 and 1 then the scheme contains implicit as well as explicit terms. Such schemes are sometimes referred to as a sem i-implicit scheme. To analyze the stability of the scheme we use von Neumann’s method. The growth factor \(G\) is (cf. Exercise 5 on page 57)
\[
G = \frac{1 - (1 - \gamma) \chi (1 - \cos \alpha \Delta x)}{1 + \gamma \chi (1 - \cos \alpha \Delta x)}.
\]

We recall that the condition \(|G| \leq 1\) or \(-1 \leq G \leq 1\) is a sufficient condition for numerical stability. From (4.65) follows that \(G \leq 1\) is always satisfied, while \(G \geq -1\) is satisfied if
\[
\chi (1 - 2\gamma) \leq 1.
\]
For $\frac{1}{2} \leq \gamma \leq 1$ (4.66) is automatically satisfied. Under these circumstances the scheme is stable regardless of the value chosen for the increments $\Delta x$ and $\Delta t$. This implies that the scheme (4.64) is unconditionally stable as long as $\frac{1}{2} \leq \gamma \leq 1$. This is not surprising because under these circumstances the weight is on the implicit part. As a corollary we note that this also proves that the pure implicit scheme (4.63), which follows from (4.64) by letting $\gamma = 1$, is indeed unconditionally stable. If however $0 \leq \gamma < \frac{1}{2}$ the weight is on the explicit part. Under these circumstances the scheme is conditionally stable under the condition (4.66). We note that for $\gamma = 0$, in which case (4.64) equals the forward in time, centered in space finite difference approximation for the diffusion equation as displayed in (4.4), we indeed retrieve the condition (4.34) of Section 4.3, that is, $\chi \leq 1$.

The value $\gamma = \frac{1}{2}$ is special. It constitutes the critical value at which the scheme (4.64) is still unconditionally stable. For this special value we get

\[\theta_{j}^{n+1} = \theta_{j}^{n} + \frac{1}{2}\kappa\Delta t \left[ \frac{\theta_{j-1}^{n+1} - 2\theta_{j}^{n+1} + \theta_{j+1}^{n+1}}{\Delta x^2} + \frac{\theta_{j-1}^{n} - 2\theta_{j}^{n} + \theta_{j+1}^{n}}{\Delta x^2} \right], \quad (4.67)\]

which is the the Crank-Nicholson scheme. The scheme is special also in another respect. Despite the fact that we employ a forward in time finite difference approximation for the time rate of change of the variable $\theta$, it becomes of second order accuracy in time as well as in space. To prove it we start by utilizing the Taylor series expansions (2.25) and (2.27) as outlined in Section 2.6 on page 18. By substituting these series into the centered differences on the right-hand side of (4.67) we first get

\[\frac{\theta_{j}^{n+1} - \theta_{j}^{n}}{\Delta t} = \frac{1}{2}\kappa \left( \partial_{t}^{2}\theta|_{j}^{n+1} + \partial_{x}^{2}\theta|_{j}^{n} \right) + O(\Delta x^2). \quad (4.68)\]

Expanding $\theta_{j}^{n+1}$ and $\partial_{x}^{2}\theta|_{j}^{n+1}$ using Taylor series gives

\[\frac{\theta_{j}^{n+1} - \theta_{j}^{n}}{\Delta t} = \partial_{t}\theta|_{j}^{n} + \frac{1}{2}\partial_{t}^{2}\theta|_{j}^{n} \Delta t + O(\Delta t^2), \quad (4.69)\]
\[\partial_{x}^{2}\theta|_{j}^{n+1} = \partial_{x}^{2}\theta|_{j}^{n} + \partial_{t}(\partial_{x}^{2}\theta)|_{j}^{n} \Delta t + O(\Delta t^2). \quad (4.70)\]

Substituting these series in (4.68) and rearranging terms we get

\[\partial_{t}\theta|_{j}^{n} = \kappa\partial_{x}^{2}\theta|_{j}^{n} - \frac{1}{2}\left\{ \partial_{t}^{2}\theta|_{j}^{n} - \kappa \partial_{t}(\partial_{x}^{2}\theta)|_{j}^{n} \right\} \Delta t + O(\Delta t^2) + O(\Delta x^2), \quad (4.71)\]

Furthermore, we apply the continuous diffusion equation (4.1) to obtain

\[\partial_{t}^{2}\theta|_{j}^{n} = \kappa \partial_{t}(\partial_{x}^{2}\theta)|_{j}^{n}. \quad (4.72)\]

Thus the second term on the right-hand side of (4.71) vanishes and (4.71) becomes

\[\partial_{t}\theta|_{j}^{n} = \kappa\partial_{x}^{2}\theta|_{j}^{n} + O(\Delta t^2) + O(\Delta x^2). \quad (4.73)\]
Thus besides being unconditionally stable, the Crank-Nicholson scheme is also of second order accuracy in time even though we employ a forward in time finite difference approximation. These two facts is why the Crank-Nicholson scheme is popular when solving true diffusive problems. However, as shown in the next section it has one disadvantage compared to the more standard schemes. It turns the otherwise parabolic equation into one which numerically look like an elliptic problem, and hence requires us to employ an elliptic solver for every time step.

4.8 A direct elliptic solver

If we inspect the Crank-Nicholson scheme (4.67) in more detail we find that including the implicit terms on the right-hand side of (4.67) turns the original parabolic equation into an apparent elliptic equation. By adding the new terms the original local algorithm is turned into a non-local or global algorithm. The implication is that the solution $\theta_{n+1}^j$ at time level $n+1$ in addition to depend on the solution at the previous time level $n$ also depends on $\theta_{n+1}^{j-1}$ and $\theta_{n+1}^{j+1}$, that is, $\theta_{n+1}^j$ depends on the solution at the adjacent space points $\pm \Delta x$ away at the new time level $n+1$. We may illustrate this by rearranging the terms in (4.67) to obtain

$$-\kappa\Delta t \frac{\Delta^2}{\Delta x^2} \theta_{j-1}^{n+1} + \left(1 + 2\kappa\Delta t \frac{\Delta^2}{\Delta x^2}\right) \theta_j^{n+1} - \kappa\Delta t \frac{\Delta^2}{\Delta x^2} \theta_{j+1}^{n+1} = \theta_j^n + \kappa\Delta t \frac{\Delta^2}{\Delta x^2} \left[\theta_{j-1}^n - 2\theta_j^n + \theta_{j+1}^n\right].$$

(4.74)

Thus we cannot solve for $\theta_{j}^{n+1}$ without knowing $\theta_{j-1}^{n+1}$ and/or $\theta_{j+1}^{n+1}$. Let us consider that we solve (4.74) for increasing values of $j$. Then for any arbitrary $j$ we have already solved for $\theta_{j-1}^{n+1}$, and it thus known. However, we have not yet solved for $j+1$, and thus $\theta_{j+1}^{n+1}$ is unknown.

Many of the model codes employed in numerical weather and numerical ocean weather prediction today employ semi-implicit methods. We are therefore in need of a method whereby such problem can be solved. Such methods are commonly referred to as elliptic solvers. Moreover, since we deal with time marching problems, we have to apply the elliptic solver for each time step. Thus we additionally require that the elliptic solver is efficient in the sense that it is fast on any computer. The most efficient elliptic solvers are those referred to as direct elliptic solvers. One such method is the so called Gauss elimination which we will us as an example. It consists of two steps. The first is called a forward sweep. Next we find the final solution by performing a backward substitution.

To get started, we first rewrite (4.74) into a more general form,

$$a_j \theta_{j-1}^{n+1} + b_j \theta_j^{n+1} + c_j \theta_{j+1}^{n+1} = h_j^n; \quad j = 2(1)J,$$

(4.75)

where $a_j$, $b_j$, and $c_j$ represents the coefficients in (4.74). The use of the subscript $j$ attached to these coefficients is to acknowledge that they in general are functions of space. Likewise $h_j$ on the right-hand side represents all “forcing” terms, that is, our knowledge of the solution at the previous time step(s). We also note that we are required to solve (4.75) within a finite domain.

---

6In the infancy of numerical weather prediction (NWP) most elliptic solvers were the iterative or indirect elliptic solvers. Even though they may be accelerated, as for instance when applying the iterative elliptic solver called “Successive over-relaxation”, they are much slower than the direct methods.
Thus at \( j = 1 \) and at \( j = J + 1 \) the variable \( \theta \) is determined by the boundary conditions. For convenience we assume that these are simple Dirichlet conditions, in which case \( \theta_1^n \) and \( \theta_{J+1}^n \) are known for all time levels \( n \).

For convenience we will also drop the superscript \( n \) and \( n + 1 \). Thus we are required to solve, 

\[
a_j \theta_{j-1} + b_j \theta_j + c_j \theta_{j+1} = h_j, \quad j = 2(1)J,
\]

under the conditions

\[
\theta_1 = \hat{\theta}_0, \quad \text{and} \quad \theta_{J+1} = \hat{\theta}_L,
\]

where \( \hat{\theta}_0 \) and \( \hat{\theta}_L \) are known functions. We observe that (4.76) may be more compactly written as

\[
\mathbf{A} \cdot \theta = h',
\]

where the tensor \( \mathbf{A} \) is the tridiagonal matrix 

\[
\mathbf{A} = \begin{bmatrix}
b_2 & c_2 & 0 & \ldots & 0 & 0 & 0 \\
b_3 & c_3 & 0 & \ldots & 0 & 0 & 0 \\
0 & b_4 & c_4 & \ldots & 0 & 0 & 0 \\
\vdots & \vdots & \vdots & \ddots & \vdots & \vdots & \vdots \\
\vdots & \vdots & \vdots & \ddots & \vdots & \vdots & \vdots \\
0 & 0 & 0 & \ldots & a_{J-1} & b_{J-1} & c_{J-1} \\
0 & 0 & 0 & \ldots & 0 & a_J & b_J
\end{bmatrix}, \quad (4.79)
\]

and the vectors \( \theta \) and \( h' \) are, respectively,

\[
\theta = \begin{bmatrix}
\theta_2 \\
\theta_3 \\
\vdots \\
\theta_{J-1} \\
\theta_J
\end{bmatrix}, \quad (4.80)
\]

and

\[
h' = \begin{bmatrix}
h_2 - a_2 \hat{\theta}_0 \\
h_3 \\
\vdots \\
h_J - c_J \hat{\theta}_L
\end{bmatrix}. \quad (4.81)
\]

**Forward sweep**

We are now ready to perform the forward sweep. The idea is to replace all elements of the matrix \( \mathbf{A} \) positioned in the lower left half with zeros. At the same time it is convenient to normalize the diagonal elements, that is, turn everyone of them into the value 1. We start with the equation for \( j = 2 \). From (4.78) follows

\[
b_2 \theta_2 + c_2 \theta_3 = h'_2. \quad (4.82)
\]
We then normalize by dividing by $b_2$

$$\theta_2 + d_2 \theta_3 = w_2,$$  \hspace{1cm} (4.83)

where

$$d_2 = \frac{e_2}{b_2} \text{ and } w_2 = \frac{h'_2}{b_2}.$$  \hspace{1cm} (4.84)

For $j = 3$ we get from (4.78)

$$a_3 \theta_2 + b_3 \theta_3 + c_3 \theta_4 = h'_3.$$  \hspace{1cm} (4.85)

Substituting for $\theta_2$ from (4.83) and normalizing gives

$$\theta_3 + d_3 \theta_4 = w_3$$  \hspace{1cm} (4.86)

where

$$d_3 = \frac{c_3}{b_3 - d_2 a_3} \text{ and } w_3 = \frac{h'_3 - a_3 w_2}{b_3 - d_2 a_3}.$$  \hspace{1cm} (4.87)

Repeating this for $j = 4$ we get

$$\theta_4 + d_4 \theta_5 = w_4$$  \hspace{1cm} (4.88)

where

$$d_4 = \frac{c_4}{b_4 - d_3 a_4} \text{ and } w_4 = \frac{h'_4 - a_4 w_3}{b_4 - d_3 a_4}.$$  \hspace{1cm} (4.89)

Thus repeating this procedure up to and including $j = J - 1$ we get

$$\theta_j + d_j \theta_{j+1} = w_j, \quad j = 2(1)J - 1.$$  \hspace{1cm} (4.90)

where the coefficients $d_j$ and $w_j$ are defined by the recursion formula

$$d_j = \begin{cases} 
\frac{e_j}{b_j} ; & j = 2 \\
\frac{c_j}{b_j - d_{j-1} a_j} ; & j = 3(1)J - 1 \\
0 ; & j = J 
\end{cases}, \quad w_j = \begin{cases} 
\frac{h'_j}{b_j} ; & j = 2 \\
\frac{h'_j - a_j w_{j-1}}{b_j - d_{j-1} a_j} ; & j = 3(1)J 
\end{cases},$$  \hspace{1cm} (4.91)

respectively. We observe that $d_J$ is set to zero. This is due to the fact that for $j = J$ (4.78) becomes

$$a_J \theta_{J-1} + b_J \theta_J = h'_J,$$  \hspace{1cm} (4.92)

and hence by substituting for $\theta_{J-1}$ from (4.90) we simply get

$$\theta_J = w_J.$$  \hspace{1cm} (4.93)

Note that all the coefficients $d_j$ and $w_j$ can be calculated once and for all.

We also notice that in matrix form (4.79) now reads

$$\mathbf{A'} \cdot \mathbf{\theta} = \mathbf{w},$$  \hspace{1cm} (4.94)
where the matrix $\mathcal{A}'$ is
\[
\mathcal{A}' = \begin{bmatrix}
1 & d_2 & 0 & \ldots & 0 & 0 \\
0 & 1 & d_3 & \ldots & 0 & 0 \\
0 & 0 & 1 & \ldots & 0 & 0 \\
\vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\
0 & 0 & 0 & \ldots & 1 & d_{J-1} \\
0 & 0 & 0 & \ldots & 0 & 1
\end{bmatrix},
\] (4.95)

and the vector $w$ is
\[
w = \begin{bmatrix}
w_2 \\
w_3 \\
\vdots \\
w_{J-1} \\
w_J
\end{bmatrix}.
\] (4.96)

Thus we have completed a forward sweep in which the equation matrix is normalized and is upper triangular only.

**Backward substitution**

We are now ready to perform the backward substitution. First we note that all the $w_j$’s and the $d_j$’s are known using the recursion formula (4.91). Second we note from (4.93) that $\theta_j$ is simply given by $w_j$ and that the latter is known from (4.91). Thus we are in a position where $\theta_j$ is known. Hence applying (4.90) for $j = J - 1$ and solving with respect to $\theta_{J-1}$ we get
\[
\theta_{J-1} = w_{J-1} - d_{J-1} \theta_J.
\] (4.97)

We may continue this and solve for $\theta_{J-2}$, $\theta_{J-3}$, \ldots and so on. Hence we can solve for all the remaining $\theta_j$’s by backward substitution into (4.90), that is,
\[
\theta_j = w_j - d_j \theta_{j+1} \quad \text{for} \quad j = J - 1(-1)2.
\] (4.98)

The Gauss elimination method is very simple to program, and it is also very efficient and fast on the computer. An example on the usefulness of this method, in which you are also required to program the method, is given in Computer Problem 5 named “Yoshida’s equatorial jet current” in the accompanied, but separate Computer Problem notes. We urge the reader to do this problem, and at least solve the resulting ODE by employing the Gauss elimination method.

**Exercises**

1. Show that the scheme (4.14) is unconditionally unstable. Hint: Show that $|G| > 1$ regardless of the choice made for $\Delta t$ and $\Delta x$.

2. Show that if $|G| = 1$ then the chosen scheme has no numerical dissipation.
3. Show that the growth factor associates with the scheme (4.43) is

\[
G = \left[ 1 + \frac{4\kappa \Delta t}{\Delta x^2} (1 - \cos \alpha \Delta x) \right]^{-\frac{1}{2}}
\]  

(4.99)

and hence that the scheme is unconditionally stable. Also show that \(|G| < 1\) for all wavelengths. Note that \(|G|\) decreases as \(\Delta t\) increases.

4. Show that the growth factor for the DuFort-Frankel scheme (4.53) is

\[
G_{1,2} = \gamma \cos \alpha \Delta x \pm \sqrt{1 - \gamma^2 (\sin \alpha \Delta x)^2} \frac{1}{1 + \gamma} \quad \text{where} \quad \gamma = \frac{2\kappa \Delta t}{\Delta x^2}. 
\]  

(4.100)

5. Show that the expression (4.65) is indeed the expression for the growth factor of the scheme (4.64) when using von Neumann’s analysis method.
Chapter 5

THE ADVECTION PROBLEM

In this chapter we will investigate possible numerical methods whereby the advection equation (3.14) may be solved by numerical means. As we did for the diffusion problem we will reduce it to its simplest form\(^1\), and hence we consider a one-dimensional advection process, that is, let \( F_A = u \theta \). Under these circumstances the advection equation (3.14) in its continuous form becomes

\[
\frac{\partial}{\partial t} \theta + \frac{\partial}{\partial x} (u \theta) = 0. \tag{5.1}
\]

Here \( u(x, t) \) is the advection speed along the \( x \)-axis. Note that in general \( u \) varies in time and space in which case the problem is non-linear. Following Einstein’s principle of making things as simple as possible, but no simpler, we will, however, in most instances below assume that \( u \) is uniform in time and space. We therefore let \( u = u_0 = \text{constant} \). Hence (5.1) becomes

\[
\frac{\partial}{\partial t} \theta + u_0 \frac{\partial}{\partial x} \theta = 0. \tag{5.2}
\]

The general true solution to (5.2) is\(^2\)

\[
\theta = \theta(x - u_0 t). \tag{5.3}
\]

We note that as long as \( \theta \) is a good function (cf. Chapter 2, Section 2.11 on page 26) the true solution may be written as a Fourier series. Thus,

\[
\theta = \sum_{n=-\infty}^{\infty} \Theta_n e^{i \alpha_n (x - u_0 t)} \tag{5.4}
\]

where \( \alpha_n \) is the wavenumber of mode number \( n \), and \( \Theta_n \) is the associated amplitude of that mode. Let us assume that the initial condition is simply a harmonic (monochromatic) wave of wavelength \( \lambda \) and amplitude \( \Theta_0 \) then initially

\[
\theta(x, 0) = \Theta_0 e^{i \frac{2\pi}{\lambda} x}. \tag{5.5}
\]

\(^1\)“Make things as simple as possible, but no simpler” Albert Einstein (1879-1955)

\(^2\)cf. Exercise no. 1 on page 86 at end of this Chapter.
To satisfy (5.4) the true solution to the advection equation also becomes a monochromatic wave of the same wavelength, that is,

$$\theta(x,t) = \Theta_0 e^{ix(x-u_0 t)}. \quad (5.6)$$

This particular solution is a wave propagating in the positive $x$ direction with a phase speed given by $u_0$. As alluded to in Chapter 1, this solution is typical of hyperbolic systems. Indeed, the solutions (5.3) and (5.6) is such that if we travel along with the advection speed we will experience no change in the property $\theta$. If we, however, observe the wave from a fixed position in space, the property $\theta$ will change in accord with (5.3) as the “wave” passes by.

In fact by using Fourier transforms as outlined in Section 2.12 we may illustrate that, regardless of the initial distribution, all solutions to (5.2) is propagated with the speed $u_0^3$. We start by noting that the solution to (5.2) may be formally written as

$$\theta(x,t) = \frac{1}{2\pi} \int_{-\infty}^{\infty} \tilde{\theta}(t,\alpha) e^{i\alpha x} d\alpha. \quad (5.7)$$

where $\tilde{\theta}(t,\alpha)$ is the Fourier transform of $\theta$, that is,

$$\tilde{\theta}(t,\alpha) = \int_{-\infty}^{\infty} \theta(x,t) e^{-i\alpha x} dx. \quad (5.8)$$

To find the Fourier transform we first transform the advection equation (5.2) and get

$$\frac{d}{dt} \tilde{\theta} = i\alpha u_0 \tilde{\theta}. \quad (5.9)$$

Solving this equation we get

$$\tilde{\theta}(t,\alpha) = \tilde{\Theta}(\alpha) e^{-i\alpha u_0 t}, \quad (5.10)$$

where $\tilde{\Theta}(\alpha)$ is the Fourier transform of the initial condition. We note that the advection equation (5.2) contains first order derivatives in time and space only. We are therefore only allowed to specify two boundary conditions. One of these is the initial condition. Let us represent the initial condition at time $t = 0$ by the $\theta_0(x)$. Hence

$$\tilde{\Theta}(\alpha) = \int_{-\infty}^{\infty} \theta_0(x) e^{-i\alpha x} dx. \quad (5.11)$$

By use of (5.7) we can therefore formally write the solution to the one-dimensional advection equation (5.2), satisfying the initial condition, as

$$\theta(x,t) = \frac{1}{2\pi} \int_{-\infty}^{\infty} \tilde{\Theta}(\alpha) e^{i\alpha(x-u_0 t)} d\alpha. \quad (5.12)$$

This formal solution underscores that the general true solution to the advection equation as given in (5.3) consists of waves of various wavelengths (wavenumber) all of which propagates at the same speed $u_0$. Furthermore, (5.12) underscores that the dominant waves, that is, the waves containing most energy, are those associated with the waves that initially have the the largest amplitudes, that is, the peaks in the Fourier spectrum of $\tilde{\Theta}(\alpha)$.  

---

3This may also be inferred from Section 3.3. In our case $u = u_0 i$ and hence $\nabla \cdot u = 0$. From (3.15) then follows that the variance of $\theta$ is conserved implying that any initial distribution is conserved as time progresses.
5.1 Finite difference forms

Our concern is to develop an algorithm by which (5.2) can be solved by numerical means. To find a suitable finite difference equation to (5.2) we follow the procedure used to derive the finite difference equation for the diffusion equation. Thus we start by replacing the terms in the advection equation by finite difference approximations. However, our experience when performing this on the diffusion equation taught us that this might not be straightforward in that some numerical schemes must be discarded because they might be numerically unstable or inconsistent.

Our task is therefore to construct a scheme that is numerically stable and consistent. The latter is done by performing a stability analysis as well as a consistency analysis. If the scheme turns out to be unstable or inconsistent then it is of no use to us and have to be discarded. Hence once a finite difference form is chosen we have to analyze it with respect to its stability and consistency. The former is performed making use of von Neumann’s method (cf. Section 4.3), while consistency is analyzed via use of Taylor series (cf. Section 2.6) as outlined in Section 4.6 on page 50.

Let us start by applying the successful forward in time and centered in space (FTCS) scheme that worked well for the diffusion equation. Thus we first construct a forward in time finite difference approximations for the time rate of change. To this end we use Taylor series expansions and get

\[ [\partial_t \theta]^n_j = \frac{\theta_{n+1}^j - \theta^n_j}{\Delta t}. \] (5.13)

In a similar fashion, we next construct a centered in space finite difference approximation for the first order space differential in (5.2), that is,

\[ [\partial_x \theta]^n_j = \frac{\theta^n_{j+1} - \theta^n_{j-1}}{2\Delta x}. \] (5.14)

By replacing the two terms in (5.2) by the finite difference approximations (5.13) and (5.14) we get the finite difference equation

\[ \frac{\theta_{j+1}^n - \theta_j^n}{\Delta t} + u_0 \frac{\theta_{j+1}^n - \theta_{j-1}^n}{2\Delta x} = 0, \] (5.15)

which we solve with respect to \( \theta_{j+1}^n \) to get

\[ \theta_{j+1}^n = \theta_j^n - u_0 \frac{\Delta t}{2\Delta x} \left( \theta_{j+1}^n - \theta_{j-1}^n \right). \] (5.16)

Since the finite difference approximation (5.16) is based on Taylor series expansions, we know a priori that the consistency requirement is satisfied. It therefore remains to analyze its stability to satisfy ourselves that the scheme is useful to us. To this end we make use of von Neumann’s method, and start by substituting the discrete Fourier component (4.23) into (5.16). After some manipulations we get

\[ \Theta_{n+1} = \Theta_n - \frac{u_0 \Delta t}{2\Delta x} \left( e^{i\alpha \Delta x} - e^{-i\alpha \Delta x} \right) \Theta_n, \] (5.17)
where the common factor \( e^{i\alpha \Delta x} \) is tossed away. Recalling the definition of the growth factor (4.24), and noting that \( e^{i\alpha \Delta x} - e^{-i\alpha \Delta x} = 2i \sin \alpha \Delta x \), we find

\[
G = 1 - i\lambda,
\]

where

\[
\lambda = \frac{u_0 \Delta t}{\Delta x} \sin \alpha \Delta x.
\]

We observe that the growth factor now is a complex number with a real part given by 1 and an imaginary part given by \( \lambda \). According to von Neumann’s method we are required to evaluate the absolute value of the growth factor. To this end we use the well known property of complex numbers, namely that its absolute value equals the square root of the sum of the squares of the real and imaginary parts\(^4\). Thus

\[
|G| = \sqrt{1 + \lambda^2}.
\]

We observe that since \( \lambda^2 \) is a positive definite the radical is always larger than or equal to one and hence \( |G| \geq 1 \). Since \( |G| = 1 \) only for the special wavenumbers that make \( \sin \alpha \Delta x = 0 \) it follows that the scheme is in general \textit{unconditionally unstable}. We then have the somewhat curious result that although the forward in time, centered in space scheme worked fine for the diffusion problem, it is totally unacceptable with regard to the advection problem.

\[
\text{Never use a forward in time, centered in space scheme for the advection problem. It always leads to an unconditionally unstable scheme.}
\]

This does not come as a total surprise. As alluded to in Chapter 2 the advection equation and the diffusion equation represent quite different physics and have quite different characteristics. While the diffusion equation is parabolic, the advection equation is hyperbolic. We should therefore expect that a FD scheme that works well for the diffusion problem not necessarily works well for the advection problem.

As listed in, e.g., O’Brien (1986) (page 165 and onwards) there are many stable schemes that have been suggested over the past to solve the advection equation. The reason is that advection is one of the most prominent atmospheric and oceanic processes. Many of the schemes are constructed to minimize unwanted properties of other schemes (e.g., numerical diffusion as detailed in Section 5.11 below, or numerical dispersion as detailed in Section 5.13 below), but a host of other schemes are constructed with focus on their efficiency on the computer.

With the advent of today’s supercomputers, with their dramatic increase in power and speed, the earlier requirements of computer efficiency is simply lessened. The focus is therefore shifting towards deriving schemes providing better conservation properties, and containing higher order accuracy, say schemes of \( \mathcal{O}(\Delta t^4, \Delta x^4) \) or higher (cf. Section 10.1 on page 143). Such schemes includes flux corrective schemes (cf. Section 5.12 on page 77), and Lagrangian schemes (cf. Section 5.9 on page 71).

---

\(^4\)Let \( A = a + ib \) be an imaginary number with real part \( a \) and imaginary part \( b \). Then \( |A| = \sqrt{AA^*} = \sqrt{a^2 + b^2} \) where \( A^* = a - ib \) is the complex conjugate of \( A \).
The leapfrog scheme

One of the first schemes employed in atmospheric and oceanic problems is the centered in time and centered in space (CTCS) scheme, a scheme we emphasized was totally useless for the diffusion problem. To construct the scheme we use Taylor series expansions. A centered in time finite difference approximation for the time rate of change is

$$\frac{\partial_t \theta^n_j}{2\Delta t} = \frac{\theta^{n+1}_j - \theta^{n-1}_j}{2\Delta t}, \quad (5.21)$$

while a centered in space finite difference approximation to the first order space derivative is

$$\frac{\partial_x \theta^n_j}{2\Delta x} = \frac{\theta^{n+1}_j - \theta^{n-1}_j}{2\Delta x}. \quad (5.22)$$

By replacing the two terms in (5.2) by their finite difference approximations (5.21) and (5.22), and rearranging terms, we get

$$\theta^{n+1}_j = \theta^{n-1}_j - u_0 \frac{\Delta t}{\Delta x} \left( \theta^{n+1}_{j+1} - \theta^{n+1}_{j-1} \right). \quad (5.23)$$

Note that since we used centered approximations to derive the finite difference equation (5.23) the truncation error is of $O(\Delta x^2)$ and $O(\Delta t^2)$. The scheme is therefore often referred to as a second order scheme. The scheme is also commonly referred to as the leapfrog scheme. The rationale is that we use information about the variable $\theta$ from all the points surrounding the $x_j, t^n$ point in $x, t$ space, that is, $\theta^n_{j\pm1}$ and $\theta^n_{j-1}$ to find $\theta^{n+1}$, but do not incorporate information of the variable from the point $x_j, t^n$ itself, that is, $\theta^n_j$. In a sense we leapfrog the point $x_j, t^n$.

The scheme is traditionally fairly popular for several reasons. For one the scheme is, as shown in Section 5.3, neutrally stable. Hence there is no numerical or artificial damping or energy dissipation associated with the scheme when going from one time level to the next, a highly desirable property. Secondly the scheme is of second order accuracy. Finally, the scheme is easy to implement and works fast and efficiently on computers. It has, however, some disadvantageous properties. First, as detailed in Section 5.13 (page 79), the scheme contains what is referred to as numerical dispersion, which sometimes leads to negative tracer concentrations. Secondly the scheme, as detailed in Section 5.14 below, also contains what is referred to as unphysical modes which has to be dealt with.

---

5In describing their model people often writes “..., while a second order scheme is employed for advective terms”.
5.3 Stability of the leapfrog scheme: The CFL condition

We note that since the leapfrog scheme is derived exclusively using Taylor series, it is a consistent scheme. What about its stability? And if stable, under what conditions is it stable?

To analyze its stability we make use of von Neumann’s method as outlined in Section 4.3. Thus we first replace the dependent variable \( \theta \) in (5.23) by its discrete Fourier component (4.23) to give

\[
\Theta_{n+1} = \Theta_{n-1} - 2i u_0 \frac{\Delta t}{\Delta x} \sin \alpha \Delta x \Theta_n
\]  

(5.24)

To find the growth factor \( G \) we first make use of (4.24) and then multiply by the growth factor to obtain

\[
G^2 + 2i \lambda G - 1 = 0,
\]  

(5.25)

where

\[
\lambda = u_0 \frac{\Delta t}{\Delta x} \sin \alpha \Delta x.
\]  

(5.26)

The solution is

\[
G_{1,2} = -i \lambda \pm \sqrt{1 - \lambda^2},
\]  

(5.27)

and hence the growth factor is a complex function. This was to be expected since the factor in front of the first order term in (5.25) is imaginary. Let us assume that the radical in (5.27) is positive. Then reusing the theorem that the absolute value of a complex number is the square root of the complex number itself multiplied by its complex conjugate we get (cf. eq. 5.20 of Section 5.1)

\[
|G_{1,2}| = \sqrt{G_{1,2} G_{1,2}^*} = \sqrt{1 - \lambda^2 + \lambda^2} = 1.
\]  

(5.28)

Since by definition \( \Theta_{n+1} \equiv |G| \Theta_n \) it follows that there is no artificial or numerical damping or dissipation involved when advancing from one time level to the next. The scheme (5.23) is therefore neutrally stable implying that the energy associated with \( \theta \) is conserved. We note that since we have assumed the velocity by which the property \( \theta \) is advected to be constant (i.e., \( u = u_0 i \)) its gradient is zero. Hence it is highly desirable that the property of the scheme is in line with the property of advection processes as outlined in Section 3.3, namely that when the gradient of the advection velocity is zero then the variance should be conserved.

To obtain this desirable property we recall that we require the radical in (5.27) to be positive\(^6\). Hence we require

\[
1 - \lambda^2 \geq 0 \quad \text{or} \quad |\lambda| \leq 1.
\]  

(5.29)

Since \(-1 < \sin \alpha \Delta x < 1\) we note that if

\[
|u_0| \frac{\Delta t}{\Delta x} \leq 1
\]  

(5.30)

then (5.29) is satisfied. Thus (5.30) is a sufficient condition for stability for the leapfrog scheme (5.23). Moreover under this condition the leapfrog scheme also becomes neutrally stable. The

\[^6\text{If the radical is negative the scheme may be shown to be unconditionally unstable, that is, } |G| > 1.\]
5.4 The upstream scheme

The upstream scheme condition (5.30) is referred to as the Courant-Friedrich-Levy condition or simply the CFL condition for stability. The ratio or non-dimensional number

\[ C = \frac{|u_0| \Delta t}{\Delta x} \]  

is usually referred to as the Courant number and the CFL condition becomes simply \( C \leq 1 \). Since \( \Delta x \) more often than not is given by the need to resolve the spatial structure or typical wavelengths of the physical problem, the CFL condition becomes a stringent upper bound on the time step \( \Delta t \), that is, \( \Delta t \leq \Delta x / |u_0| \). Therefore the larger the advection speed the smaller the time step, and the smaller the grid size the smaller the time step.

The latter fact has some implications. In the atmosphere and ocean the dominant wavelength (or scale) is associated with the first baroclinic Rossby radius. At Norwegian latitudes this length is about 500 km in the atmosphere and as small as 10 km in the ocean. Thus the time step for oceanic advective processes are much smaller than the similar time step we have to use for advective processes in the atmosphere.

5.4 The upstream scheme

Because of these somewhat unwanted properties of the leapfrog scheme the so called upwind or upstream scheme became quite popular early on, and is quite common even today. As the name indicates the scheme make use of information exclusively from points upstream to calculate the value at the new time level. It is a two time level scheme which is forward in time and one-sided in space. If the advection velocity is positive it is backward in space, and forward in space if the advection velocity is negative. Thus, again using Taylor series expansions, we get

\[ \theta_j^{n+1} = \theta_j^n - |u_0| \frac{\Delta t}{\Delta x} \begin{cases} \theta_j^n - \theta_{j-1}^n & ; u_0 \geq 0 \\ \theta_{j+1}^n - \theta_j^n & ; u_0 < 0 \end{cases} \]  

(5.32)

The scheme is conditionally stable under the CFL condition, that is, \( |u_0| \Delta t \leq \Delta x \) or \( \Delta t \leq \Delta x / |u_0| \) (cf. Exercise 2 on page 87). We note for later convenience that (5.32) may be written as

\[ \theta_j^{n+1} = (1 - C) \theta_j^n + C \begin{cases} \theta_{j-1}^n & ; u_0 \geq 0 \\ \theta_{j+1}^n & ; u_0 < 0 \end{cases} \]  

(5.33)

where \( C \) is the Courant number as defined in (5.31).

One of the major advantages of the upwind scheme is that it conserves the fact that tracer concentration is a positive definite quantity. Furthermore we observe that it is a consistent scheme since it is derived using Taylor series. Being a two level scheme it also works fast end efficient on the computer. Despite of these circumstances the upwind scheme has one major drawback. It contains, as detailed in Section 5.11 on page 75, what we refer to as numerical diffusion. Depending on the choices we make regarding the time step and the space increment the numerical diffusion may be large and sometimes larger than the actual physical diffusion of the original problem. It therefore tends to even out the solution artificially as time progresses. In particular
areas where large gradients appear, e.g., frontal areas, are prone to artificial diffusion. Thus fronts are diffused which in turn inhibits realistic baroclinic instability processes. In addition we observe that the upwind scheme has truncation errors that are first order in time and space, that is, its accuracy is $O(\Delta t)$ and $O(\Delta x)$, which is one order of magnitude less than the leapfrog scheme. Because of these rather disadvantageous properties this author does not recommend the use of the upstream scheme alone. Its usefulness is that it forms the basis for more complicated schemes, e.g., the Lax-Wendroff scheme.

## 5.5 The diffusive scheme

As shown in Section 5.1 the FTCS scheme applied to the advection equation gives an unconditionally unstable scheme. In an attempt to avoid this numerical instability, but retain a forward in time scheme it was early on suggested to replace $\theta^n_j$ in (5.16) by \[ \frac{1}{2}(\theta^n_{j+1} + \theta^n_{j-1}) \]. Thus we get

\[ \theta^{n+1}_j = \frac{1}{2}(\theta^n_{j+1} + \theta^n_{j-1}) - u_0 \frac{\Delta t}{2\Delta x} (\theta^n_{j+1} - \theta^n_{j-1}). \] (5.34)

To be useful to us we must satisfy ourselves that the scheme is stable and consistent. The latter is not obvious in this case since the scheme is no longer based on Taylor series. To analyze its consistency we first subtract $\theta^n_j$ on both sides of (5.34) to get

\[ \theta^{n+1}_j - \theta^n_j = \frac{1}{2}(\theta^n_{j+1} - 2\theta^n_j + \theta^n_{j-1}) - u_0 \frac{\Delta t}{2\Delta x} (\theta^n_{j+1} - \theta^n_{j-1}) \] (5.35)

Then using Taylor series we observe that the terms on the left-hand side becomes

\[ \theta^{n+1}_j - \theta^n_j = \partial_t \theta^n_j \Delta t + \frac{1}{2} \partial^2 \theta^n_j \Delta t^2 + O(\Delta t^3) \] (5.36)

while the first and second term on the right-hand side become

\[ \frac{1}{2}(\theta^n_{j+1} - 2\theta^n_j + \theta^n_{j-1}) = \frac{1}{2} \partial^2 \theta^n_j \Delta x^2 + O(\Delta x^4), \] (5.37)

and

\[ u_0 \frac{\Delta t}{2\Delta x} (\theta^n_{j+1} - \theta^n_{j-1}) = u_0 \Delta t \left( \partial_x \theta^n_j + \frac{1}{6} \partial^3 \theta^n_j \Delta x^2 + O(\Delta x^4) \right), \] (5.38)

respectively. Thus substituting (5.36) - (5.38) into (5.35) and rearranging terms we get

\[ \partial_t \theta^n_j + u_0 \partial_x \theta^n_j = \frac{1}{2} \frac{\Delta x^2}{\Delta t} (1 - C^2) \partial^2 \theta^n_j + O(\Delta x^2) + O(\Delta t). \] (5.39)

where $C$ is the Courant number as defined in (5.31). Hence we find that the scheme is only conditionally consistent in that $\Delta x^2$ has to go to zero faster than $\Delta t$. We also observe that the root of this inconsistency is the first term on the right-hand side of (5.39). We observe that this term acts as a diffusion term with a diffusion coefficient of $\kappa = \frac{1}{2} \frac{\Delta x^2}{\Delta t} (1 - C^2)$. The
finite difference equation (5.34) in which we employ finite space and time increments, therefore contains a diffusive term not present in the continuous advection equation. This is why the numerical scheme (5.34) is commonly referred to as the diffusive scheme.

We now turn to analyze the stability of the diffusive scheme. Again using von Neumann’s method we find that the growth factor is given by

\[ G = \sqrt{1 - (1 - C^2) \sin^2 \alpha \Delta x}, \]  

(5.40)

where \( C \) is the Courant number. Since \( \sin^2 \alpha \Delta x \) is a positive definite it follows that \( |G| \leq 1 \) iff \( C \leq 1 \). Hence the scheme is conditionally stable under the condition that the Courant number is less than or equal to one. We notice that this condition is exactly the same as the one we derived for the leapfrog and the upwind schemes.

### 5.6 The Lax-Wendroff scheme

To avoid or lessen the impact of the first order numerical diffusion inherent in the diffusive scheme, to make it consistent, and to increase its accuracy, Richtmyer and Morton (1967) advocated the use of a scheme based on the work of Lax and Wendroff (1960) now commonly referred to as the Lax-Wendroff scheme. It is a two step scheme which combines the diffusive scheme and the leapfrog scheme by first performing a diffusive step (also known as the predictor step) and then add a leapfrog step (also known as the corrective step). In summary we first derive the solution at the mid time level \( t_{n+\frac{1}{2}} \) and at the mid increments in space \( x_{j+\frac{1}{2}} \) (cf. the crosses marked in the dashed grid of Figure 5.1) employing the diffusive scheme. Then we use the results as the basis for the corrective step. In the latter we employ the leapfrog scheme to calculate the value of the variable at the new time level \( t_{n+1} \) and at the regular grid points in space, that is, \( x_j \) (cf. the circled points in the solid grid in Figure 5.1).

![Grid Layout](image)

Figure 5.1: Displayed is the grid layout for the Lax-Wendroff scheme. The solid lines denote the grid through the circled \( x_j, t^n \) points. The dashed lines denote the grid through the \( x_{j+\frac{1}{2}}, t^{n+\frac{1}{2}} \)-points which are marked with a cross.
Thus in the predictor step we construct a forward in time, centered in space finite difference equation employing the diffusive scheme. Note that in the this step we only proceed to time level \( n + \frac{1}{2} \), that is, we compute \( \theta_{j+\frac{1}{2}}^{n+1} \) using (5.34). Hence we get

\[
\theta_{j+\frac{1}{2}}^{n+1} - \frac{1}{2} \left( \theta_{j+1}^{n} + \theta_{j}^{n} \right) - \frac{u_0}{\Delta x} \left( \theta_{j+1}^{n} - \theta_{j}^{n} \right) = 0, \tag{5.41}
\]

or

\[
\theta_{j+\frac{1}{2}}^{n+1} = \frac{1}{2} \left( \theta_{j+1}^{n} + \theta_{j}^{n} \right) - \frac{u_0}{2\Delta x} \left( \theta_{j+1}^{n} - \theta_{j}^{n} \right). \tag{5.42}
\]

We note that even though the forward in time, centered in space scheme is unstable the trick of replacing \( \theta_j^n \) by half the sum of its nearest space neighbors makes the scheme stable under the condition that the Courant number is less than or equal to one as shown in Section 5.5.

The second step is to employ the leapfrog scheme (5.23) to find \( \theta_{j}^{n+1} \), that is, the solution at time level \( n + 1 \) at the point \( j \) based on the values found at the mid time level. Thus we get\(^7\)

\[
\theta_j^{n+1} = \theta_j^n - \text{sgn}(u_0)|u_0| \frac{\Delta t}{\Delta x} \left( \theta_{j+\frac{1}{2}}^{n+\frac{1}{2}} - \theta_{j-\frac{1}{2}}^{n+\frac{1}{2}} \right). \tag{5.43}
\]

If we now eliminate the dependence on \( n + \frac{1}{2} \) and \( j + \frac{1}{2} \) by substituting (5.42) into (5.43) we get

\[
\theta_j^{n+1} = \theta_j^n - \frac{1}{2} \text{sgn}(u_0)C \left( \theta_{j+1}^{n} - \theta_{j-1}^{n} \right) + \frac{1}{2} C^2 \left( \theta_{j+1}^{n} - 2\theta_j^n + \theta_{j-1}^{n} \right), \tag{5.44}
\]

where \( C \) is the Courant number. We observe that the last term on the right-hand side of (5.44) looks like an FDA of the second order derivative of \( \theta \) with respect to \( x \), that is, a diffusive term, and it may therefore appear that the Lax-Wendroff scheme has some inherent numerical diffusion. This is however not the case as shown in the next paragraph.

We observe that the Lax-Wendroff scheme is an explicit scheme, and hence if it is numerically stable it is most probably only conditionally stable. Another question is whether the scheme is consistent? In contrast to the leapfrog scheme and the upwind scheme, who are both derived using Taylor series, the latter is not obvious. In fact we showed in the preceding section that the first step using a diffusive scheme is indeed inconsistent. To analyze the consistency of the Lax-Wendroff scheme we first substitute the respective Taylor series into (5.44). The result is

\[
\partial_t \theta|_j^n + u_0 \partial_x \theta|_j^n = -\frac{1}{2} \left( \partial_t^2 \theta|_j^n - u_0 \partial_x^2 \theta|_j^n \right) \Delta t + \mathcal{O}(\Delta t^2) + \mathcal{O}(\Delta x^2). \tag{5.45}
\]

Since (5.2) implies that \( \partial_t^2 \theta = -u_0 \partial_x (\partial_t \theta) = u_0^2 \partial_x^2 \theta \), we may neglect the second term on the right-hand side, and hence we get

\[
\partial_t \theta|_j^n + u_0 \partial_x \theta|_j^n = \mathcal{O}(\Delta t^2) + \mathcal{O}(\Delta x^2). \tag{5.46}
\]

---

\(^7\)The function \( \text{sgn} \) return the sign of its argument. Thus \( \text{sgn}(u_0) \) returns the sign of the velocity \( u_0 \), that is, \( u_0 = \text{sgn}(u_0)|u_0| \).
We first observe that (5.2) is recovered when $\Delta x \to 0$ and $\Delta t \to 0$ independently. Hence the scheme is indeed consistent. Second we observe that the terms neglected are $O(\Delta x^2)$ and $O(\Delta t^2)$, and hence that the scheme is second order accurate. Moreover, and equally important we got rid of the apparent diffusive term mentioned in the preceding paragraph.

To analyze its stability we use von Neumann’s method. Thus we start by substituting the variables by their respective discrete Fourier components and then find the equation for the growth factor $G$. Thus we get

$$G = (1 - C^2 + C^2 \cos \alpha \Delta x) - i \text{sgn}(u_0) C \sin \alpha \Delta x$$

which is a complex function. The magnitude of the growth factor is therefore

$$|G| = \sqrt{(1 - C^2 + C^2 \cos \alpha \Delta x)^2 + C^2 \sin^2 \alpha \Delta x}$$

which after some manipulation may be written

$$|G| = \sqrt{1 - (1 - C^2)C^2(1 - \cos \alpha \Delta x)^2}.$$  

(5.49)

We notice that the sign of the last term is determined by the factor $1 - C^2$ since the remaining factors are positive definite quantities. Thus as long as $1 - C^2$ is positive we observe from (5.49) that $|G| \leq 1$, and hence that the Lax-Wendroff scheme is conditionally stable under the condition that the Courant number is less than one. Some other good features about the scheme are the absence of a temporal unphysical mode present in the leapfrog scheme (cf. Section 5.14 on page 82) and that it avoids the diffusion characteristics of the upwind and diffusive schemes (cf. Section 5.11 on page 75).

### 5.7 An implicit scheme

As for the diffusion equation we may also construct an implicit scheme for the advection equation. This is easily constructed by using a backward in time, centered in space scheme. Thus

$$\theta_{j+1}^n = \theta_j^n - u_0 \frac{\Delta t}{2\Delta x} (\theta_{j+1}^{n+1} - \theta_{j-1}^{n+1}), \quad \begin{cases} j = 2(1)J - 1 \\ n = 0(1) \ldots \end{cases}$$

(5.50)

We observe that this scheme is $O(\Delta t)$ and $O(\Delta x^2)$, that is first order in time and second order in space accurate.

To investigate the stability of (5.50) we make use of von Neumann’s method. Thus we get

$$\Theta_{n+1} = \Theta_n - iC \sin \alpha \Delta x \Theta_{n+1},$$

(5.51)

which results in a complex growth factor given by

$$G = \frac{1}{1 + iC \sin \alpha \Delta x}.$$  

(5.52)
5.8 Solutions to the initial problem of CTCS schemes

THE ADVECTION PROBLEM

Figure 5.2: Comparison of the numerical solution to the advection equation (5.1) using the leapfrog, the upwind and the Lax-Wendroff schemes with a Courant number of one half ($C = 0.5$). The solution is shown after 10 cycles (periodic boundary condition). The true solution is a bell function as shown by the black solid curve. We note that both the leapfrog and the Lax-Wendroff schemes give rise to numerical dispersion, while the upwind scheme gives rise to numerical diffusion.

Thus the magnitude of the growth function is

$$|G| = \frac{1}{\sqrt{1 + C^2 \sin^2 \alpha \Delta x}}$$

which satisfies $|G| \leq 1$ for all finite time steps. Thus the implicit scheme, as expected, is unconditionally stable, and hence avoids the restrictive CFL condition.

Finally we note that the time step, just like the grid size, must be sufficient to resolve the typical periods of the physical problem. Commonly in atmospheric and oceanic applications the typical period is much longer than the Nüquist frequency $2\Delta t$, and hence the CFL condition in most cases puts a much more stringent requirement on $\Delta t$ than the requirement of resolving the typical periods of the physical problem. Thus for most meteorological and oceanographic problems the resolution requirement is on the grid size. We notice, however, that if $\Delta t$ becomes too large the growth factor will be small implying that the scheme will contain a large numerical dissipation.

5.8 Solutions to the initial problem of CTCS schemes

Although the leapfrog scheme is conditionally and neutrally stable it is not without disadvantageous properties. One of the problems with the scheme is associated with the number of initial
conditions required, a problem already touched upon in Section 4.1 regarding the diffusion problem. Thus the question is how to start the time marching procedure. The forward in time centered in space scheme is one possibility, but as shown in Section 5.1 this is unconditionally unstable. However, we may nevertheless make use of this scheme when applied to a single time step.

Thus we start by using the scheme

$$
\theta_j^i = \theta_j^0 - u_0 \frac{\Delta t}{2\Delta x} \left( \theta_{j+1}^0 - \theta_{j-1}^0 \right). \quad (5.54)
$$

For the time level 2 and onwards we then use the leapfrog scheme (5.23). The step (5.54) and more generally the forward in time, centered in space scheme is usually referred to as the Euler scheme. We emphasize that although the Euler scheme is unconditionally unstable it does not ruin the solution when applied for one time step only. It may even be used from time to time to avoid the unphysical mode inherent in the leapfrog scheme (cf. Section 5.14).

### 5.9 The method of characteristics

To help understand why the upwind scheme becomes unstable when the CFL condition is not satisfied we will first solve the advection equation using the method of characteristics.

#### Theory

Let the slopes

$$
\frac{D^* x}{dt} = u(x,t) \quad (5.55)
$$

define special curves in the $t, x$ space (cf. Fig. 5.3), and let us simultaneously define the special differential operator

$$
\frac{D^*}{dt} \equiv \partial_t + \frac{D^* x}{dt} \partial_x. \quad (5.56)
$$

Then the advection equation (5.2) may be rewritten to yield

$$
\frac{D^* \theta}{dt} = 0 \quad \text{along the slopes} \quad \frac{D^* x}{dt} = u_0. \quad (5.57)
$$

We commonly refer to the slopes defined by (5.55) as the *characteristics* and (5.55) itself as the *characteristic equation*. Since the solutions to (5.1) also are solutions to (5.57) we often refer (5.57) as the *compatibility equation*.

We observe that (5.57) tells us that $\theta$ is conserved along the characteristics (5.55). Thus if we know the solution at time $t = 0$, that is, $\theta(x, t = 0)$ for $0 < x < L$, then the solution at any later time $t > 0$ and at any particular point $x$ in space is found by simply following the characteristic back from the point $x, t$ toward the initial time $t = 0$ as illustrated in Figure 5.3. In our case with $u_0 = \text{constant}$ the characteristics deform to straight lines with positive slopes $1/u_0$ when $u_0 > 0$. From Figure 5.3 we may conclude that after a critical time $t = t_c = L/u_0$ all information about the initial distribution of $\theta$ is lost. Indeed for $t > t_c$ it follows that the solution...
5.9 Method of characteristics

THE ADVECTION PROBLEM

\[ \frac{\partial \theta}{\partial t} + c \frac{\partial \theta}{\partial x} = 0 \]

\( x = 0 \)
\( x = L \)

\( t = t_c \)
\( t = 0 \)

Figure 5.3: Sketch of the characteristics in the \( x, t \) plane. For \( u = u_0 = \text{constant} > 0 \) the characteristics are the straight lines sloping to the right in \( x, t \) space as given by (5.55). If \( x = L \) marks the end of the computational domain, then all information about the initial condition is lost for times \( t > t_c \).

in the computational domain \( 0 < x < L \) is determined wholly by the boundary condition at \( x = 0 \). Since (5.1) only contains the first derivative with respect to \( x \), only one condition in \( x \) is allowed. The boundary at \( x = L \) is therefore open in the sense that there is no boundary condition that replaces the differential equation there. The physical space therefore, in principle, continues to infinity. Thus the boundary \( x = L \) is a numerical boundary necessitated by the fact that any computer, however large, are limited in its capacity. This problem is especially compound for oceanographic models, since the oceanic spatial scales are small compared to the similar scales in the atmosphere. We will come back to this problem in Chapter 7 where we will investigate details concerning conditions constraining the solutions at open boundaries.

We note in passing that since (5.1) and (5.57) are compatible, a solution to (5.57) is also automatically a solution to (5.1). We may therefore solve (5.57) employing numerical methods as well as (5.1) in which case it is referred to as the method of characteristics. We will make use of this fact to give a physical interpretation of the CFL criterion. We also note that the method of characteristic may be applied to much more complex problems and systems than the simple advection equation (e.g., Lister, 1966; Røed and O’Brien, 1983). It is used here merely as an example.

**Numerical considerations**

Since in this simple case (5.57) tells us that \( \theta \) is conserved along the characteristics, the problem is reduced to find the characteristics in the \( x, t \) space, that is, to solve (5.55). As is now common
Figure 5.4: Sketch of the method of characteristics. The distance between the grid points are $\Delta t$ in the vertical and $\Delta x$ in the horizontal direction. The sloping solid line is the characteristic through the point $j, n + 1$. It is derived from (5.55) and the slope is given by $1/u$ ($u > 0$). The point labeled $Q$ is therefore a distance $u\Delta t$ to the left of $x_j$. As long as $u\Delta t < \Delta x$ then $Q$ is located between $x_{j-1}$ and $x_j$. If however $u\Delta t > \Delta x$ then the point $Q$ is located to the left of $x_{j-1}$.

we then divide the computational domain in the $x, t$ space into a grid as displayed in Figure 5.4. Let us for a minute consider that $u_0 = u(x, t)$, that is, is a function of space and time, and that it is known for all $x_j$ and at all time levels $t^n$. Then at the time level $t^{n+1}$ the characteristic through $x_j$ is simply given by the characteristic equation (5.55). To find the characteristic we make a simple finite difference approximation of (5.55) considering that $u_j^{n+1}$ can be approximated by $u_j^n$. Hence

$$\frac{x_j - x_Q}{\Delta t} = u_j^n \quad \text{or} \quad x_Q = x_j - u_j^n \Delta t$$

(5.58)

where $x_Q$ is the point at which the characteristic crosses the time level $t^n$. Since we know both $u_j^n$ and $x_j$, (5.58) is really an equation which determines the location of $x_Q$. We note that as long as $u_j^n \Delta t < \Delta x$ then $x_Q$ is located between the grid points $x_j$ and $x_{j-1}$. Since $\theta_j^n$ is known for all grid points, we may interpolate linearly between the adjacent grid points to find $\theta_Q^n$, or the value of $\theta$ at the location $x_Q$ at time level $t = t^n$. To this end we may use a two point linear interpolation. Thus

$$\theta_Q^n = \theta_{j-1}^n + \frac{x_Q - x_{j-1}}{\Delta x} (\theta_j^n - \theta_{j-1}^n)$$

(5.59)
5.10 Physical interpretation of the CFL condition

Figure 5.4 is drawn for \( u_0 > 0 \), implying that \( u_0 = |u_0| \), and may be used to visualize the CFL criterion for the upwind scheme. First we note that since \( u_0 > 0 \) (5.58) says that \( x_Q < x_j \). Moreover (5.58) also implies that the distance between \( x_Q \) and \( x_j \) is \( u_0 \Delta t \). Thus if we additionally desire that \( x_{j-1} \leq x_Q \) then \( u_0 \Delta t \leq \Delta x \). If we compare this result with the upwind scheme as given in (5.32) we observe that for \( u_0 > 0 \) the information used to compute \( \theta_{j+1}^n \) does originate by weighting the two points \( \theta_j^n \) and \( \theta_{j-1}^n \). In fact we may rewrite (5.32) for \( u_0 > 0 \) to give

\[ \theta_{j+1}^n = (1 - C_j) \theta_j^n + C_j \theta_{j-1}^n, \]

which matches (5.62) exactly. Thus from (5.63) follows that the upwind scheme may be interpreted as the value of \( \theta \) at the time level \( n+1 \), that is, \( \theta_{j+1}^n \), is found by a simple weighting of the values \( \theta_j^n \) and \( \theta_{j-1}^n \) using the Courant number as weight. What the method of characteristics (5.62) reveals is that the latter interpretation is only valid as long as \( |u_0| \Delta t \leq \Delta x \). This is exactly what the CFL criterion demands in order to make the numerical upwind scheme stable, that is, the Courant number must be less than one or that (5.30) must be satisfied.

Moreover, if \( |u_0| \Delta t > \Delta x \) then the characteristic through \( x_j \) at time level \( n+1 \) (cf. Figure 5.4) will cross the time level \( n \) to the left of \( x_{j-1} \), that is, \( x_{j-2} < x_Q < x_{j-1} \). Under these circumstances the upwind scheme will still use (5.63) to calculate \( \theta \) at the new time level, that is, continue to use the weighted information using values at \( x_j \) and \( x_{j-1} \) at the previous time level. This is obviously wrong and use of (5.63) will lead to a major error. If this is allowed to continue for time step after time step the error accumulates and will finally give rise to a numerical instability.

The speed defined by \( \Delta x \) and \( \Delta t \), that is, \( \Delta x/\Delta t \), is often referred to as the signal speed of the grid or simply the grid speed. The CFL criterion (5.30) may therefore be interpreted as a condition which constrains the grid speed to be larger than the advection speed \( u_0 \). In other words, the advection speed must be small enough to let the area of dependence be between within \( x_{j-1} \) and \( x_{j+1} \) at time level \( n \).
5.11 Numerical diffusion

Although the leapfrog scheme is neutrally stable we have just shown that it has at least one major disadvantage; it is dispersive. In particular, as displayed in Figure 5.7, this is true when the resolution is poor, that is, in areas where $\Delta x$ is inadequate to resolve the dominant wavelength. Also the impact of the dispersion increases with decreasing Courant number.

As a result the upstream scheme was for a long time the favored advection scheme. Unfortunately also the upstream scheme is far from perfect. It contains what is referred to as numerical diffusion. The name derives from the fact that the error made by truncating the Taylor series to first order to arrive at the finite difference equation (5.35) acts similar to physical diffusion, that is, the inherent truncation error tend to diminish differences in the tracer distribution. This is exemplified in Figure 5.5 where an initial narrow, peak like tracer distribution spreads out while being advected. In contrast the analytic solution that the numerical solution tries to mimic is one in which the initial tracer distribution is advected without change. We underscore that this does not imply that any tracer is lost. The numerical diffusion process, just like its physical counterpart, does conserve the total tracer content. What happens is that the numerical diffusion smooth out any differences in the initial tracer concentration. Thus it redistributes the initial tracer distribution while conserving the initial total tracer content. This is evident in Figure 5.5. Comparing the area under the dash-dot curve and the area under the solid curve they are actually the same. Note again that this redistribution is artificial and arises due to the application of the upstream scheme.

To analyze the origin of the numerical diffusion in the upstream scheme let us reconsider
5.11 Numerical diffusion

THE ADVECTION PROBLEM

(5.32). We first rewrite it in terms of an advective flux defined by

\[ F_j^n = \frac{1}{2} [(u_0 + |u_0|)\theta_j^n + (u_0 - |u_0|)\theta_{j+1}^n] \frac{\Delta t}{\Delta x}. \tag{5.64} \]

We note that since \( u_0 = \text{sgn}(u_0)|u_0| \). Hence the last term on the right-hand side of (5.64) is zero when \( u_0 \geq 0 \) and the first term is zero when \( u_0 < 0 \). Under these circumstances \( F_j^n = \text{sgn}(u_0)C\theta_j^n \). Thus (5.32) becomes

\[ \theta_j^{n+1} = \theta_j^n - (F_j^n - F_{j-1}^n), \tag{5.65} \]

which is valid regardless of the sign of \( u_0 \). If we substitute each of the terms in (5.65) by its associated Taylor series expansion, that is,

\[ \theta_j^{n+1} = \theta_j^n + \partial_t \theta_j^n \Delta t + \frac{1}{2} \partial_t^2 \theta_j^n \Delta t^2 + O(\Delta t^3), \]

\[ \theta_j^{n+1} = \theta_j^n \mp \partial_x \theta_j^n \Delta x + \frac{1}{2} \partial_x^2 \theta_j^n \Delta x^2 \pm O(\Delta x^3), \tag{5.66} \]

we get

\[ \partial_t \theta_j^n + \frac{1}{2} \partial_x^2 \theta_j^n \Delta t + O(\Delta t^2) = -u_0 \partial_x \theta_j^n + \frac{1}{4} |u_0| \partial_x^2 \theta_j^n \Delta x + O(\Delta x^2). \tag{5.67} \]

We note that by differentiating (5.2) we get \( \partial_t^2 \theta_j^n = u_0^2 \partial_x^2 \theta_j^n, \) and hence by rearranging terms that

\[ \partial_t \theta_j^n = -u_0 \partial_x \theta_j^n + \frac{1}{2} |u_0| (\Delta x - |u_0| \Delta t) \partial_x^2 \theta_j^n + O(\Delta x^2) + O(\Delta t^2). \tag{5.68} \]

Defining

\[ \kappa^* = \frac{1}{2} (1 - C)|u_0| \Delta x, \tag{5.69} \]

where \( C \) is the Courant number as defined in (5.30), (5.68) may be written

\[ \partial_t \theta_j^n = -u \partial_x \theta_j^n + \kappa^* \partial_x^2 \theta_j^n + O(\Delta x^2) + O(\Delta t^2). \tag{5.70} \]

Thus to second order in time and space we solve the following equation

\[ \partial_t \theta + u \partial_x \theta = \kappa^* \partial_x^2 \theta. \tag{5.71} \]

We recognize (5.71) as an advection-diffusion equation (cf. Chapter 3) with a diffusion coefficient \( \kappa^* \) given by (5.69). The terms of order \( O(\Delta x) \) and \( O(\Delta t) \) which we neglected when employing the upstream scheme therefore give rise to a diffusion. This diffusion is unphysical and an artifact that appears due to the numerical method used. It is therefore a numerical diffusion inherent in the scheme. The strength of the numerical diffusion is determined by the diffusion coefficient defined in (5.69). We note that the diffusion is insignificant if the Courant number is close to or equals one. This corresponds to the upper limit of the CFL criterion (5.30) for stability, and is associated with a near neutrally stable scheme. The diffusion term also goes to zero when \( \Delta x \) and \( \Delta t \) goes to zero. In that sense the upstream scheme is consistent.
5.12 Flux corrective schemes

In contrast to the second order leapfrog scheme the first order upwind (or upstream) scheme has the advantage that it is a positive definite scheme. Thus if the distribution of say \( \theta(x,t) \) at some arbitrary time \( t \) is such that \( \theta \geq 0 \) for all \( x \) then also \( \theta \geq 0 \) for all later times \( t = t + n\Delta t, \quad n = 1, 2, \ldots \). Another important property, as exemplified in Figure 5.5 on page 75, is that the position of the peak values are correctly propagated at any time without any dispersion. These are valuable properties well worth retaining in any scheme. The question arises if it is possible to construct a scheme that retains these properties while at the same time avoids, or at least minimizes, the numerical diffusion inherent in the scheme?

There are several schemes that offers a solution. Here we will present one of them called MPDATA\(^8\), a scheme first suggested by Smolarkiewicz (1983) (see also Smolarkiewicz and Margolin, 1997). The key element in MPDATA is to correct the diffusive flux inherent in upstream scheme. MPDATA therefore belongs to a class of schemes known as flux corrective schemes. To illustrate the method we first note that the advection equation (5.1) may be written

\[
\partial_t \theta + \partial_x F_A = 0, \tag{5.72}
\]

where \( F_A = u\theta \). As shown in Section 5.11 solving (5.72) using the upwind scheme results in a solution that satisfies (5.71) rather than (5.72). Thus, rather than solving (5.72) we appear to solve

\[
\partial_t \theta + \partial_x (F_A + F^*_D) = 0, \tag{5.73}
\]

where \( F^*_D = -\kappa^* \partial_x \theta \) and \( \kappa^* \) is given by (5.69). Thus the upwind scheme introduces an artificial or numerical diffusion represented by the diffusive flux \( F^*_D \). To avoid this unwanted diffusion Smolarkiewicz (1983) suggested to solve

\[
\partial_t \theta + \partial_x (F_A + F^*_A) = 0 \tag{5.74}
\]

rather than (5.72). Here \( F^*_A = u^* \theta \) is an artificially introduced advective flux where \( u^* \) is called the anti-diffusion velocity. The idea is to let the artificially introduced advective flux \( F^*_A \) exactly oppose the diffusive flux \( F^*_D = -\kappa^* \partial_x \theta \) introduced by the upwind scheme. We achieve this if we let \( F^*_A = -F^*_D \), which conveniently defines the anti-diffusive velocity by

\[
u^* \equiv \frac{\kappa^* \partial_x \theta}{\theta}. \tag{5.75}\]

To repeat, the idea of Smolarkiewicz (1983) is to introduce an artificial, advective flux that exactly compensate for the artificial diffusion generated by the first order upwind scheme. Theoretically adding the advective flux \( F^*_A = u^* \theta \) to the left-hand side of the continuous advection equation (5.72), as displayed in (5.74), exactly neutralizes the numerical diffusion arising when employing the upwind scheme. The introduction of the advective flux is therefore called a flux correction method.

---

\(^8\) MPDATA is an abbreviation of “Multiple Positive Definite Advection-Transport Algorithm”
5.12 Flux correction

We note that according to (5.75) $u^* = 0$ when $\partial_x \theta = 0$. Thus the propagation of the extreme values are not affected by adding the anti-diffusive velocity, or corrective flux. The positions of the extrema after time $t$ are therefore conserved. Moreover, we observe that $u^*$ is proportional to $\partial_x \theta$. It is therefore maximum where the gradient is steepest and its sign follows the sign of the slope. If for instance the distribution after some time is as displayed in Figure 5.5 on page 75, the effect of adding the artificial, advective flux is nil where the distribution has its maximum value. Thus the propagation of the maximum in the initial distribution is unaffected and its position is conserved. Moreover, we observe that $u^*$ is proportional to $\partial_x \theta$. It is therefore maximum where the gradient is steepest and its sign follows the sign of the slope. If for instance the distribution after some time is as displayed in Figure 5.5 on page 75, the effect of adding the artificial, advective flux is nil where the distribution has its maximum value. Thus the propagation of the maximum in the initial distribution is unaffected and its position is conserved. To the left of the “top” $\partial_x \theta > 0$. Hence, an artificial flux is added with just the right amount and sign to offset the numerical diffusive flux. To the right of the top $\partial_x \theta$ changes sign, and hence the anti-diffusive flux changes sign as well. Thus, as expected, the anti-diffusive flux helps to revoke the diffused gradients, regardless of the sign of the slope, and the enhancement is most effective where $\partial_x \theta$ is steepest. Moreover the anti-diffusive flux is just right to neutralize the artificial or numerical diffusive flux introduced when using the upwind scheme. Finally we observe that when computing the anti-diffusive velocity numerically we will have a problem if $\theta$ tends to zero faster than $\partial_x \theta$ in which case the anti-diffusive velocity $u^*$ tends to infinity. To avoid this Smolarkiewicz (1983) suggested to add a small number to the denominator in (5.75) when computing the anti-diffusive velocity.

The numerical implementation suggested by Smolarkiewicz (1983) is equally simple. He suggested to perform the compensation in two steps by using a so called predictor-corrector method. In the first step, the predictor step, we compute a prediction $\theta^*_j$ based on the true advection equation (5.72) using a low order advection algorithm\(^9\). Using the upwind scheme for this purpose we get

$$\theta^*_j = \theta^n_j - \left( F_A^n_{|j} - F_A^n_{|j-1} \right),$$

(5.76)

where

$$F_A^n_{|j} = \frac{1}{2} \left[ (u^n_j + |u^n_j|)\theta^n_j + (u^n_j - |u^n_j|)\theta^n_{j+1} \right] \Delta t \Delta x.$$  

(5.77)

We note that this step retains all the advantageous properties of the upwind scheme. We know, however, that the predictor solution $\theta^*_j$ is “infected” by a numerical diffusion which in its continuous form is represented by a diffusive flux that reads $F^*_D = -\kappa^* \partial_x \theta$, where $\kappa^*$ is given (5.69). This causes the prediction $\theta^*_j$ in general, and in particular for Courant numbers less than one, to appear less steep and generally smoother than its analytic or continuous counterpart. This is particularly evident in areas where the initial distribution has steep gradients as for instance visualized in Figure 5.5

In the second step, the corrector step, we solve the advection equation (5.74) without the original advection term, that is

$$\partial_t \theta + \partial_x F^*_A = 0,$$

(5.78)

again using the low order upwind scheme. Hence the corrected solution $\theta^{n+1}_j$ reads

$$\theta^{n+1}_j = \theta^*_j - \left( F^*_A_{|j} - F^*_A_{|j-1} \right),$$

(5.79)

where

$$F^*_A_{|j} = \frac{1}{2} \left[ (u^*_j + |u^*_j|)\theta^*_j + (u^*_j - |u^*_j|)\theta^*_j+1 \right] \Delta t \Delta x.$$  

(5.80)

\(^9\)A low order scheme is an algorithm with a truncation error of $O(\Delta t)$ and $O(\Delta x)$.
To compute the anti-diffusive velocity $u_j^*$ we may for instance use a centered scheme when computing the gradient $\partial_x \theta^*$. Thus using the analytic expression (5.75) we obtain

$$u_j^* = \frac{1}{4} (1 - C_{n,j}^n) |u_j^n| \left( \frac{\theta_{j+1}^* - \theta_{j-1}^*}{\theta_j^* + \epsilon} \right),$$

(5.81)

where $C_{n,j}^n = |u_j^n| \Delta t / \Delta x$ is the local Courant number. Note that we have added, as suggested by Smolarkiewicz (1983), a small number $\epsilon$ in the denominator to ensure that $u_j^*$ goes to zero when both $\theta_j^*$ and $\partial_x \theta_j^*$ goes to zero at the same time. If we make use of (5.81) to compute the anti-diffusive flux its maximum values correspond to areas where the predictor slopes are maximum. Thus in the second step the gradients are re-steeepened. Moreover, it has no effect where the predictor slopes are zero. Thus the position of the maximum is unchanged during the corrector step. As an example look at the dotted curve in Figure 5.5. If this was the predictive step the largest correction will be affected along the two flanks and thus steepen the diffused gradients. Note also that since the area under the curve is conserved when employing the upwind scheme, the maximum value increases during the corrector step. The solution therefore retains all the advantageous properties of the upwind scheme, and appears to avoid the artificial slackening of the steep gradients when applying the upwind scheme only. Moreover we observe that the corrector step makes the solution correct to $O(\Delta x^2)$ and $O(\Delta t^2)$. Hence MPDATA is a second order scheme that in theory compensates exactly for the artificial diffusive flux inherent in the lower order upwind scheme.

It should be noted, that since we employ an upwind scheme also in the corrector step the MPDATA data method also contains some artificial diffusion, however, second order smaller than the one using the upwind scheme only. This artificial diffusion may in turn be further minimized by rerunning the corrector step once more using the corrected solution as input. This is the beginning of an iterative procedure where the number of iterative steps are determined by the required accuracy only. A simpler and cheaper method (in terms of consumed computer time) is to slightly overestimate the anti-diffusive velocity by multiplying (5.81) by a scaling factor, a method already suggested by Smolarkiewicz (1983). Thus we redefine the anti-diffusive velocity to read

$$u_j^* = \frac{1}{4} S_c (1 - C_{n,j}^n) |u_j^n| \left( \frac{\theta_{j+1}^* - \theta_{j-1}^*}{\theta_j^* + \epsilon} \right),$$

(5.82)

where $S_c$ is the scaling factor. As an example Figure 5.6 shows the solution to a pure advection equation employing MPDATA, in which the initial distribution is a narrow Gaussian bell function. In the upper panel the scaling factor is set to one (no scaling), while in the lower panel a scaling factor $S_c = 1.3$ is used.
waves propagates away from the original splash point in such a way that that longer waves leads progressively shorter waves. The reason for this is that the phase speed, say $c$, depends on the wavenumber (or wavelength), that is, waves of different wavelengths propagate at different speeds. Regarding gravity waves in deep waters the longer the wavelength the faster the phase speed. Thus the longer waves will lead the progressively shorter waves. The same is also true for other types of waves for instance planetary Rossby waves.

Mathematically this is expressed through the dispersion relation $\omega = \omega(\alpha)$ where $\omega$ is the frequency and $\alpha$ is the wavenumber. Recall that the phase speed is

$$c = \frac{\omega}{\alpha}. \quad (5.83)$$

If the frequency is a linear function of $\alpha$, then the phase speed becomes a constant and all waves propagate at the same speed, that is, $\partial_\alpha c = 0$. The solution is then said to be non-dispersive. In the general case, however, $\omega$ is a non-linear function of the wavenumber $\alpha$. Then $\partial_\alpha c \neq 0$ and hence the phase speed depends on the wavelength. The solution is then said to be dispersive. Recall that the energy contained in the wave propagates with the group velocity defined by

$$c_g = \partial_\alpha \omega = \alpha \partial_\alpha c + c. \quad (5.84)$$

Hence if the wave is non-dispersive $c_g = c$, that is, the energy contained in the wave propagates at the same speed as the wave itself. On the other hand if the wave is dispersive then both the
phase speed and the group velocity depends on the wavelength. Moreover if $\partial_\alpha c < 0$, which is the case for gravity waves, then the waves travel at speeds faster than their group velocity. Thus gravity waves tend to travel faster than their energy is propagated\(^\text{10}\).

If we apply a wave solution to the advection equation (5.1), that is, $\theta = \Theta_0 e^{i\alpha(x-ct)}$, we find that the phase speed is $c = u$. Thus if $u = u_0$, that is, is uniform in time and space the phase speed becomes constant. Under these circumstances all the waves propagates with the same velocity, namely $u_0$, and hence the true solution to the advection equation is non-dispersive. The question then arises. Is this true for the numerical solution?

To investigate this we apply a similar analysis based on the finite difference approximation to the advection equation (5.1). Let us consider the leapfrog scheme (5.23). We know that this scheme is neutrally stable. Furthermore, as long as all the gradients are well resolved by our grid it is a superb scheme in the sense that it is a stable and consistent scheme with no numerical dissipation or diffusion. But what about the dispersiveness? To get started we first substitute the discrete Fourier component

$$\theta^n_j = \Theta_0 e^{i\alpha j(\Delta x - c_n \Delta t)}, \quad (5.85)$$

into (5.23). This gives

$$2i \sin(\alpha c \Delta t) = -2iu_0 \frac{\Delta t}{\Delta x} \sin(\alpha \Delta x) \quad (5.86)$$

which is written

$$c = \frac{1}{\alpha \Delta t} \arcsin \left( u_0 \frac{\Delta t}{\Delta x} \sin \alpha \Delta x \right). \quad (5.87)$$

Obviously $\partial_\alpha c \neq 0$, and hence the leapfrog scheme is dispersive. Since this dispersiveness is due to the finite difference approximation, that is, associated with our numerical solver, we commonly refer to this as numerical dispersion. We note that as $\Delta x$ and $\Delta t$ goes to zero, but such that the stability condition is satisfied, that is,

$$C \leq 1 \quad \text{or} \quad |u_0| \Delta t \leq \Delta x \quad (5.88)$$

then $c \to u_0$ and the wave becomes non-dispersive.

This is visualized in Figure 5.7 showing the normalized phase speed as a function of the normalized wave number for various Courant numbers. The figure clearly exhibit the dispersive nature of the leapfrog scheme. We also observe from Figure 5.7, as was first noted by Grotjhan and O’Brien (1976), that the dispersiveness gets worse the less the Courant number. In fact from (5.84) follows that $c_g$ becomes zero when

$$\alpha \partial_\alpha c = -c. \quad (5.89)$$

By use of (5.87) it follows that this is true for wavenumbers satisfying $\cos(\alpha \Delta x) = 0$, that is, for wavenumbers

$$\alpha_m = \frac{1}{2}(2m - 1)\pi, \quad m = 1, 2, \ldots \quad (5.90)$$

\(^{10}\)For capillary waves $\partial_\alpha c > 0$ and hence a capillary wave travels at a speed slower than its energy.
5.14 Unphysical solutions

Figure 5.7: Numerical dispersion for the leapfrog scheme. The curves depict the numerical phase speed as a function of the wavenumber based on (5.87) for various values of the Courant number $C = u \Delta t/\Delta x$. The vertical axis indicates the phase speed $c$ normalized by the advection speed $u$. The horizontal axis indicates the wavenumber normalized by $\pi/\Delta x$ where $\Delta x$ is the space increment or the grid size. The analytic dispersion curve is just a straight line corresponding to the phase speed $c = u$, that is $c/u = 1$. Note that as the wavenumber increases (that is the wavelength decreases) the numerical phase speed deviates more and more from the correct analytic phase speed for all values of the Courant number. For wave numbers which gives $\alpha \Delta x/\pi > 0.5$, that is for waves of wavelengths $\lambda < 4\Delta x$ the slope of the curves indicates that the group velocity is negative. Thus for waves with wavelengths shorter than $4\Delta x$ the energy is propagating in the opposite direction of the waves.

We therefore conclude that the longest wave for which $c_g = 0$ is for $m = 1$, that is $\lambda = 2\pi/\alpha_1 = 4\Delta x$. As displayed in Figure 5.7 this corresponds to the normalized wavenumber $\alpha' = 0.5$. For higher wavenumbers, that is, waves whose wavelength are shorter than $4\Delta x$ the group velocity actually becomes negative. Thus if the wave is poorly resolved the leapfrog scheme will actually propagate energy opposite to the wave itself. This is clearly unphysical and must be avoided.

It is therefore extremely important that the scales that dominates the property that is advected is well resolved. Let the dominant wavenumber be $\alpha$. Then by looking at Figure 5.7 $\Delta x$ must be chosen so that $\alpha \Delta x < 0.3\pi$ for Courant numbers close to one and even less if the Courant number is smaller.

5.14 Unphysical solutions and numerical modes

In Section 5.11 and Section 5.13 we showed that two of the most popular numerical schemes for solving the advection equation, namely the upwind scheme and the leapfrog scheme, harbored some unwanted properties. While the upwind or upstream scheme was shown to contain numer-
ical diffusion, the leapfrog scheme turned out to contain numerical dispersion. Although these properties, as their name suggests, have a physical interpretation they are nonetheless results of the employed scheme and hence unphysical or a numerical artifact. In Section 5.12 we showed how the numerical diffusion to some extent could be avoided by use of flux correction. Below (Section 10.5 on page 154) we will show that there exists a method called the spectral method which avoids the dispersion inherent in the leapfrog scheme. However, this method is applicable only for global models. For limited area models there is no such remedy available.

We will now show that the leapfrog scheme contains an additional unphysical property, namely false numerical modes. Let us start by assuming that the initial condition is as given in (5.5), that is, a harmonic wave with wave number \( \alpha \) and amplitude \( \theta_0 \). The true solution is then given by (5.6), that is, the solution is a single monochromatic wave with wavenumber \( \alpha \) propagating with the phase speed \( c = u_0 \) in the positive \( x \) direction.

To reveal that the leapfrog scheme actually contains two modes we solve the leapfrog scheme analytically. We start by assuming that the initial condition is a monochromatic wave with wavenumber \( \alpha \) as given by (5.5). Recalling from Section 4.2 (page 41) that any good function can be written in terms of an infinite sum of waves, we note that if we can find the solution for one monochromatic wave we find the solution to any arbitrary initial condition by summing up (in wavenumber space) over all possible wavenumbers. Furthermore we recall from the stability analysis of Section 4.3 on page 42 using induction that the solution may be written

\[
\theta^n_j = G^n \Theta_0 e^{i \alpha j \Delta x},
\]

(5.91)

where \( G \) is the growth factor and \( \Theta_0 \) is the initial amplitude. We underscore that \( G \) in general is a complex number. Recall that the growth factor for the leapfrog scheme actually has two solutions for the growth factor as given by (5.27) on page 64. For our purpose we rewrite these two solutions as

\[
G_1 = g_r - ig_i, \quad \text{and} \quad G_2 = -(g_r + ig_i),
\]

(5.92)

where the real and imaginary parts are given by

\[
g_r = \sqrt{1 - \lambda^2} \quad \text{and} \quad g_i = \lambda,
\]

(5.93)

respectively, and where \( \lambda \) is given by (5.26). Note that we have assumed that the CFL condition for stability is satisfied so that \( g_r \) is a real number. Recalling that any complex number \( P = a + ib \) may be written

\[
P = |P| e^{i \phi}
\]

(5.94)

where

\[
|P| = \sqrt{PP^*} = \sqrt{a^2 + b^2} \quad \text{and} \quad \phi = \arcsin \left( \frac{b}{|P|} \right),
\]

(5.95)

we note that \( |G_{1,2}| = 1 \) and that the two solutions become

\[
G_1 = e^{-i \alpha c \Delta t} \quad \text{and} \quad G_2 = (-1) e^{i \alpha c \Delta t}
\]

(5.96)

where \( c = c(\alpha) \) is the dispersive phase speed defined in (5.87). Substituting these solutions into (5.91) we obtain

\[
\theta^n_j = \Theta_1 e^{i \alpha (j \Delta x - cn \Delta t)} + (-1)^n \Theta_2 e^{i \alpha (j \Delta x + cn \Delta t)},
\]

(5.97)
where $\Theta_1$ and $\Theta_2$ are as yet two unknown constants. Since we have only one initial condition at our disposal, we can only obtain a relationship between them. Utilizing the initial condition (5.5) we find
\[ \theta_j^n = (\Theta_2 - \Theta_0)e^{i\alpha(j\Delta x - cn\Delta t)} + (-1)^n\Theta_2 e^{i\alpha(j\Delta x + cn\Delta t)}. \] (5.98)

We note that the first term on the right-hand side of (5.98) is a wave propagating in the positive $x$ direction with phase speed $c$ and amplitude $\Theta_2 - \Theta_0$. In contrast the second term on the right-hand side is a wave propagating in the negative $x$ direction, but with the same phase speed. Furthermore, the latter has an amplitude alternating between $\pm \Theta_2$. We conclude that by making use of the leapfrog scheme, the finite difference solution contains two solutions in the form of two waves propagating in opposite directions with the dispersive phase speed given by (5.87). In contrast the true solution (5.6) to the advection equation contains only one wave that propagates in the positive $x$ direction with phase speed $u_0$. The two waves that occur in the finite difference solution is due to the fact that the leapfrog scheme is of second order. As such it requires us to give two boundary conditions in time. We are therefore in need of an additional condition to determine the remaining constant.

The latter problem is associated with the initial boundary problem discussed in Section 5.8 above, but manifests itself in the unknown constant $\Theta_2$ in (5.98). One remedy suggested in Section 5.8 is to apply an Euler step as the first step. Note that this was suggested to start the time marching problem, otherwise we had to assign a value to $\theta$ at the time step prior to the initial time. We recall that the Euler step (5.54) is
\[ \theta_j^1 = \theta_j^0 - u_0 \frac{\Delta t}{2\Delta x} (\theta_{j+1}^0 - \theta_{j-1}^0). \] (5.99)

Substituting the initial condition (5.5) into (5.99) and (5.98) (letting $n = 1$) the additional condition becomes
\[ \Theta_0[1 - i \sin(\alpha c \Delta t)]e^{i\alpha j\Delta x} = (\Theta_2 - \Theta_0)e^{-i\alpha c \Delta t} - \Theta_2 e^{i\alpha c \Delta t} \] (5.100)
and hence that
\[ \Theta_2 = \Theta_0 \frac{1 - \cos(\alpha c \Delta t)}{2 \cos(\alpha c \Delta t)}. \] (5.101)

The complete finite difference solution is then
\[ \theta_j^n = \Theta_0 \left\{ \frac{1 + \cos(\alpha c \Delta t)}{2 \cos(\alpha c \Delta t)} e^{i\alpha(j\Delta x - cn\Delta t)} + (-1)^{n+1}\frac{1 - \cos(\alpha c \Delta t)}{2 \cos(\alpha c \Delta t)} e^{i\alpha(j\Delta x + cn\Delta t)} \right\}. \] (5.102)

We observe, as mentioned in the previous Section 5.13, that as long as the stability criteria (5.30) is satisfied then $c \to u_0$ when $\Delta t$ and $\Delta x$ goes to zero independently. Under these circumstances we take note that $[1 + \cos(\alpha c \Delta t)]/2 \cos(\alpha c \Delta t) \to 1$ and $(1 - \cos(\alpha c \Delta t))/2 \cos(\alpha c \Delta t) \to 0$. Thus as we allow $\Delta t \to 0$ and $\Delta x \to 0$ the first term on the right-hand side of (5.102) approaches the true solution while the second term vanishes. We conclude therefore that the second

\*\*11 Note that from (5.86) follows $\lambda = u_0 \frac{\Delta t}{\Delta x} \sin(\alpha \Delta x) = \sin(\alpha c \Delta t)$ \*\*
term is an unphysical mode while the first term is the physical mode. The occurrence of the two
modes in the leapfrog scheme are sometimes referred to as “time splitting” in the literature.

It is of utmost importance that we do control the unphysical mode. Unless we do it may
create a lot of noise in our calculations, or in extreme cases even numerical instabilities. This is
particularly true since it alternates between attaining positive and negative values.

5.15 How to get rid of the unphysical mode: The Asselin filter

The simplest way to get rid of the numerical or unphysical mode in the leapfrog scheme is, from
time to time, to make use of a one-sided, forward in time, centered in space step from time to time
as outlined in (5.16), that is, perform an Euler step. Although, as shown in Section 5.1, the Euler
scheme is numerically unstable, we may still apply it for single time steps without destroying the
stability of for instance the leapfrog scheme.

Another method, originally suggested by Robert (1966) and further developed by Asselin
(1972), is to apply a time filtering technique. That a time filter will work the trick is obvious
since the unphysical mode alternates between positive and negative values from time step to time
step. We start by remarking that in general a time filter invoking the neighboring time levels only
may be written

\[
\theta(x,t) = \gamma \theta(x,t + \Delta t) + (1 - 2\gamma) \theta(x,t) + \gamma \theta(x,t - \Delta t),
\]  

(5.103)

where \(\theta(x,t)\) is the filtered function and \(\gamma\) is a weighting parameter. By use of the notation
introduced in Section 2.9 we may rewrite (5.103) to yield

\[
\theta^n_j = \theta^n_j + \gamma \theta^{n+1}_j - 2\theta^n_j + \theta^{n-1}_j).
\]  

(5.104)

Note that if \(\gamma = 0\) we retrieve the original function, while for \(\gamma = \frac{1}{4}\) the filter is the standard
1-2-1 filter,

\[
\theta^n_j = \frac{1}{2} \theta^{n+1}_j + 2\theta^n_j + \theta^{n-1}_j,
\]  

(5.105)

which gives twice the weight to the mid time level \(n\). To investigate the properties of the filter
let us study one period or frequency \(\omega\) only. We start by representing the function \(\theta\) by it Fourier
component

\[
\theta^n_j = \hat{\theta}_j e^{i\omega n \Delta t}.
\]  

(5.106)

Substituting this into (5.104) gives

\[
\hat{\theta}^n_j = R \hat{\theta}^n_j,
\]  

(5.107)

where the ratio

\[
R(\gamma) = \frac{\theta^n_j}{\theta^n_j} = 1 - 2\gamma + 2\gamma \cos \omega \Delta t
\]  

(5.108)

is the response function of the filter. For the 1-2-1 filter, that is, for \(\gamma = \frac{1}{4}\), the response function
becomes

\[
R(\frac{1}{4}) = \frac{1}{2} (1 + \cos \omega \Delta t)
\]  

(5.109)
5.15 The Asselin filter

Since the period is \( T = \frac{2\pi}{\omega} \) we notice that for the standard 1-2-1 filter \( R = 0 \) for waves of period \( T = 2\Delta t \), while \( R = \frac{1}{2} \) for waves of period \( T = 4\Delta t \). Thus waves or noise on the Nyquist frequency, that is, waves of periods \( 2\Delta t \), cannot exist and hence vanish. Since the unphysical mode inherent in the leapfrog scheme alternates between attaining negative and positive values from time step to time step, its dominate waveperiod is exactly \( 2\Delta t \). Hence the 1-2-1 filter is perfect in this sense. Noise of scales close to slightly longer periods, say \( 4\Delta t \), is damped to half of their original energy. The effect on the longer periods are minimal. It was these advantageous properties of the the 1-2-1 filter that lead Robert (1966) and Asselin (1972) to suggest to use this method to damp the unphysical mode inherent in the leapfrog scheme.

In practice we do this as follows. Let us assume that the filtered solution have been determined for time level \( n - 1 \), that is, assume that \( \overline{\theta}^{n-1} \) is known and has been stored. Let us furthermore assume that the unfiltered value at time level \( n \), that is, \( \theta^n \), has been stored as well. We then first apply the leapfrog scheme to compute the function \( \theta \) at the new time level \( n + 1 \), that is, \( \theta^{n+1} \), using the unfiltered values of \( \theta \) at time level \( n \) and the filtered values of \( \theta \) at time level \( n - 1 \). For instance, if we consider the simple advection equation (5.1) and make use of the leapfrog scheme (5.23) to advance to the next time level we obtain

\[
\theta^{n+1}_j = \theta^{n-1}_j - u \frac{\Delta t}{\Delta x} (\theta^n_{j+1} - \theta^n_{j-1}).
\]  

(5.110)

Note that we in this first step have used the filtered value at time level \( n - 1 \) to compute \( \theta^{n+1} \). The next step is to compute the filtered values at time level \( n \) using the filter (5.104), that is,

\[
\overline{\theta}^n_j = \theta^n_j + \gamma \left[ \theta^{n+1}_j - 2\theta^n_j + \overline{\theta}^{n-1}_j \right].
\]  

(5.111)

In the third and final step we replace the filtered values at time level \( n - 1 \) by the new filtered values at time level \( n \). We can safely do this because after we have computed the new filtered values at time level \( n \) the filtered values at time level \( n - 1 \) are obsolete. We may now proceed to the next time level to compute \( \theta^{n+2} \), and so on.

We take note, however, that the Asselin filter affects the numerical stability and that it produces numerical diffusion. It is possible to show that while the numerical diffusion increases with increasing values of the weighing parameter \( \gamma \) the critical value for stability decreases. The latter implies that the stability criterion becomes more strict and that we have to diminish the time step \( \Delta t \). This fact entails that although we wish to employ the simple 1-2-1 filter since it exactly kills the unphysical mode, it becomes unstable unless we decrease the time step. Decreasing the time step in turn means that our computations becomes less efficient. It is therefore common to apply a lower value for the weighing function, say \( \gamma = 0.08 \). Note that even a weak Asselin filter eventually modifies the longer wave periods by diffusion. Hence we must apply the Asselin filter with care and not necessarily for every time step.

Exercises

1. Show that the true solution to (5.2) is indeed (5.3). Hint: Make use of Fourier series.
2. Show that the CFL criterion for the leapfrog scheme, the diffusive scheme and the upwind scheme all are given by (5.30).

   Hint: Express the growth function in terms of $G = \sqrt{1 - (1 - C)f}$ where $C = |u_0|\Delta t/\Delta x$ is the Courant number and $f = f(C, \alpha, \Delta x)$.

3. Let us assume that the Courant number equals one. Show that under these circumstances the upstream scheme has no truncation errors. Use the method of characteristics to illustrate why this is the case.

4. Show that (5.97) is a solution to (5.23). Moreover, show that (5.102) follows from (5.97) when the initial distribution is given by (5.99), and where (5.16) is made use of to find $\theta_j^{-1}$.

   Hint: Show first that $G_{1,2}$ from (5.27) may be written

   $G_1 = e^{-ix}, G_2 = e^{i(x+\pi)}$  \hspace{1cm} (5.112)

   where $x$ is given by (5.98).
Chapter 6

THE SHALLOW WATER PROBLEM

Since advection and mixing (diffusion) are the foremost processes by which tracers are transported and spread in the atmosphere and ocean, we maintain that they are two of the most fundamental and important balances to treat correctly when solving the governing equations numerically. Most commonly the two processes are combined in the so called advection-diffusion equation, an equation treated in Chapter 3. Inherent in this equation is the velocity by which the tracer is advected. In the preceding chapter (Chapter 5) we assumed this velocity to be a known function and mostly treated it as a constant. In reality the velocity is part of the dynamics of the atmosphere-ocean system, and hence a function of time and space. The prognostic equation from which it is determined is the momentum equation (1.1) presented on page 2 in Chapter 1. Accordingly we need to get insight into how to solve the momentum equations numerically in addition to the advection-diffusion equation.

As alluded to in Chapter 3 an essential element of the atmosphere-ocean dynamics, that makes it stand out from ordinary fluid dynamics, is the effect of the Earth’s rotation giving rise to the so called Coriolis force\(^1\). The appearance of this latter force makes it possible to obtain time-independent solutions to the momentum equation that differs from the trivial state of rest. This stationary solution is made possible by balancing the Coriolis and the pressure force as displayed in (1.39) on page 9 and we commonly refer to it as the geostrophic balance.

Consequently we devote this chapter to methods whereby we may determine the advection velocity by solving the momentum equation numerically. As in the preceding chapters we continue to follow Albert Einstein’s mantra of making things as simple as possible, but no simpler. Thus we turn our attention to the shallow water equations derived in Section 1.5 on page 7. The equations themselves, as presented in (1.23) - (1.24) or their depth integrated versions (1.30) and (1.31) on page 8, are indeed simple, yet they include the essence of atmosphere-ocean dynamics. Most importantly they retain the possibility of a geostrophic balance. We note that the momentum equation (1.30) involves two unknowns, namely the the horizontal velocity \(u\) and the height

\(^1\)Gaspard-Gustave de Coriolis or Gustave Coriolis (21 May 1792 - 19 September 1843) was a French mathematician, mechanical engineer and scientist. He is best known for his work on the supplementary forces that are detected in a rotating frame of reference, and one of those forces nowadays bears his name. (Source: Wikipedia). Note that this force is virtual in the sense that its presence is caused by our choice of coordinate system, namely one fixed to the rotating Earth.
of a fluid column (or pressure). Thus we need the continuity equation (1.31) to close the system.

Even though the shallow water equations are simple, and contains a much reduced momentum equation, they are complex enough to let us appreciate the methods whereby the governing equations, including the momentum equation, are solved numerically. One of the reasons for this is that the fully three-dimensional, barotropic/baroclinic equations of motion can be described in terms of so called vertical normal modes, where each mode is governed by a set of shallow water equations. Another way of illustrating this is to discretize a numerical model into say $N$ vertical layers where the density is constant within each layer. For each layer we then get a set of shallow water equations to solve that are coupled to the other layers through the pressure forcing and through the vertical mixing term. In either case we end up with a set of $N$ shallow water equations to solve. Each of them has a so called “equivalent depth” (or equivalent geopotential height) corresponding roughly to the height of the coordinate surface above ground/bottom. Note that the shallow water equations as presented in (1.30) and (1.31) correspond to a one layer model in which the surface is the Lagrangian (movable) vertical coordinate. They also represent the first and foremost vertical normal mode in the normal mode approach. To obtain the effect of baroclinicity we may introduce a second layer or a second normal mode. The interface between the two layers is then the second movable Lagrangian vertical coordinate.

In this Chapter we therefore study the shallow water equations as given in (1.33) and (1.34). For a one-layer model of uniform density $\rho_0$ they read,

$$\partial_t \mathbf{u} + \mathbf{u} \cdot \nabla H \mathbf{u} + f \mathbf{k} \times \mathbf{u} = -\nabla H \phi + \frac{\tau_s - \tau_b}{\rho_0 h} + \frac{\mathbf{X}}{h}, \quad (6.1)$$

$$\partial_t \phi + \mathbf{u} \cdot \nabla H \phi = -\phi \nabla H \cdot \mathbf{u}, \quad (6.2)$$

where $\phi = gh = g(H + \eta)$ is the geopotential and $h$ is the geopotential height. In a one layer model like the present one the geopotential height is simply the depth (or height) of fluid column where $H = H(x, y)$ is the equilibrium depth and $\eta = \eta(x, y, t)$ the deviation of the top surface away from a reference geopotential level as illustrated in Figure 6.1. Furthermore $\tau_s$ is the stress at the top surface (in the ocean referred to as the wind stress, that is, the traction the atmosphere exercises on the ocean surface), while $\tau_b$ is the stress at the bottom, e.g., friction. We note that the wind stress then acts as an energy source, while the bottom friction acts as an energy sink, that is, dissipates energy. Finally, $\mathbf{X}$ contains lateral (horizontal) viscosity arising from lateral viscous processes, e.g., momentum diffusion (commonly referred to as eddy viscosity). Thus the eddy viscosity acts to diffuse or even out small scale motion in the atmosphere and ocean. Thus it is no different from the diffusion of tracers as treated in Chapter 4.

The set (6.1) and (6.2), consisting of the two components of the horizontal momentum equation and the continuity equation (conservation of mass), highlights the importance of geostrophy, and is a particularly useful and simple yet complicated enough set of equations to simulate many dynamical processes active in the atmosphere and ocean. Equally important, they allow us to

---

2 An elegant derivation is given in the Appendix to Lighthill (1969)

3 Note that the layers may move up and down in the vertical. The interfaces between the layers are therefore movable and forms a vertical Lagrangian coordinate system. We will return to this in Chapter 8 on page 127 where we treat various vertical coordinate systems.

4 This was first introduced by Jules Charney in his fundamental paper published in 1955 (Charney, 1955)
introduce methods whereby the momentum and continuity equations may be solved by numerical means, without having to concern ourselves with the thermodynamic processes. It should be stressed though that thermodynamic processes are indeed important processes in the atmosphere as well as the ocean. Their influence on the dynamics is through the active tracers like potential temperature, humidity and salinity. These tracers are in turn determined by the tracer equations, that is, by advection-diffusion equations treated in the previous chapters (Chapters 4 and 5).

In addition we note that the set (6.1) and (6.2) contains multiple dependent variables, that is, \( \phi, u, \) and \( v \), and hence conveniently introduces methods whereby partial differential equations containing more than one dependent variable may be solved numerically. Moreover, as is evident from (6.1) and (6.2), the shallow water equations is a set of non-linear equations. Since there is a fundamental difference between linear and non-linear systems, we split the presentation below into linear and non-linear versions of the shallow water equations. Furthermore, to highlight some salient fact of importance when solving the shallow water equations by numerical means, we also separate between non-rotating and rotating systems.

Since the mixing and friction terms are energy source and/or sink terms we may, without loss of generality, neglect these terms, as we for instance did in Section 3.4. Under these circumstances the set (6.1) and (6.2) becomes

\[
\begin{align*}
\partial_t u + u \cdot \nabla_H u + f k \times u &= -\nabla_H \phi, \\
\partial_t \phi + u \cdot \nabla_H \phi &= -\phi \nabla_H \cdot u.
\end{align*}
\]

We observe that (6.3) and (6.4) are non-linear, and that they contain three dependent variables, namely the two horizontal, depth integrated, velocity components \( u, v \) and the geopotential \( \phi \). We also notice that the set (6.3) and (6.4) is complete in the sense that we have three equations to solve for the three unknowns \( u, v, \phi \). By the same token we observe that to find a solution for the unknowns we need three initial conditions and six boundary conditions to determine the
integration constants. Finally, we note that the set (6.3) and (6.4) contains three independent variables, namely \(x, y\) and \(t\).

To repeat, the shallow water equations are not only a set of equations that describe important and fundamental aspects of the atmosphere-ocean dynamics. It is also a convenient set of equations whereby the numerical treatment of equations containing more than one dependent variable as well as several independent variables may be introduced within a geophysical fluid context.

### 6.1 Linearization of the shallow water equations

To linearize the system (6.3) - (6.4) we start by assuming that the dependent variables can be written in terms of a basic state plus a perturbation, that is,

\[
u = \bar{u} + u' \quad \text{and} \quad \phi = \bar{\phi} + \phi' \tag{6.5}\]

where the perturbed velocity (or basic state) is \(\vec{u} = \bar{u} + \vec{v}\) and the perturbed geopotential is \(\bar{\phi}\). Note that both \(\bar{u}\) and \(\bar{\phi}\) may be functions of the independent variables \(x, y\) and \(t\). Substituting these expressions into (6.3) - (6.4) we get

\[
\begin{align*}
\partial_t \bar{u} + \partial_t u' &+ \bar{u} \cdot (\nabla_H \bar{u} + \nabla_H u') + u' \cdot (\nabla_H \bar{u} + \nabla_H u') \\
- f k \times \bar{u} - f k \times u' &= -\nabla_H \bar{\phi} - \nabla_H \phi', \tag{6.6} \\
\partial_t \bar{\phi} + \partial_t \phi' &+ \bar{u} \cdot (\nabla_H \bar{\phi} + \nabla_H \phi') + u' \cdot (\nabla_H \bar{\phi} + \nabla_H \phi') \\
&+ \bar{\phi}(\nabla_H \cdot \bar{u} + \nabla_H \cdot u') + \phi'(\nabla_H \cdot \bar{u} + \nabla_H \cdot u') = 0 \tag{6.7}
\end{align*}
\]

By nature a perturbation is defined as being “small”. Here we define small to imply that products of the perturbations can be dropped in comparison with the perturbation itself, that is,

\[
|u'^2| \ll |u'|, \quad |\phi'^2| \ll |\phi'|. \tag{6.8}
\]

Thus (6.6) and (6.7) reduces to

\[
\begin{align*}
\partial_t \bar{u} + \partial_t u' &+ \bar{u} \cdot (\nabla_H \bar{u} + \nabla_H u') + u' \cdot \nabla_H \bar{u} \\
- f k \times \bar{u} - f k \times u' &= -\nabla_H \bar{\phi} - \nabla_H \phi', \tag{6.9} \\
\partial_t \bar{\phi} + \partial_t \phi' &+ \bar{u} \cdot (\nabla_H \bar{\phi} + \nabla_H \phi') + u' \cdot \nabla_H \bar{\phi} \\
&+ \bar{\phi}(\nabla_H \cdot \bar{u} + \nabla_H \cdot u') + \phi'\nabla_H \cdot \bar{u} = 0 \tag{6.10}
\end{align*}
\]

By the same token the basic state is assumed to be slowly varying in time and space implying that time and space derivatives of the basic state are small with respect to the derivatives of the perturbations. Hence

\[
\begin{align*}
|\partial_t \bar{u}| &\ll |\partial_t u'|, \quad |\nabla_H \bar{u}| \ll |\nabla_H u'|, \quad |\nabla_H \cdot \bar{u}| \ll |\nabla_H \cdot u'|, \\
|\partial_t \phi| &\ll |\partial_t \phi'|, \quad |\nabla_H \bar{\phi}| \ll |\nabla_H \phi'|. \tag{6.11}
\end{align*}
\]
Invoking these inequalities in (6.9) and (6.10) we get

\[ \partial_t \mathbf{u} + \bar{\mathbf{u}} \cdot \nabla_H \mathbf{u} + f \mathbf{k} \times \mathbf{u} + \nabla_H \phi = -f \mathbf{k} \times \bar{\mathbf{u}} - \nabla_H \tilde{\phi}, \]

\[ \partial_t \phi + \bar{\mathbf{u}} \cdot \nabla_H \phi + \bar{\phi} \nabla_H \cdot \mathbf{u} = 0, \]

where we have dropped the primes on the perturbation quantities for clarity. The two equations (6.12) and (6.13) thus describe the time evolution of the perturbations away from the specified basic state.

Note that we have kept the term \( \nabla_H \tilde{\phi} \) on the right-hand side of (6.12) even though the inequalities in (6.11) encourage us to neglect it compared to \( |\nabla_H \phi'| \). The reason is the effect of the Earth’s rotation. To be consistent we do require that the basic state is in dynamical balance. This implies that the basic state must be a solution to our equation when the perturbations are zero. Thus from (6.6) and (6.7) follows that the basic state must be the solution to the set

\[ \partial_t \bar{\mathbf{u}} + \bar{\mathbf{u}} \cdot \nabla_H \bar{\mathbf{u}} - f \mathbf{k} \times \bar{\mathbf{u}} = -\nabla_H \tilde{\phi}, \]

\[ \partial_t \tilde{\phi} + \bar{\mathbf{u}} \cdot \nabla_H \tilde{\phi} + \tilde{\phi} \nabla_H \cdot \bar{\mathbf{u}} = 0. \]

If we assume that the non-linearity of the basic state is small, that is, require the Rossby number to be small (cf. Section 1.6 on page 9), it follows that the basic state must be in geostrophic balance, that is, must satisfy the equation

\[ \bar{\mathbf{u}} = \frac{1}{f} \mathbf{k} \times \nabla_H \tilde{\phi}. \] (6.16)

Under these circumstances the right-hand side of (6.12) is zero, while allowing a basic state that is different from the trivial state of rest, that is, \( \bar{\mathbf{u}} \neq 0 \). Hence the second terms on the right-hand side of (6.14) and (6.15) must be kept, and the set (6.12) and (6.13) finally becomes,

\[ \partial_t \mathbf{u} + \bar{\mathbf{u}} \cdot \nabla_H \mathbf{u} + f \mathbf{k} \times \mathbf{u} + \nabla_H \phi = 0 \]

\[ \partial_t \phi + \bar{\mathbf{u}} \cdot \nabla_H \phi + \tilde{\phi} \nabla_H \cdot \mathbf{u} = 0. \]

(6.17)

(6.18)

We note that the possibility of having a basic state that differs from a state of rest, does not exclude it. If we indeed assume that the basic state is one at rest (\( \bar{\mathbf{u}} = 0 \)) then (6.16) requires \( \tilde{\phi} \) to be constant, say \( \tilde{\phi} = \phi_0 = gH_0 \), where and \( H_0 \) is a constant geopotential height. Substituting this simple basic state into (6.17) and (6.18) we arrive at a particular simple linear subset, namely,

\[ \partial_t \mathbf{u} + f \mathbf{k} \times \mathbf{u} = -\nabla_H \phi, \]

\[ \partial_t \phi + \phi_0^2 \nabla_H \cdot \mathbf{u} = 0, \]

where \( \phi_0 = \sqrt{gH_0} \) is known as the phase speed (propagation speed) of inertia-gravity waves.

Let us for a moment in addition assume that the time rate of change of the geopotential in (6.20) is so small that the motion is effectively divergence free, that is, \( \nabla_H \cdot \mathbf{u} = 0 \). We may then introduce a streamfunction \( \psi \) by defining

\[ \mathbf{u} = \mathbf{k} \times \nabla_H \psi, \quad \text{or} \quad u = -\partial_y \psi \quad \text{and} \quad v = \partial_x \psi. \] (6.21)
If we then use the operator $k \cdot \nabla_H \times$ on (6.19), remembering that the Coriolis parameter is a function of the latitude $y$, we get

$$\partial_t \zeta + \beta \partial_x \psi = 0, \quad (6.22)$$

where $\zeta = k \cdot \nabla_H \times u = \nabla_H^2 \psi$ is the vorticity, $\beta = \partial_y f$ is the rate of change of the Coriolis parameter with latitude, and where we have made use of (6.21) and that $\nabla_H \cdot u = 0$. We note that (6.22) is the barotropic vorticity equation whose solutions contains the Rossby waves. In fact if substitute for a Fourier component, that is, let

$$\psi = \psi_0 e^{i(\alpha x + ly - \omega t)}, \quad (6.23)$$

the we arrive at the well known dispersion relation for Rossby waves, viz.

$$c = \frac{\omega}{\alpha} = -\frac{\beta}{(\alpha^2 + l^2)}. \quad (6.24)$$

As in the preceding chapters we further assume that the problem is one-dimensional. This is achieved by letting $\partial_y = 0$. We also assume that motion is such that $\partial_y f = 0$, normally referred to as the $f$-plane approximation. Note that this assumption exclude Rossby waves to be part of the solution. The set (6.17) and (6.18) then becomes

$$\begin{align*}
\partial_t u + \bar{u} \partial_x u - f v + \partial_x \phi &= 0, \\
\partial_t v + \bar{u} \partial_x v + f u &= 0, \\
\partial_t \phi + \bar{u} \partial_x \phi + \bar{\phi} \partial_x u &= 0.
\end{align*} \quad (6.25-6.27)$$

while the subset (6.19) and (6.20) becomes

$$\begin{align*}
\partial_t u - f v + \partial_x \phi &= 0, \\
\partial_t v + f u &= 0, \\
\partial_t \phi + \bar{c}_0^2 \partial_x u &= 0.
\end{align*} \quad (6.28-6.30)$$

For later reference we note that the geostrophic velocity components $\bar{u}$ and $\bar{v}$ are given by

$$\bar{u} = -\frac{1}{f} \partial_y \bar{\phi}, \quad \text{and} \quad \bar{v} = \frac{1}{f} \partial_x \bar{\phi}, \quad (6.31)$$

respectively. Thus if we let the assumption $\partial_y = 0$ also apply to the basic state then $\bar{u} = 0$. Under these circumstances the second terms on the left-hand side of (6.25) and (6.26) vanish as well. However, we require merely that the perturbations $u$, $v$, and $\phi$ are independent of $y$. Thus to be consistent we must require that $\bar{u} = \bar{u}(x,t)$, that is, that $\partial_y \bar{u} = 0$. From (6.31) we note that this implies that $\bar{\phi}$ must be a linear function of $y$.

### 6.2 Linear, non-rotating shallow water equations

We will first investigate possible numerical methods to solve the one-dimensional, linear shallow water equations (6.25) - (6.27). We first note that by neglecting rotation we cannot assume a
THE SHALLOW WATER PROBLEM

6.2 Linear, non-rotating

basic state in geostrophic balance. Hence $\bar{u} = \bar{v} = 0$, and thus we are left with (6.28) - (6.30) as our governing equations. We note that under these circumstances (6.29) gives $v = \text{constant} = 0$ and hence that (6.29) becomes obsolete. Thus the motion is described by,

$$
\begin{align*}
\partial_t u &= -\partial_x \phi, \\
\partial_t \phi &= -c_0^2 \partial_x u.
\end{align*}
$$

(6.32)

(6.33)

Thus we are left with two equations for the two unknowns $u$ and $\phi$. We observe that the system is hyperbolic. This is for instance easily shown if we substitute (6.32) into (6.33). We then get

$$
\partial_t^2 \phi - c_0^2 \partial_x^2 \phi = 0,
$$

(6.34)

which is the linear wave equation.

To explore possible solutions to (6.32) and (6.33) we investigate one Fourier component. Recall that the full solution may be represented is an infinite sum over all wavelengths in which each Fourier component has its own amplitude. Thus we assume that the solution may be written in forms of waves, that is,

$$
\phi = \phi_0 e^{i\alpha(x-ct)} \quad \text{and} \quad u = u_0 e^{i\alpha(x-ct)}
$$

(6.35)

where $u_0$ and $\phi_0$ are arbitrary wave amplitudes and $c = \omega/\alpha$ is the wave speed. Substituting (6.35) into (6.32) and (6.33) and tossing away common non-zero factors we get

$$
\begin{align*}
-cu_0 + \phi_0 &= 0, \\
c_0^2 u_0 - c\phi_0 &= 0.
\end{align*}
$$

(6.36)

(6.37)

Solving with respect to $u_0$ or $\phi_0$ then gives the dispersion relation

$$
c^2 - c_0^2 = 0.
$$

(6.38)

This equation has two solutions for the wave speed $c$, namely

$$
c_{1,2} = \pm c_0.
$$

(6.39)

Thus the true solution for say the geopotential is

$$
\phi = \phi_1 e^{i\alpha(x-c_0 t)} + \phi_2 e^{i\alpha(x+c_0 t)},
$$

(6.40)

that is, two ordinary gravity waves propagating with a constant phase speed $c_0$ in opposite directions.\(^5\)

---

\(^5\)Note that $u$ has a similar solution
Finite difference form

Our task is to solve (6.32) and (6.33) numerically to give a solution that mimics the true solution. As alluded to in Section 2.4 the advection equation bears a strong resemblance to the wave equation and belongs to the class of hyperbolic partial differential equations. We therefore expect the leapfrog scheme, that is, a centered in time, centered in space (CTCS) scheme to work well. Using centered finite difference approximations for the differential terms we get

\[ \frac{u_{j}^{n+1} - u_{j}^{n-1}}{2\Delta t} + \frac{\phi_{j+1}^{n} - \phi_{j-1}^{n}}{2\Delta x} = 0, \]  
\[ \frac{\phi_{j}^{n+1} - \phi_{j}^{n-1}}{2\Delta t} + c_{0}^{2} \frac{u_{j+1}^{n} - u_{j-1}^{n}}{2\Delta x} = 0. \]  

(6.41)  
(6.42)

Since we made use of Taylor series to derive the approximations, we know a priori that the scheme is consistent. To satisfy ourselves that the system is numerically stable we use von Neumann’s analysis method. Thus we start by substituting the discrete Fourier components

\[ u_{j}^{n} = U_{n} e^{i\alpha j \Delta x} \]  
\[ \phi_{j}^{n} = \Phi_{n} e^{i\alpha j \Delta x} \]  

(6.43)

into (6.41) and (6.42). We then get

\[ U_{n+1} - U_{n-1} = -2i\gamma \Phi_{n}, \]  
\[ \Phi_{n+1} - \Phi_{n-1} = -2i\gamma c_{0}^{2} U_{n}, \]  

(6.44)  
(6.45)

where

\[ \gamma = \frac{\Delta t}{\Delta x} \sin \alpha \Delta x. \]  

(6.46)

Since \( n \) is only a time step counter we may rewrite (6.44) to give

\[ U_{n+2} - U_{n} = -2i\gamma \Phi_{n+1}, \]  

(6.47)

and similarly

\[ U_{n} - U_{n-2} = -2i\gamma \Phi_{n-1}. \]  

(6.48)

We then subtract (6.48) from (6.47) to get

\[ U_{n+2} - 2U_{n} + U_{n-2} = -2i\gamma (\Phi_{n+1} - \Phi_{n-1}). \]  

(6.49)

Finally we substitute for \( (\Phi_{n+1} - \Phi_{n-1}) \) using (6.45) to get

\[ U_{n+2} - 2\lambda U_{n} + U_{n-2} = 0, \quad \lambda = 1 - 2c_{0}^{2} \gamma^{2}. \]  

(6.50)

Next we define a growth factor \( G \) by \( U_{n+2} \equiv GU_{n} \). Substitution into (6.50) then results in a second order equation to solve for the growth factor. Solving it we get the two solutions

\[ G_{1,2} = -\lambda \pm \sqrt{\lambda^{2} - 1} = -\lambda \pm i\sqrt{1 - \lambda^{2}}. \]  

(6.51)
Thus the growth factor is complex and has an imaginary part if and only if \( \lambda^2 \leq 1 \). Under these circumstances
\[
|G_{1,2}| = \sqrt{\lambda^2 + 1 - \lambda^2} = 1,
\]
and the scheme is neutrally stable as expected employing a CTCS scheme to solve for a hyperbolic system. Recall that this result depends on the fact that the radical in (6.51) is positive definite. Thus stability is ensured if and only if \(-1 \leq \lambda \leq 1\). The right-hand inequality is trivially satisfied, while the left-hand inequality requires
\[
c_0^2 \left( \frac{\Delta t}{\Delta x} \right)^2 \leq 1 \quad \Rightarrow \quad C = c_0 \frac{\Delta t}{\Delta x} \leq 1, \quad \text{or} \quad \Delta t \leq \frac{\Delta x}{c_0}
\]
where \( C \) is the Courant number. Thus we arrive at the result that the scheme is neutrally stable (contains no numerical dissipation) under the condition that the Courant number is less than or equal to one. We observe, however, that the velocity that enters the definition of the Courant number is the phase speed \( c_0 \) rather than the advection velocity \( u \). Since the phase speed normally is much larger than the advection speed, that is, \( c_0 \gg |u| \), it follows that the constraint on the time step is much more stringent for the momentum equation than for the tracer equation.

### 6.3 Staggered grids

We note that the system (6.32) and (6.33) contains four integration constants, namely, two in time and two in space. The system therefore allows us to specify two boundary conditions in \( x \), no more no fewer. If we were to solve (6.32) and (6.33) for \( x \in (-L, L) \) and for \( t \in (-T, T) \), we then have three options. The first option is to specify one condition at \( x = 0 \) and one condition at \( x = L \). The second is to specify two conditions at \( x = 0 \), while the third option is to specify two conditions at \( x = L \). Let us assume that there are impermeable walls at \( x = 0 \) and \( x = L \). Then the natural condition is no flow through them, that is,
\[
|u|_{x=0} = 0 \quad \text{and} \quad |u|_{x=L} = 0 \quad \forall t.
\]
Letting \( x_j = (j - 1)\Delta x \) in which \( x_1 = 0 \) and \( x_J = L \) (6.54) numerically translates to
\[
u^n_i = 0 \quad \text{and} \quad v^n_j = 0 \quad \forall n.
\]

To show that this causes a problem when employing the CTCS scheme we first rewrite (6.41) and (6.42) to give
\[
u_j^{n+1} = u_j^{n-1} - \frac{\Delta t}{\Delta x} \left( \phi_{j+1}^n - \phi_{j-1}^n \right);
\]
\[
\phi_j^{n+1} = \phi_j^{n-1} - c_0^2 \frac{\Delta t}{\Delta x} \left( u_{j+1}^n - u_{j-1}^n \right).
\]

\textit{6} Remember that \( c_0 \) is defined as a positive definite quantity. Thus we do not use the absolute value in the definition of the Courant number as we did for the advection problem.

\textit{7} A typical current speed in the ocean is 0.1 m/s while typically internal wave phase speed is of order 1 m/s. In the atmosphere a typical wind speed is 10 m/s while the phase speed is typically of order 100 m/s or more.
where \( n = 0(1)N \) and \( j = 2(1)J - 1 \). To find the value at an arbitrary time level \( n \) at the first “wet” point \( j = 2 \) we use (6.56) and (6.57) and get\(^8\)

\[
\begin{align*}
\phi_{n+1}^j &= \phi_n^{j-1} - c_0^2 \frac{\Delta t}{\Delta x} (u_{n+1}^j - u_{n-1}^j), \\
\phi_{n+1}^{J-1} &= \phi_n^{J-2} - c_0^2 \frac{\Delta t}{\Delta x} (u_{n+1}^{J-1} - u_{n-1}^{J-1}).
\end{align*}
\]

The term that poses a problem is \( \phi_1^n \) in (6.58). It is unknown since the present scheme does not allow us to calculate it, and the number of allowable boundary conditions do not allow us to specify it. Please observe that we have a similar problem at the other boundary \( x = L \). Evaluating (6.56) and (6.57) for the last “wet” point \( j = J - 1 \) we get

\[
\begin{align*}
u_{n+1}^{J-1} &= u_{n+1}^{J-2} - \frac{\Delta t}{\Delta x} (\phi_n^{J-1} - \phi_{n-1}^{J-2}), \\
\phi_{n+1}^{J-1} &= \phi_n^{J-2} - c_0^2 \frac{\Delta t}{\Delta x} (u_{n+1}^{J-1} - u_{n-1}^{J-1}).
\end{align*}
\]

Again the term that poses a problem is the term \( \phi_1^n \).

If we try to remedy this problem by using option (ii) above, that is, specify \( \phi \) at \( x = 0 \) as well, the problem aggravates at \( x = L \). However tempting we are not allowed to specify more than two conditions in space total. If we continue and specifies \( \phi \) at the boundaries in addition to \( u \), we run into the problem of over-specifying the system, a dangerous path. We do get numbers out of the computer, they may even look reasonable, but they are wrong. This author strongly advocates against exploring such an avenue.

To avoid the problem Mesinger and Arakawa (1976) suggested to use what is referred to as staggered grids. Instead of calculating the two variables \( u \) and \( \phi \) at the same points, we simply stagger one of them with respect to the other, say one half grid length along the \( x \)-axis. Thus we calculate \( u \) at \( x_{j+\frac{1}{2}} \)-points and \( \phi \) at \( x_j \)-points. The CTCS finite difference approximation to (6.32) and (6.33) then reads

\[
\begin{align*}
u_{j+\frac{1}{2}}^{n+1} &= \nu_{j+\frac{1}{2}}^{n-1} - 2 \frac{\Delta t}{\Delta x} (\phi_{j+\frac{1}{2}}^{n} - \phi_{j}^{n}), \\
\phi_{j}^{n+1} &= \phi_{j}^{n-1} - 2 c_0^2 \frac{\Delta t}{\Delta x} (u_{j+\frac{1}{2}}^{n} - u_{j-\frac{1}{2}}^{n}).
\end{align*}
\]

Note the appearances of the factor 2 in the second term on the right-hand side of (6.62) and (6.63). It appears because the distance between two adjacent points in the finite difference approximation of \( \partial_x \) in the staggered formulation is \( \Delta x \) rather than \( 2\Delta x \), while the centered in time scheme still carries \( 2\Delta t \).

To avoid the cumbersome use of the \( j + \frac{1}{2} \) notation, we may visualize the points associated with \( \phi_j \) as being located in one grid and the points associated with \( u_{j+\frac{1}{2}} \) to be located in another grid of the same size, but staggered on half grid length as illustrated in Figure 6.2. Furthermore,

\(^8\)Note that for \( n = 0 \) the terms invoking the time level \( n - 1 \) is evaluated at \(-1\). This problem is handled as we did for the advection, that is, is avoided if we use forward in time scheme for the first time step.
as illustrated in Figure 6.2, we may visualize $\phi_j$ and $u_{j+\frac{1}{2}}$ as being located in the same cell numbered $j$. By using the cell numbering structure rather than the point numbering structure, recalling that (6.62) is evaluated at $u$-points and (6.63) at $\phi$-points, (6.62) and (6.63) become

$$u_{j+1}^n = u_j^n - 2\frac{\Delta t}{\Delta x} \left( \phi_{j+1}^n - \phi_j^n \right), \quad (6.64)$$
$$\phi_{j+1}^n = \phi_j^{n-1} - c_o^2 \frac{2\Delta t}{\Delta x} (u_j^n - u_{j-1}^n), \quad (6.65)$$

which avoids the cumbersome one-half notation. We emphasize though that (6.64) and (6.65) are correct if and only if the two grids are staggered exactly one half grid length.

It remains to answer whether the staggering has an impact on the stability. To investigate this we first note that care has to be exercised when constructing the discrete Fourier components. Since they are associated with the point numbering we have to take into account that the $u$ and $\phi$ points are one half grid-length apart. Thus the discrete Fourier components are

$$\phi_j^n = \Phi_n e^{i\alpha_j \Delta x} \quad \text{and} \quad u_j^n = U_n e^{i\alpha (j + \frac{1}{2}) \Delta x} \quad (6.66)$$

respectively. Substituting these expressions into (6.64) and (6.65), or (6.62) and (6.63), we get

$$U_{n+1} - U_{n-1} = -4i\gamma \Phi_n, \quad (6.67)$$
$$\Phi_{n+1} - \Phi_{n-1} = -4i\gamma c_o^2 U_n. \quad (6.68)$$

where

$$\gamma = \frac{\Delta t}{\Delta x} \sin \left( \frac{\alpha \Delta x}{2} \right) \quad (6.69)$$

Eliminating $U_n$ we get

$$\Phi_{n+2} - 2\Phi_n + \Phi_{n-2} = -16\gamma^2 c_o^2 \Phi_n. \quad (6.70)$$

Moreover, defining a growth factor by

$$G \equiv \frac{\Phi_{n+2}}{\Phi_n} \quad (6.71)$$
and using the formula $2 \sin^2 \frac{\psi}{2} = 1 - \cos \psi$ we get

$$G^2 - 2\lambda G + 1 = 0,$$

where

$$\lambda = 1 - 4c_0^2 \left( \frac{\Delta t}{\Delta x} \right)^2 (1 - \cos \alpha \Delta x).$$

The growth factor therefore has two complex conjugate solutions given by

$$G_{1,2} = \lambda \pm i\sqrt{1 - \lambda^2}.$$  

Thus as long as $\lambda^2 \leq 1$ we get $|G_{1,2}| = 1$. The staggered scheme is therefore neutrally stable as expected. However, this is only true as long as $\lambda^2 \leq 1$, which requires that

$$C \leq \frac{1}{2}, \quad \text{or} \quad \Delta t \leq \frac{\Delta x}{2c_0}.$$  

We observe that (6.75) is a more stringent CFL condition compared to the CFL condition (6.53) associated with the non-staggered grid. This is no surprise. When we stagger the grids we effectively decreases the distance between two adjacent points to one-half the original grid length. Thus the distance between two adjacent points in the staggered grid, say $\Delta x_{\text{stagg}}$, is simply $\Delta x_{\text{stagg}} = \Delta x/2$. Using $\Delta x_{\text{stagg}}$ instead of $\Delta x$ the CFL condition becomes the expected

$$\Delta t \leq \frac{\Delta x_{\text{stagg}}}{c_0}.$$  

### 6.4 Linear and rotating shallow water equations

We are now ready to analyze the effect of the Earth’s rotation by retaining the Coriolis terms. We first study their linear, one-dimensional version, that is, (6.28) - (6.30) on page 94.

As we did for the non-rotating case (cf. Section 6.2) we start by analyzing the various wave motions supported by (6.28) - (6.30). Thus we assume a wave solution,

$$X = X_0 e^{i\alpha(x - ct)},$$

where $\alpha$ is the wavenumber in the $x$ direction and $X$ denotes a vector consisting of the three dependent variables, that is,

$$X = \begin{bmatrix} u \\ v \\ \phi \end{bmatrix} \quad \text{and} \quad X_0 = \begin{bmatrix} u_0 \\ v_0 \\ \phi_0 \end{bmatrix},$$

Recall that any solution of (6.28) - (6.30), under the assumption that the solution is a good function, may be represented by an infinite number of waves of different wavelengths and amplitudes in accord with Section 2.11 on page 26.
where $X_0$ is the amplitude.

Inserting (9.21) into (6.28) - (6.30) we get

\begin{align*}
-i\alpha c u_0 - f v_0 + i\alpha \phi_0 &= 0, \\
-i\alpha c v_0 + f u_0 &= 0, \\
c\phi_0 - c_0^2 u_0 &= 0,
\end{align*}

To find the dispersion relation we first multiply (6.79) by $-i\alpha c$ and (6.80) by $f$ and add the results together. By substituting the result into (6.81) we then get

\begin{equation}
\left\{ c^2 - c_0^2 \left[ 1 + \left( \frac{1}{\alpha L_R} \right)^2 \right] \right\} = 0,
\end{equation}

where $L_R = c_0/f$ is the Rossby radius of deformation. Thus one solution is $c_1 = 0$, a stationary wave. The two remaining solutions are

\begin{equation}
c_{2,3} = \pm c_0 \sqrt{1 + \left( \frac{1}{\alpha L_R} \right)^2}.
\end{equation}

These are gravity waves modified by the Earth’s rotation and are commonly referred to as inertia-gravity waves. The gravity waves are associated with wave speeds $c = \pm c_0$, and are thus associated with the first term in the radical. The inertia part is associated with oscillating frequencies $\omega = f$ or phase speed $c = f/\alpha$, that is, frequencies proportional to the inertia frequency or inertial oscillation, and is contained in the second term under in the radical.

The same procedure may also be used if the basic state is in geostrophic balance, that is, using the linearized equations (6.25) - (6.27) on page 94. Substituting (6.78) into (6.25) - (6.27) we get

\begin{align*}
\bar{\alpha} (\bar{u} - c) u_0 - f v_0 + i\alpha \phi_0 &= 0, \\
f u_0 + i\alpha (\bar{u} - c) v_0 &= 0, \\
\bar{\phi} u_0 + (\bar{u} - c) \phi_0 &= 0.
\end{align*}

Following the same procedure the dispersion relation becomes

\begin{equation}
(\bar{u} - c) \left[ -\alpha^2 (\bar{u} - c)^2 + f^2 + \alpha^2 \bar{\phi} \right] = 0.
\end{equation}

Thus again there are three solutions for the phase speed $c$, namely

\begin{align*}
c_1 &= \bar{u}, \\
c_{2,3} &= \bar{u} \pm c_0 \sqrt{1 + \left( \frac{f}{\alpha c_0} \right)^2},
\end{align*}

where

\begin{equation}
c_0 = \sqrt{\bar{\phi}} = \sqrt{gH}.
\end{equation}
The first solution is an infinitely long wave in which the motion is in geostrophic balance commonly referred to as the Rossby mode. We easily derive this interpretation by substituting \( c_1 = \bar{u} \) from (6.88) into (6.84). The latter equation then becomes

\[-fv + i\alpha \phi = 0 \quad \text{or} \quad v = \frac{1}{f} i\alpha \phi. \quad (6.91)\]

Using the Fourier solution backwards we thus recover the geostrophic balance (1.40), that is,

\[v = \frac{1}{f} \partial_x \phi. \quad (6.92)\]

The two other solutions represented by \( \pm c_2^0 \sqrt{1 + \left( \frac{f}{\alpha c_0} \right)^2} \) are the inertia-gravity wave modes referred to in the previous paragraph, where the inertia part is associated with oscillating frequencies \( \omega = f \) and the two gravity waves with wave speeds \( c = \pm c_0 \).

**Finite difference form**

To solve (6.25) - (6.27) numerically using a finite difference method we replace the derivatives by finite difference approximations. To this end we first note that commonly the inertia-gravity mode has a much higher wave speed than the Rossby mode, that is \(|c_0| \gg |\bar{u}|\), and that these modes therefore becomes unstable first. For simplicity we therefore put \( \bar{u} = 0 \), that is, we are back to the problem with a basic state at rest. Since the problem is hyperbolic we apply the centered in time, centered in space (CTCS) leapfrog scheme on a non-staggered grid. We then get

\[u^{n+1}_j - u^{n-1}_j = 2f \Delta t v^{n}_j - \frac{\Delta t}{\Delta x} (\phi^{n+1}_j - \phi^{n-1}_j), \quad (6.93)\]

\[v^{n+1}_j - v^{n-1}_j = -2f \Delta t u^{n}_j, \quad (6.94)\]

\[\phi^{n+1}_j - \phi^{n-1}_j = -c^2_0 \frac{\Delta t}{\Delta x} (u^{n+1}_j - u^{n-1}_j). \quad (6.95)\]

Replacing the variables by their discrete Fourier components

\[u^n_j = U_n e^{i\alpha j \Delta x}, \quad v^n_j = V_n e^{i\alpha j \Delta x}, \quad \phi^n_j = \Phi_n e^{i\alpha j \Delta x}, \quad (6.96)\]

we get

\[U_{n+1} - U_{n-1} = 2f \Delta t V_n - 2i\Phi_n \frac{\Delta t}{\Delta x} \sin \alpha \Delta x, \quad (6.97)\]

\[V_{n+1} - V_{n-1} = -2f \Delta t U_n, \quad (6.98)\]

\[\Phi_{n+1} - \Phi_{n-1} = -2ic^2_0 U_n \frac{\Delta t}{\Delta x} \sin \alpha \Delta x. \quad (6.99)\]
Eliminating $V_n$ and $\Phi_n$ we get\(^\text{10}\)

$$U_{n+2} - 2\lambda U_n + U_{n-2} = 0,$$

(6.100)

where

$$\lambda = 1 - 2(C^2 \sin^2 \alpha \Delta x + f^2 \Delta t^2)$$

(6.101)

where in turn

$$C = c_0 \frac{\Delta t}{\Delta x}$$

(6.102)

is the Courant number as before. Defining the growth factor as $G \equiv U_{n+2}/U_n$ we get two complex conjugate solutions $G_{1,2} = \lambda \pm i\sqrt{1 - \lambda^2}$ provided the radical is a positive definite quantity. As expected $|G_{1,2}| = 1$ and hence the CTCS scheme is neutrally stable. The impact of throwing in the Coriolis effect is inherent in the expression for $\lambda$. To ensure that the growth factor is complex we must require $\lambda^2 \leq 1$. This is satisfied if

$$c_0^2 \left( \frac{\Delta t}{\Delta x} \right)^2 \sin^2 \alpha \Delta x + f^2 \Delta t^2 \leq 1.$$  

(6.103)

Thus the CFL criterion for stability becomes,

$$\Delta t \leq \frac{\Delta x}{c_0 \sqrt{1 + \left( \frac{\Delta x}{L_R} \right)^2}}.$$  

(6.104)

Since the first term in the radical dominates it follows that in practice

$$\Delta t \leq \frac{\Delta x}{c_0} \text{ or } C \leq 1.$$  

(6.105)

In an atmospheric model the largest equivalent depth is approximately 10 km giving a speed of the inertia-gravity waves of the order of 300m/s. This is considerably more than the wind speed and sets strong limitations to how long time steps we can take. In an ocean model the situation is the same. Although the oceanic equilibrium depth is one order of magnitude less, about 1km, the wave speed is still about 100m/s considerably larger than a typical ocean current speed of 0.1m/s. Thus both in the ocean and the atmosphere the CFL condition limits the time steps to minutes and sometimes seconds\(^\text{11}\). The inertia-gravity waves contain tidal motion and in the ocean also the storm surge signal. Thus in contrast to atmosphere models ocean model application are restricted to such limitations on the time step since we have to simulate these important oceanic features explicitly. In the atmosphere these motions are not part of the signal and thus may be treated semi-implicitly (cf. Section 6.6 on page 106).

\(^{10}\)This most efficiently done by first raising $n$ by one in (6.97) giving an expression involving $U_{n+2}, U_n, V_{n+1}$ and $\Phi_{n+1}$. Next decreasing $n$ by one gives an expression containing $U_n, U_{n-2}, V_{n-1}$ and $\Phi_{n-1}$. Subtracting and using (6.98) and (6.99) then results in (6.100).

\(^{11}\)For an ocean of equilibrium depth $H = 4 \cdot 10^3\text{m}, g = 10\text{ms}^{-2}$ and non-eddy resolving grid size $\Delta x = 20\text{km}$ follows from (6.105) that $\Delta t < 141\text{s}$ or slightly more than two minutes.
6.5 Non-linear dynamics

The next subset of the shallow water equations we consider is one in which we keep the non-linear terms in the system (6.3) - (6.4). Again assuming a one-dimensional system, that is, we neglect all terms differentiated with respect to $y$, we get

\[
\begin{align*}
\partial_t u + u \partial_x u - fv &= -\partial_x \phi, \quad (6.106) \\
\partial_t v + u \partial_x v + fu &= 0, \quad (6.107) \\
\partial_t \phi + u \partial_x \phi &= -\phi \partial_x u. \quad (6.108)
\end{align*}
\]

We note that we retain three equations with three unknowns $u$, $v$ and $\phi$ as dependent variables. The independent variables are time $t$ and the horizontal direction $x$.

Rewriting (6.106) - (6.108) in terms of the volume flux $U = hu$, $V = hv$, and noting that $\phi = gh$, we get

\[
\begin{align*}
\partial_t U + \partial_x \left( \frac{U^2}{h} \right) - fV &= -\frac{1}{2} g \partial_x h^2, \quad (6.109) \\
\partial_t V + \partial_x \left( \frac{UV}{h} \right) + fU &= 0, \quad (6.110) \\
\partial_t h + \partial_x U &= 0. \quad (6.111)
\end{align*}
\]

The advantage when using (6.109) - (6.111) instead of (6.106) - (6.108) is that the continuity equation (6.111) becomes linear, and that the non-linear terms in (6.109) and (6.110) are all written in flux form. The latter gives better conservation properties of the associated numerical scheme.

**Finite difference form**

Let us first neglect the Coriolis terms. The system (6.106) - (6.108) then reduces to

\[
\begin{align*}
\partial_t u + u \partial_x u + \partial_x \phi &= 0, \quad (6.112) \\
\partial_t \phi + u \partial_x \phi + \phi \partial_x u &= 0, \quad (6.113)
\end{align*}
\]

and as we have experienced before (6.107) becomes redundant. Thus we are left with a system of two equations for the two unknowns $u$ and $\phi$. We note that written in flux form they read

\[
\begin{align*}
\partial_t U + \partial_x \left( \frac{U^2}{h} \right) &= -\frac{1}{2} g \partial_x h^2, \quad (6.114) \\
\partial_t h &= -\partial_x U. \quad (6.115)
\end{align*}
\]

The system is still hyperbolic so it is natural to employ a CTCS (leapfrog) scheme. Hence employing an unstaggered grid we get

\[
\begin{align*}
U_j^{n+1} &= U_j^n + AU_j^n + PR_j^n \quad (6.116) \\
h_j^{n+1} &= h_j^{n-1} - \frac{\Delta t}{\Delta x} (U_{j+1}^n - U_{j-1}^n), \quad (6.117)
\end{align*}
\]
where
\[ AU_j^n = -\frac{\Delta t}{\Delta x} \left( \left[ \frac{U^2}{h} \right]_{j+1} - \left[ \frac{U^2}{h} \right]_{j-1} \right) \quad \text{and} \quad PR_j^n = -\frac{g\Delta t}{2\Delta x} \left( [h^2]_{j+1} - [h^2]_{j-1} \right) \] (6.118)

If we retain the Coriolis terms the system is still hyperbolic. Thus we continue to apply a CTCS scheme. In fact the scheme is not much different from (6.116) and (6.117). If we discretize (6.109) - (6.111) on an unstaggered grid using a CTCS scheme we thus get

\[ U_j^{n+1} = U_j^{n-1} + 2f\Delta t V_j^n + AU_j^n + PR_j^n \] (6.119)
\[ V_j^{n+1} = U_j^{n-1} + 2f\Delta t U_j^n + AV_j^n \] (6.120)
\[ h_j^{n+1} = h_j^{n-1} - \frac{\Delta t}{\Delta x} (U_j^n - U_{j-1}^n) \] (6.121)

where
\[ AV_j^n = -\frac{\Delta t}{\Delta x} \left( \left[ \frac{UV}{h} \right]_{j+1} - \left[ \frac{UV}{h} \right]_{j-1} \right) \] (6.122)

Thus the what is added to (6.116) and (6.117) is a third equation, namely (6.120). In addition we notice the Coriolis terms.

To avoid the use of more boundary conditions in space than allowed, we have to stagger the grids for \( h \) and \( U,V \) as explained in Section 6.3. Using the cell structure of Figure 6.2 and using \( j \) as the cell counter we then get

\[ U_j^{n+1} = U_j^{n-1} + 2f\Delta t V_j^n + AU_j^n + PR_j^n \] (6.123)
\[ V_j^{n+1} = U_j^{n-1} + 2f\Delta t U_j^n + AV_j^n \] (6.124)
\[ h_j^{n+1} = h_j^{n-1} - \frac{2\Delta t}{\Delta x} (U_j^n - U_{j-1}^n) \] (6.125)

where \( AU_j^n \) and \( AV_j^n \) are as before, while \( PR_j^n \) because of the staggering becomes

\[ PR_j^n = -\frac{g\Delta t}{\Delta x} \left( [h^2]_{j+1} - [h^2]_{j-1} \right) \] (6.126)

**Numerical stability**

The question then arises if the schemes are stable. Since the equations are non-linear the analysis is not as straightforward as for a linear system. In fact throwing in the non-linear terms adds to the complexity of the possibilities for an unstable solution. The reason is, as alluded to earlier on in Section 3.2 on page 32 and Section 4.6 on page 50, that in contrast to linear dynamics non-linear dynamics allow energy to be exchanged between waves. Thus the non-linear terms are able to redistribute energy among the different wavelengths present in the problem, something which is impossible in a linear system. In fact as time evolves the non-linear terms acts to cascade the energy progressively towards smaller and smaller wavelengths in accord with the Taylor rhyme quoted on page 32.
When we solve non-linear problems numerically using finite difference approximations to replace the differential operators inherent in our equations, we simply do not resolve waves with wavelengths shorter than $2\Delta x$. In the real world, however, the energy cascade continues beyond this wavelength and progressively towards shorter and shorter waves. In our numerical solution the grid resolution inhibits this cascading across the resolution limit, that is, across the $2\Delta x$ wavelength. The energy thus accumulates in the wavelength bands closest to our grid resolution, that is, in the $2\Delta x < \lambda < 4\Delta x$ wavelength band. Eventually this causes any scheme that works well for a linear system to blow up. This is called non-linear instability and is treated in more detail in Section 10.3 on page 149.

The processes responsible for continue the energy cascade across and beyond the grid resolution wavelength acts on scales that are shorter than our $2\Delta x$ grid resolution. Such processes are commonly referred to as sub-grid-scale (SGS) processes. To numerically allow for a continuation of the energy cascade across the the $2\Delta x$ limit we have to parameterize (or mimic) the SGS processes. Commonly we parameterize the SGS processes in terms of some kind of diffusion. As we were showing in Section 4.2, diffusion acts to smooth out noise, that is, diffusion is selective and helps to smooth out the shortest waves. We may therefore add diffusive terms to our governing, non-linear equations to avoid the continues accumulation of energy in the short wavelength band. In our relatively simple one-dimensional cases this entails adding terms of the form $\kappa \partial_x^2 u$, where $\kappa$ is a diffusion coefficient. As alluded to in Chapter 4 the diffusion inherent in such a term is scale selective, that is, damps out the shortest way most efficiently since all the Fourier components will be damped by a factor $e^{-\kappa \alpha^2 t}$. The results are however sensitive to the choice of diffusion coefficient. Too high and the energy is damped too fast. This will give a result where too much energy is absorbed by diffusion and the solution will too smooth. Too low and the energy contained in low wave number band increases with time. The results is that the non-linear instability sooner or later kicks in and the solution becomes unstable.

### 6.6 The semi-implicit method

From the analysis above we notice that by introducing a pressure force (in addition to advection) the CFL criterion becomes much more stringent (shorter time-step). It is therefore tempting to treat terms responsible for this behavior implicitly while we treat other terms explicitly. Such a method is commonly referred to as a **semi-implicit method**.

For clarity we start with a one dimensional shallow water problem, that is,

\[
\begin{align*}
\partial_t u &= A_u - \partial_x \phi, \\ 
\partial_t v &= A_v, \\ 
\partial_t \phi &= A_\phi - \Phi \partial_x u,
\end{align*}
\]

where $A_u$, $A_v$, and $A_\phi$ include the non-linear as well as the Coriolis terms. We learned previously that the terms responsible for this behavior was the pressure terms. From Section 5.2 we learned that treating any term implicitly avoid this restriction on the time step. It is therefore tempting to treat the pressure terms, that is, the $\partial_x \phi$ and $\Phi \partial_x u$ terms implicitly while integrating the remaining terms explicitly. The finite difference approximation form of the equations above on an
unstaggered grid then becomes,

$$\frac{u^{n+1} - u^{n-1}}{2\Delta t} = [A_u]^n - [\partial_x \phi]^{n+1}, \quad (6.130)$$

$$\frac{v^{n+1} - v^{n-1}}{2\Delta t} = [A_v]^n, \quad (6.131)$$

$$\frac{\phi^{n+1} - \phi^{n-1}}{2\Delta t} = [A_\phi]^n - \Phi [\partial_x u]^{n+1}, \quad (6.132)$$

To proceed, the first two equations are solved with respect to $u_j^{n+1}$ and $v_j^{n+1}$ respectively giving,

$$u^{n+1} = u^{n-1} + 2\Delta t [A_u]^n - 2\Delta t [\partial_x \phi]^{n+1} \quad (6.133)$$

$$v^{n+1} = v^{n-1} + 2\Delta t [A_v]^n \quad (6.134)$$

derivation with respect to $x$ and insertion into the equation for $\phi$ gives a Helmholtz equation,

$$\Phi \Delta t^2 [\partial_x^2 \phi]^{n+1} - \phi^{n+1} = B \quad (6.135)$$

where $B$ contains known quantities at time levels $n, n-1, \ldots$

With proper boundary conditions ($\phi$ or its normal derivative at lateral boundaries), these equations may easily be solved by standard numerical methods called elliptic solvers. One such elliptic solver, is the direct elliptic solved treated in Section 4.8 (Gauss elimination), but there are also a host of iterative (non-direct) solvers to choose. The most commonly in use is called the successive over-relaxation (SOR) method. Having obtained $\phi^{n+1}$, we easily find $u^{n+1}$ and $v^{n+1}$.

This method is widely used in atmospheric models as we do not have to take the gravity mode speed $c_0 = \sqrt{\Phi} = \sqrt{gH}$ into account when estimating an upper bound for the time step. Thus we avoid the restrictive CFL condition and thus makes it is possible to use longer time-steps.

We emphasize that we cannot employ this method in the ocean. Treating the fast barotropic waves implicitly would then ruin the inertia-gravity waves which carries information about such important signals as tides and storm surges.

### 6.7 The Semi-Lagrangian method

The semi-Lagrangian method is constructed in order to take even longer time steps for advection than the Leapfrog method permits. In addition it is very accurate. For each time step one calculates where parcels arriving at a grid point came from. These departure points will normally differ from the standard grid-points and the values of the prognostic variables at time level $n$ must be estimated by interpolation from the surrounding grid point values. The new value at the grid points at time level $n+1$ due to advection are then simply the value at the departure point at time level $n$. The one dimensional rendition of this method is the method of characteristics as outlined in Section 5.9.

By using an interpolation method with sufficient accuracy, for instance cubic interpolation, the method is more accurate than using second order differences and leapfrog time-integration.
In addition the stability is independent of the length of the time step as long as the departure-points values are calculated by interpolation (not extrapolation). However, the accuracy strongly depends on how accurate one can estimate the departure points and with very long time steps the estimates will be inaccurate. Another drawback is that the method requires additional computations and this overhead may be as costly as taking a shorter time step with a simpler method. The method is widely used in atmospheric models; often in combination with the semi-implicit method.
Chapter 7

OPEN BOUNDARY CONDITIONS AND NESTING TECHNIQUES

As is well known, computers, however large, can only hold a finite number of numbers in their so-called random access memory (RAM). Thus even the biggest computers are limited in their capacity. This is perhaps the main reason why numerical oceanography is less mature than numerical meteorology. Recall that at sub-polar latitudes the typical length scale of a synoptic low in the atmosphere is about 500 - 1000 km and the typical time scale a few days. In contrast the scales of a low in the ocean at the same latitude are about 10 - 50 km in length and a few weeks to months in time. These differences are associated with the difference in the Rossby radius of deformation that in the ocean is about two orders of magnitudes smaller than in the atmosphere.

To illustrate this point let us consider a global model with a grid size of about 2 degrees (Figure 7.1 upper panel). A mesh size of 2 degrees, or about 200 km, entails that the grid size is about one fifth of the the atmospheric Rossby radius. This is a tolerable grid size for a numerical atmosphere model. If we, however, scale this to the Rossby radius of deformation in the ocean, the grid for the atmosphere model would look like the one displayed in the lower panel of Figure 7.1 which has a grid size of about 3-4 times the Rossby radius of deformation. No meteorologist in his right mind would consider it to be an adequate grid for a numerical weather prediction (NWP) model. To obtain a similar tolerable resolution in the ocean we have to employ grids of mesh sizes 2-4 km, or 1/200th of a degree. Thus the need for RAM is much higher for an ocean model than for an atmosphere model covering the same region. In addition comes the fact that with decreased mesh size the time step is much smaller as well in order to satisfy the CFL criterion. In practice it therefore takes a much larger computational effort to provide say a 24 hour “weather” forecast for the ocean for a given area on a given computer than to provide a similar weather prediction. To enable computers to provide numerical ocean weather forecasts as fast as today’s NWP models for the same area we therefore need much faster computers. To make things even worse recall that the time scale in the ocean is much longer than in the atmosphere. A weather prediction of say ten days corresponds to an ocean forecast of at least one month.

Global weather predictions with more than adequate resolution to resolve the atmospheric weather systems are common today. Such forecasts are run by several national institutes as well
Figure 7.1: Upper panel shows the Earth’s surface covered by a 2 degree mesh. Lower panel shows a similar mesh of 30 degrees mesh size. The figure conveniently illustrates how a 2 degree mesh in the ocean would look like in the atmosphere scaled by the Rossby radius of deformation.
OPEN BOUNDARY CONDITIONS

as at the European Centre for Medium-range Weather Forecasts (ECMWF). However, there are still local effects, notably processes associated with irregular topography, that is not yet resolved properly by the global model. Thus most national institutes providing public meteorological services run a limited area model which is “nested” into a global model\(^1\). Nesting a finer mesh model (sometimes referred to as the inner model) into a coarser mesh model (sometimes referred to as the outer model) entails that we want to transfer results from the outer model to the inner model and vice versa. This has to be done via their common boundary. Thus one need a condition at these boundaries. Since fluid is allowed to pass freely through these boundaries they are referred to as open boundaries. Nesting of a finer mesh model into a coarser model is also sometimes referred to as dynamical downscaling since the inner model provides a solution that is a dynamically consistent downscaling of the coarser solution to a scale that takes into account the finer scales, for instance due to a more realistic representation of the topography.

The situation for the ocean is even more profound. Global numerical ocean weather predictions (NOWP) are not yet feasible at all on today’s computers. Thus none of the global ocean models that exist today resolve the oceanic weather at higher latitudes. This is particularly true for the oceanic component of the coupled global climate models, or so called AOGCMs (Atmosphere - Ocean Global Climate Models). To provide synoptic NOWPs in limited areas we therefore have to make use of nesting techniques or dynamical downscaling for the time being.

As a consequence both ocean and atmosphere models have to deal with open boundaries. At these boundaries the governing equations are still valid, which makes them very different from natural boundaries where the governing equations are replaced by a natural boundary condition as for instance the condition of no normal flow through a solid, impermeable boundary. Nevertheless, since our computational domain ends at the open boundary we have to provide a boundary condition there. Such conditions are called open boundary conditions or OBCs for short. The mathematical challenge is to construct OBCs which ensure that a solution to the governing equations exists and is unique. From a physical point of view though we would like the solution to be as close to the “correct” solution as possible. The correct solution refers to the one we would have obtained if the model was global with natural boundary conditions applied along its boundaries only. It should be emphasized that this is a dilemma since the solution to the governing equations is determined not only by the equations themselves, but also by the boundary conditions as mentioned in Section 2.5. Thus when applying OBCs to determine our solution we are not ensured that the solution we obtain is the correct one. In fact it is impossible in general to prove that the solution we obtain exists and is unique. The latter is only possible in special cases, for instance for very simplified linear systems.

One of the first to derive a numerical solution of a NWP problem applying OBCs was Jules Charney, Ragnar Fjørtoft and John von Neumann in the late 1940s. Their attempt was published in the famous paper by Charney et al. (1950). Recall that just after the second world war the digital computers were in their infancy. They therefore employed a very simple atmospheric model compared to today’s standards. In fact they attempted to solve the barotropic, quasi-geostrophic equations formulated in terms of a potential vorticity equation for a computational

\(^1\)At the Norwegian Meteorological Institute the limited area model at the time of writing is HIRLAM - the High Resolution Limited Area Model
domain limited to a square located in the middle of the North Atlantic Ocean. Thus they had to apply OBCs at all four sides of the domain. They opted for a set of OBCs which was to specify the potential vorticity at the inflowing boundaries and to apply a radiation condition (see Section 7.2 below) at outflowing boundaries. Later Platzman (1954) showed that their solution was unstable when applying the OBC they had chosen. Thus the problem of specifying OBCs that renders the solution unique and correct is not new. Since these early papers on the problems encountered in specifying OBC there has been a variety of papers treating the problem. For a review see Chapman (1985); Reed and Cooper (1986, 1987); Palma and Matano (2000) and most recently Blayo and Debreu (2006).

7.1 What is an open boundary?

To help in choosing an OBC that would render the solution stable, the following definition of an open boundary was offered by Reed and Cooper (1986):

An open boundary is a computational boundary at which disturbances originating in the interior of the computational domain are allowed to leave it without disturbing or deteriorating the interior solution.

Thus we demand that the conditions we impose at open boundaries satisfies certain requirements. These requirements may later be used to construct criteria by which the success of our employed OBC may be assessed. We emphasize that an open boundary is an artificial boundary between what we call our computational domain (or our area of interest) and a domain exterior to our computational domain.

One obvious requirement, which follows directly from the definition above, is that disturbances originating in the interior of our domain propagating toward the open boundary should be allowed to pass through to the exterior without distorting or disturbing the interior solution. Equally obvious is that disturbances originating in the exterior domain is free to enter our domain without distortions. The latter is sometimes hard to achieve since we do not always have sufficient knowledge about the exterior solution.

To illustrate this let us consider a Kelvin wave created in the interior of our domain and propagating towards the open boundary. The condition we impose at the open boundary should then be able to let the wave pass through and not be reflected, that is, none of the energy contained in the Kelvin wave should be allowed to be radiated back into the interior. Likewise a Kelvin wave created in the exterior domain, for instance by a model covering a larger domain that encompasses our computational domain, should be free to enter without being distorted or damped. Next we require that the chosen OBC leads to a stable solution (numerically). Moreover, from a mathematical point of view we require that the OBC together with the governing equations leads to a mathematical problem that is well posed or at least well-posed enough so that a solution exists and is unique.

In the following we will give some details regarding some of the common OBCs developed over the years. In the two last sections we will give details about two of the most promising
OPEN BOUNDARY CONDITIONS 7.2 Radiation conditions

OBCs that are useful as nesting techniques as well as OBCs. These are the Flow Relaxation Scheme (e.g., Engedahl, 1995a) and the Perfectly Matched Layer approach (e.g., Navon et al., 2004; Blayo and Debreu, 2006). While the former OBC was developed in the meteorological community (Davies, 1976, 1985), the latter comes from the electromagnetic field (Berenger, 1994).

7.2 Radiation conditions

Many of the processes in the ocean and atmosphere are processes involving wave propagation in one way or another. The early attempts at developing OBCs therefore based their OBC formulations on simple wave equation. In its simplest form the wave equation reads,

\[ \partial_t \phi + c_\phi \partial_n \phi = 0 \] (7.1)

Here \( \phi \) represents the dependent variable, \( c_\phi \) is the component of the phase velocity normal the boundary associated with the variable \( \phi \), while \( \partial_n \) denotes the derivative normal to the open boundary. Imposing (7.1) as an OBC it becomes what is known as the radiation condition. When use is made of (7.1) as an OBC we fundamentally assume that the disturbances passing through the open boundary consists of waves. Note that the disturbances passing through the boundary may consist of several waves of different wavelengths, and hence that (7.1), strictly speaking, is only valid for one Fourier component only. It is thus only suitable for linear problems in which there is no energy exchange between wave numbers.

One of the first obstacles in employing (7.1) as our OBC is that we do not know the phase velocity \( c_\phi \). From Section 5.10 we recall that \( c_\phi \) is the slope of the characteristics. Thus if the choice of \( c_\phi \) perfectly matches the slope of the characteristics then (7.1) is a perfect open boundary condition. However, it is only for very simple, physical problems, e.g., for a monochromatic wave problem, that we are able to determine the characteristics a priori, and hence \( c_\phi \) is generally unknown.

We immediately recognize that (7.1) contains two special cases. The first case is \( c_\phi = 0 \), while the second is the opposite, namely when the phase velocity \( c_\phi \to \infty \). In the former case we notice that the characteristics are straight vertical lines in \( x, t \) space, and that we may integrate (7.1) in time to give,

\[ \phi = \text{const.} \] (7.2)

Thus under these circumstances the dependent variable is known for all times at the OBC, and we recall from Section 2.5 that the OBC is a Dirichlet condition. Commonly this is referred to as a clamped condition since the dependent variable \( \phi \) does not change in time at the OB.

In the latter case when \( c_\phi \to \infty \) we notice that characteristics are horizontal straight lines in the \( x, t \) space. We notice from (7.1) that if \( \partial_t \phi \) should remain finite we must require that the gradient \( \partial_n \phi \) must be zero, and hence that

\[ \partial_n \phi = 0. \] (7.3)

Usually the condition (7.3) is referred to simply as a gradient condition. We recall from Section 2.5 that such a condition was referred to as a Neumann condition.
If the phase velocity is finite and differs from zero, then we have a true radiation condition. The problem is then reduced to determine the phase velocity $c_\phi$. If the solution is in the form of known waves, say a barotropic Kelvin wave\(^2\). Under these circumstances the phase velocity is known, and in the case of a Kelvin wave it is,

$$c_\phi = c_0 = \sqrt{gH} \quad (7.4)$$

where $g$ is the gravitational acceleration, and $H$ is the equilibrium depth of a fluid column.

Let us consider a problem of a fluid contained in a channel of equilibrium depth $H$. Furthermore, let us consider a frictionless motion and let $h$ denote the total depth or layer thickness of a fluid column and $u$ the speed of the fluid column\(^3\). Moreover, let us consider that the motion is on a non-rotational Earth, and that the fluid has constant and uniform density. Then the governing equations may be written (cf. Section 6.2)

$$\partial_t u = -g \partial_x h \quad (7.5)$$
$$\partial_t h = -H \partial_x u \quad (7.6)$$

The classic method to solve the above set is to first differentiate (7.5) with respect to $x$ and (7.6) with respect to $t$ and then add the results. The result is

$$\partial_t^2 h - c_0^2 \partial_x^2 h = 0, \quad (7.7)$$

that is, a wave equation with a phase speed equal to $c_0$ as given in (7.4). The set (7.5) and (7.6) thus requires two boundary conditions in space. Let us assume that the channel has two open boundaries at $x = 0$ and $x = L$. The natural boundary condition at these two boundaries is then the radiation condition (7.1) with a phase speed of $\pm c_0$, respectively.

Recalling that the phase velocity is determined by the slope of the characteristics, we may also use the method of characteristics to find it. As we will show this also leads us to impose the radiation condition at the two open boundaries. We start by multiplying (7.6) by an as yet unknown function $\lambda$ and add the result to (7.5). We then get

$$\partial_t u + \lambda H \partial_x u + \lambda \left( \partial_t h + \frac{g}{\lambda} \partial_x h \right) = 0 \quad (7.8)$$

Next we define the operator $\frac{D^x}{dt}$ such that

$$\frac{D^x}{dt} = \partial_t + D^x \partial_x \quad (7.9)$$

that is, an individual derivative in the special direction $D^x$ (cf. Section 5.10). We may now formulate an equation for the unknown function $\lambda$ by requiring that the characteristics satisfies the following equations

$$\frac{D^x}{dt} = \lambda H, \quad (7.10)$$

\(^2\)A barotropic Kelvin wave is common phenomena in oceanography. It belongs to the class of planetary gravity waves. Kelvin waves are commonly filtered out in meteorology models.

\(^3\)Since we consider a frictionless motion we may safely assume that $u$ is independent of depth.
and
\[
\frac{D^* x}{dt} = \frac{g}{\lambda}.
\] (7.11)

Under these circumstances we may write equation (7.8) as
\[
\frac{D^* u}{dt} + \lambda \frac{D^* h}{dt} = 0.
\] (7.12)

Equating the right-hand sides of (7.10) and (7.11) and solving for the unknown function \( \lambda \) we get the two solutions
\[
\lambda_{1,2} = \pm \frac{c_0}{H},
\] (7.13)

where \( c_0 = \sqrt{gH} \). Hence (7.12) becomes
\[
\left( \frac{D^*}{dt} \right)_{1,2} \left( u \pm c_0 \frac{h}{H} \right) = 0
\] (7.14)
valid along the characteristics given by
\[
\left( \frac{D^* x}{dt} \right)_{1,2} = \pm c_0.
\] (7.15)

We observe that (7.14) and (7.15) are formally two equations. By use of (7.9) and (7.15) we thus get
\[
\partial_t \left( u + c_0 \frac{h}{H} \right) + c_0 \partial_x \left( u + c_0 \frac{h}{H} \right) = 0,
\] (7.16)
\[
\partial_t \left( u - c_0 \frac{h}{H} \right) - c_0 \partial_x \left( u - c_0 \frac{h}{H} \right) = 0.
\] (7.17)

We recall from Section 5.9 that (7.16) and (7.17) (or eq. 7.14) are the compatibility equations, while (7.15) are the two characteristic equations. While (7.16) describes a wave propagating in the positive \( x \)-direction with phase velocity \( c_0 \), we observe that (7.17) describes a wave propagating in the opposite direction, but with the same phase velocity. In particular we notice that (7.14) expresses that the specific combinations of the dependent variables \( u \) and \( h \), namely \( u \pm c_0 \frac{h}{H} \), are conserved along the characteristics (7.15).

Let us assume that our problem is to solve (7.5) for \( 0 < x < L \) and that the two boundaries \( x = 0 \) and \( x = L \) are open. Let us in addition assume that a motion is generated in the interior of the domain, e.g., in the form of an initial deviation of the layer thickness \( h \) locally. The question then arises: what is the correct boundary condition to impose on the two open boundaries? We know from the two compatibility equations (7.14) that the information about the deviation will propagate along the two characteristics given by (7.15). Towards the right-hand boundary at \( x = L \) the information will propagate along \( (\frac{D^* x}{dt})_1 = c_0 \), and towards the left-hand boundary \( x = 0 \) along \( (\frac{D^* x}{dt})_2 = -c_0 \). To avoid reflection we must impose a condition that ensures that information cannot propagate back into our interior domain. Since information propagates along
the characteristics, we must ensure that no characteristics at \( x = 0, L \) slopes towards the interior. Consequently we require that

\[
\left( \frac{D^*x}{dt} \right)_2 = 0 \quad \text{at} \quad x = L
\]  

and

\[
\left( \frac{D^*x}{dt} \right)_1 = 0 \quad \text{at} \quad x = 0.
\]

Substituting this into the leftmost expressions in respectively (7.16) and (7.17) we get

\[
\partial_t \left( u - c_0 \frac{h}{H} \right) = 0 \quad \text{at} \quad x = L
\]  

and

\[
\partial_t \left( u + c_0 \frac{h}{H} \right) = 0 \quad \text{at} \quad x = 0.
\]

We now integrate (7.20) and (7.21) in time and get

\[
u = c_0 \frac{h}{H} + \text{const.}, \quad \text{at} \quad x = L
\]  

and

\[
u = -c_0 \frac{h}{H} + \text{const.} \quad \text{at} \quad x = 0.
\]

This is in fact the radiation condition. Indeed if we substitute the expression (7.6) for \( h \) into (7.22) we get (7.1) with \( \phi = u \) and \( c_\phi = c_0 \).

The advantage of using the method of characteristics to derive the non-reflective boundary condition is that it gives us insight into how to construct open boundary conditions in general. This was for instance exploited by Røed and Cooper (1987) to construct a weakly reflective boundary condition for a more general problem including the effect of Earth’s rotation based on earlier work by Hedstrøm (1979) (cf. Section 7.5).

### 7.3 Implementation of the radiation condition

We now consider the numerical implementation of the one-dimensional version of the radiation condition (7.1), and that the space variable is \( x \). In this we essentially follow the implementation given in Røed and Cooper (1987). Recall that the radiation condition then reads

\[
\partial_t \phi + c_\phi \partial_x \phi = 0.
\]

To get started let us assume that the computational domain is \( x \in <0, L> \) and \( t \in <0, T> \). The boundaries are then at \( x = 0, L \), where we assume that \( x = L \) is an open boundary, while \( x = 0 \) is a natural boundary. Furthermore we construct a grid in the \( x, t \) coordinates where \( x_j = (j - 1) \Delta x \) and \( t^n = n \Delta t \) (cf. Fig. 7.2).
Figure 7.2: Sketch of the mesh in the $t, x$ plane close to the right-hand open boundary. The computational domain is then to the left of $x = L$. The letters $J$, $J - 1$, and $J - 2$ denote grid points respectively at the open boundary, the first and second points inside the computational domain, while $n$, $n - 1$, and $n + 1$ denote the time levels.

Since (7.24) is an advection equation it is natural that we use one of the stable schemes developed in Section 5.1. We emphasize that it is important that the interior scheme and the scheme we use to solve the radiation equation has the same accuracy. Thus if the interior scheme is of second order accuracy in time and space then it is natural that we choose the leapfrog scheme. If the interior scheme is first order in time and space then it is natural that we choose a similar scheme for the radiation condition (7.24), say the upwind scheme. In the following we assume that the latter is true.

We then proceed using the upwind scheme for the radiation condition at $x = L$, that is for $j = J$. Assuming that $c_\phi \geq 0$, and following the notation of Figure 7.2, we get

$$\frac{\phi_{j}^{n+1} - \phi_j^n}{\Delta t} + c_\phi \frac{\phi_j^n - \phi_{j-1}^n}{\Delta x} = 0 \quad (7.25)$$

or

$$\phi_{j}^{n+1} = (1 - r_\phi)\phi_j^n + r_\phi\phi_{j-1}^n \quad (7.26)$$

where

$$r_\phi = c_\phi \frac{\Delta t}{\Delta x}. \quad (7.27)$$

Equation (7.26) says that the radiation condition in essence is an interpolation of values from the interior and at previous times. The problem is that we don’t know the weighting function, that is, the phase velocity $c_\phi$? As suggested by Orlanski (1976) we might solve (7.20) with respect to the phase velocity (or $r_\phi$) and get

$$r_\phi = -\frac{\phi_{j}^{n+1} - \phi_j^n}{\phi_j^n - \phi_{j-1}^n}. \quad (7.28)$$
7.3 Implementation

However, since we do not know the solution at the boundary at time level \( n+1 \) this expression is useless. Our only way of determining \( c_\phi \) (or \( r_\phi \)) is to use our knowledge about the solution at previous times. We then have several options. One is to use interior points at the same time level, in which case

\[
 r_\phi = -\frac{\phi_{n+1} - \phi_{n-1}}{\phi_{J-1} - \phi_{J-2}}. \tag{7.29}
\]

A second is to use information at previous times at same points in space,

\[
 r_\phi = -\frac{\phi_n - \phi_{n-1}}{\phi_{J-1} - \phi_{J-2}}. \tag{7.30}
\]

Both of these expressions provides an expression for the phase velocity. But which is the correct one? If we interpret (7.29) and (7.30) in terms of characteristics as in the previous section (see also Section 5.9), we notice that (7.29) assumes that the slope of the characteristic through the \((x_J, t^{n+1})\) point to a first approximation equals the slope through the \((x_{J-1}, t^{n+1})\) point, while (7.30) assumes that it to a first approximation equals the slope through the point \((x_J, t^n)\). Following this argument a third option is to assume that the characteristic through \((x_J, t^{n+1})\) continues backward in times and crosses the time level \( n - 1 \) between \(x_{J-1}\) and \(x_{J-2}\). This is tantamount to assume that to a first approximation the slope through \((x_J, t^{n+1})\) equals the slope through \((x_{J-1}, t^n)\). We then get the following expression for the phase velocity

\[
 r'_\phi = -\frac{\phi_{n-1} - \phi_{n-1}}{\phi_{J-1} - \phi_{J-2}}. \tag{7.31}
\]

On purpose we have used a prime for this expression, since it is a predictor for the phase velocity. We must require that the number returned is not negative. Hence we correct the result by defining \( r_\phi \) (no prime attached) as

\[
 r_\phi = \begin{cases} 
   r'_\phi & ; & 0 \leq r'_\phi \\
   0 & ; & r'_\phi < 0 
\end{cases}. \tag{7.32}
\]

As argued by Røed and Cooper (1987) we think (7.32) is a better approximation. Consequently, we use the expression (7.32) to substitute for \( r_\phi \) in (7.26) when determining the new boundary value \( \phi_{n+1} \) at time level \( n+1 \). If our open boundary was at \( x = 0 \) the inequality sign in (7.32) must be reversed to ensure that the phase velocity then is negative.

As alluded to it is only pure wave problems where processes like non-linear interactions, friction, wind forcing and the Coriolis acceleration are neglected that satisfies the radiation condition (7.1). All realistic models employed today within oceanography or meteorology are much more complex, and includes at least the processes just mentioned, and in most cases many more. The radiation condition is therefore far from being a perfect open boundary condition\(^4\).

Since the radiation condition in most cases is far from being perfect the meteorological and oceanographic communities has developed several other optional OBCs (e.g., Chapman, 1985; Røed and Cooper, 1986, 1987; Palma and Matano, 2000; Blayo and Debreu, 2006). In the following sections we study some of the more popular ones.

\(^4\)In fact it may be shown that there is no such thing as a perfect boundary condition mathematically speaking, since the problem in a geophysical context is ill posed.
7.4 The sponge

One of the most popular ones is the so called "sponge condition." In essence the method is to extend the computational domain outside of the area of interest (interior domain) to include an area where the energy contained in that part of the solution leaving the interior domain is gradually decreased. In practice we achieve this by gradually increasing the relative importance of those terms associated with diffusive or frictional processes as the solution is advected or propagated into the exterior or extended domain (sometimes referred to as the sponge domain).

As an example let us study the problem formulated in association with (7.5), but with some friction added. Thus the governing equations become

\[ \partial_t u = -g \partial_x h - \gamma u, \]  
\[ \partial_t h = -H \partial_x u, \]

(7.33)
(7.34)

where \( \gamma \) is a constant in the interior domain, say \( \gamma = \gamma_0 \). As in the former problem we solve (7.33) and (7.34) within the interior domain \( x \in (-L, L) \) where \( x = 0 \) and \( x = L \) are open boundaries. Outside of the interior domain we extend the computational domain to include areas within which we let the frictional parameter \( \gamma \) increase gradually. For instance at the right-hand open boundary \( x = L \) we add a buffer from \( x = L \) to \( x = LL \) in which we let the value of \( \gamma \) increase exponentially, viz.,

\[ \gamma = \begin{cases} \gamma_0; & 0 \leq x \leq L \\ \gamma_0 e^{\lambda(x-L)}; & L < x \leq LL \end{cases}, \]

(7.35)

where the parameter \( \lambda \) determines how fast or quickly the frictional effect increases.

We can derive an analytic solution to (7.33) and (7.34). We start by differentiating (7.34) with respect to time, and then substitute for \( \partial_t u \) from (7.33). We then get

\[ \partial_t^2 h + \gamma \partial_t h = gH \partial_x^2 h. \]

(7.36)

Searching for wave like solution we let

\[ h = h_0 e^{\omega t} e^{i \alpha x}. \]

(7.37)

Substituting this expression into (7.36) we get the dispersion relation

\[ \omega^2 + \gamma \omega + gH \alpha^2 = 0. \]

(7.38)

\[ h = h_0 e^{-\frac{1}{2} \gamma t} e^{i \alpha(x-ct)}, \]

(7.39)

where

\[ c = \sqrt{gH - \left( \frac{\gamma}{2 \alpha} \right)^2}, \]

(7.40)

\( \alpha \) is the wavelength and \( h_0 \) is a constant. We observe that as time evolves the amplitude of the solution within the sponge domain decreases. Furthermore, we notice that phase velocity
(7.40) decreases with increasing $\gamma$, that is, decreases as the wave propagates deeper into the sponge area. Thus as the solution to the system (7.33) and (7.34) propagates into the sponge where the frictional parameter $\gamma$ increases in accord with (7.35), then it will be suppressed while decelerated. If we want to retain these properties it is important that $\gamma$ increases slowly within the sponge. To achieve this the exterior, sponge layer (or zone) added to the interior domain must be of a certain finite length. If $\gamma$ increases too quickly then the wave solution will not behave like given by (7.39) and possible reflection of wave energy might be generated.

Since application of the sponge condition as an open boundary condition requires that the sponge zone is of a certain extension, it adds a large amount of computer time to our problem, and hence slows down the wall clock time. Another problem with the sponge condition is that if the solution consists of forced waves (cf. Røed and Cooper, 1986), for instance is governed by equations like

\begin{align}
\partial_t u &= -g \partial_x h - \gamma u + \tau \\
\partial_t h &= -H \partial_x u
\end{align}

(7.41) \hspace{1cm} (7.42)

where $\tau$ represents the forcing, then the solution in the sponge layer of the former wave solution and a solution dominated by a balance between the forcing term and the frictional term, that is, $u = \tau/\gamma$, which implies that as $\gamma$ increases $u$ decreases so that mass (volume) accumulates within the sponge zone. For longer term integrations this accumulation of volume changes the pressure forcing and sooner or later this will have an impact on the interior solution as well.

### 7.5 A weakly reflective OBC

As alluded to in Section 7.2 we may use the method of characteristics to construct a weakly reflective OBC also for problems including non-linearities, Coriolis affects and forcing (cf. Røed and Cooper, 1987).

As an example let us again study the problem governed by (7.5), where we add non-linear terms as well the Coriolis term and forcing,

\begin{align}
\partial_t u + u \partial_x u - fv &= -g \partial_x h + F^x, \\
\partial_t v + u \partial_x v + fu &= F^y, \\
\partial_t h + \partial_z (hu) &= 0,
\end{align}

(7.43) \hspace{1cm} (7.44) \hspace{1cm} (7.45)

where $F^x, F^y$ are the forcing terms, and $f$ is the Coriolis parameter. We immediately recognize the system (7.43) - (7.45) as the shallow water equations for a barotropic fluid. Multiplying (7.45) by a yet unknown function $\lambda$, and adding to (7.43), and defining a common operator by

\begin{align}
\frac{D^*}{dt} = \partial_t + \frac{D^* x}{dt} \partial_x,
\end{align}

(7.46)

we get that $\lambda$ must satisfy the equation

\begin{align}
\frac{D^* x}{dt} = u + \lambda h = u + \frac{q}{\lambda},
\end{align}

(7.47)
Thus
\[ \lambda_{1,2} = \pm \frac{c}{h}, \quad c = \sqrt{gh}, \] (7.48)
and hence the compatibility equations associated with (7.43) and (7.45) and the appropriate characteristic equations are
\[
\left( \frac{D^*}{dt} \right)_{1,2} (u \pm 2c) = fv + F^x \quad \text{along} \quad \left( \frac{D^*x}{dt} \right)_{1,2} = u \pm c. \quad (7.49)
\]
To avoid reflections at, say \( x = L \), we require
\[
\left. \left( \frac{D^*x}{dt} \right) \right|_{2}^{1} \bigg|_{x=L} = 0. \quad (7.50)
\]
By substitution of this expression in (7.49) the weakly reflective open boundary condition becomes,
\[
\partial_t (u + 2c) + (u + c)\partial_x (u + 2c) = fv + F^x, \quad (7.51)
\]
\[
\partial_t (u - 2c) = fv + F^x, \quad (7.52)
\]
valid at \( x = L \). A finite difference approximation to these equation then gives a weakly reflective OBC.

### 7.6 The Flow Relaxation Scheme

We now construct an OBC that was first suggested by Davies (1976). As shown below it is somewhat similar to the sponge OBC in two respects. First it requires us to extend the computational domain to include an exterior domain or buffer zone. Second it is in essence a sponge in which the solution is suppressed as it progresses into the buffer zone. The method is called the Flow Relaxation Scheme commonly abbreviated to FRS (Davies, 1976; Martinsen and Engedahl, 1987; Cooper and Thompson, 1989; Engedahl, 1995a; Shi et al., 1999, 2001). In particular the two latter references are useful in that they give a detailed description of the FRS and in addition gives a nice example of its use as an OBC.

One of the advantages of the FRS compared to for instance the sponge is that it allows us to specify an exterior solution. The FRS can therefore also be used as a one way nesting condition, in which an exterior solution is specified by, e.g., a coarser grid model covering a much larger area. The FRS as a nesting technique is for instance used as the main method whereby information from global and semi-global model are transferred to regional models at Meteorologisk institutt. This is true for both their numerical weather prediction (NWP) models as well as their numerical ocean weather prediction (NOWP) models.\footnote{cf. http://met.no/}

In essence the method just modifies the numerical solution in a buffer zone, the so called FRS zone, for each time step, based on the interior solution and a specified exterior solution. The
7.6 Flow relaxation

OPEN BOUNDARY CONDITIONS

FRS zone commonly consists of a small number of grid points, say 5-20. We emphasize that the FRS zone is an extension of the interior domain and thus extends the computational domain (cf. Figure 7.3). Within the FRS zone the solution is relaxed towards an a priori specified exterior solution, often referred to as the outer solution. The relaxation is performed by specifying a weighing function that for each grid point in the FRS zone computes a weighted mean between the specified outer solution and the interior solution computed from the governing equations.

Let \( \phi(x, t) \) be the dependent variable in our problem and let the interior domain be \( x \in (-M, L) \), where \( x = 0 \) is an open boundary. As displayed in Figure 7.3 the FRS zone extends the interior domain so the computational domain is increased to the left by adding an FRS zone starting at \( x = -M \) and ending at \( x = 0 \). As usual we define the grid points by \( x_j = (j - 1)\Delta x \), where the index \( j = 1(1)J \) covers all grid points of the computational domain, that is, both the FRS zone and the interior domain. Then \( j = J \) specifies the right-hand boundary of the computational domain, that is, \( x_J = L \), while \( j = 1 \) specifies the left-hand boundary of the FRS zone, that is, \( x_1 = -M \). Let the index of grid point where the FRS zone is attached to the interior domain (\( x = 0 \)) be \( j = JM \). Then \( x_JM = 0 \) defines the point which is at the left-hand open boundary. As alluded to the FRS allows us to specify an outer solution, which can be the result of another numerical model covering a larger domain than our interior domain. We denote this exterior solution by \( (\phi_e)^n_j \) which emphasize that the outer solution is a function of space and time. Let us now assume that we have computed all the our dependent variable \( \phi^n_j \) at all points, including the FRS zone at time level \( n \). Furthermore using the governing equation of our model we can derive a solution at the time level \( n + 1 \) including within the FRS zone, except of course at \( j = 1 \). We denote this predictor be denoted by \( \phi^*_j \). The next step is to correct the solution in

Figure 7.3: Sketch of the FRS zone, the interior domain and the computational domain. Also shown are the appropriate indices.
the FRS zone by computing our dependent variable as a weighted mean between our predicted solution \( \phi_j^* \) and the specified outer solution \( (\phi_e)_j^n \) to derive the final or corrected solution at time level \( n+1 \). We do this by employing the formula

\[
\phi_j^{n+1} = (1 - \alpha_j)\phi_j^* + \alpha_j(\phi_e)_j^{n+1}; \quad j = 1(1)JM,
\]

(7.53)

where \( \alpha_j \) denotes the weights. By letting \( \alpha_j \) vary from \( \alpha_1 = 1 \) at the outer edge of the FRS zone, \( (j = 1) \) to \( \alpha_{JM} = 0 \) at the open boundary at \( j = JM \), we notice from (7.53) that at the outer edge of the FRS zone the numerical solution at time level \( n+1 \) equals the specified outer solution, or \( \phi_1^{n+1} = (\phi_e)_1^{n+1} \). Similarly we notice that at the open boundary the solution equals the interior solution, or \( \phi_{JM}^{n+1} = \phi_{JM}^* \).

Experiments employing the FRS shows that the solution is sensitive to the distribution of the specified weighting function \( \alpha \) throughout the FRS zone. It is also sensitive to the width of the FRS zone, e.g., Martinsen and Engedahl (1987), Engedahl (1995a). They found that distributing \( \alpha \) applying a hyperbolic tangent function, that is,

\[
\alpha_j = 1 - \tanh \left( \frac{j - 1}{2} \right); \quad j = 1(1)JM,
\]

(7.54)

is a good choice. Furthermore they concluded that for oceanic application the width of the FRS zone should be at least seven grid points, that is, \( JM \geq 7 \).

One of the disadvantages of employing FRS as an OBC (or nesting technique) is that the computational domain is increased, and hence that the computational burden is increased. This disadvantage is somewhat suppressed by the fact that the FRS allows us to specify an outer solution, which as shown below can be used to effectively minimize the error. Another disadvantage is that solution when applying the FRS as an OBC does not conserve fundamental properties such as volume (or mass).

As an example let us study the numerical solution of the continuous problem

\[
\partial_t \phi = \mathcal{L}[\phi] \quad ; \quad x \in <0, L>,
\]

(7.55)

where \( \mathcal{L} \) is a spatial differential operator. Furthermore, let us assume that the open boundary is at \( x = 0 \) and that \( x = L \) is a natural boundary. As above we let \( (\phi_e)_j^n; \quad j = 1(1)JM \) denote the specified exterior solution. If we solve (7.55) applying a forward in time finite difference scheme then we get

\[
\phi_j^* = \phi_j^n + \Delta t \mathcal{L}_j^n; \quad j = 2(1)J - 1,
\]

(7.56)

where \( \phi_j^* \) is the predictor. We then correct the predictor by applying the relaxation formula (7.53). We then get

\[
\phi_j^{n+1} = (1 - \alpha_j)\phi_j^* + \alpha_j(\phi_e)_j^{n+1}; \quad j = 1(1)J,
\]

(7.57)

where we without loss of generality have extended (7.53) to the entire computational domain. To ensure that we do no corrections to the predictor within the interior domain we redefine the weighting function \( \alpha_j \) as

\[
\alpha_j = \begin{cases} 
1 - \tanh \left( \frac{j - 1}{2} \right) & ; \quad j = 1(1)JM - 1 \\
0 & ; \quad j = JM(1)J 
\end{cases}
\]

(7.58)
We may now use the expression on the right-hand side of (7.56) to substitute for $\phi^*_j$ in (7.57). If we in addition add the zero $\alpha_j \phi_{j}^{n+1} - \alpha_j \phi_{j}^{n}$ we get

$$\frac{\phi_{j}^{n+1} - \phi_{j}^{n}}{\Delta t} = L_{j}^{n} + \gamma_j \left[(\phi_e)^{n+1} - \phi_{j}^{n+1}\right]$$

(7.59)

where the coefficient $\gamma_j$ is defined by,

$$\gamma_j = \frac{\alpha_j}{1 - \alpha_j}.$$  

(7.60)

If we now let $\Delta t$ and $\Delta x$ tend to zero, we notice that (7.59) is a forward in time, finite difference approximation to the continuous equation

$$\partial_t \phi = L[\phi] + \gamma (\phi_e - \phi) ; \ x \in (-\infty, \infty).$$

(7.61)

We observe that except for the additional “frictional” term $\gamma (\phi_e - \phi)$ equation (7.61) equals (7.55). We also notice that the additional term is proportional to the difference between the interior solution and the exterior solution, and that the proportionality factor (7.60) varies from zero at the open boundary ($x = 0$) to infinity at the edge of the FRS zone. Thus the relative importance of the frictional term increases as we progress into the FRS zone. If we now specify an exterior solution as constant and zero, then (7.61) is turned into

$$\partial_t \phi = L[\phi] - \gamma \phi.$$  

(7.62)

Under these circumstances the FRS acts like a sponge with a frictional parameter $\gamma$ which gradually and monotonically increases towards infinity as we progress into the sponge, not unlike the exponential function specified in Section 7.4.

To illustrate the non-conservative properties of the FRS we use the example problem governed by (7.5) and (7.6). To get started we notice first that as $\Delta x$ and $\Delta t$ tend to zero (7.57) takes on the form

$$\phi = (1 - \alpha)\phi^* + \alpha \phi_e.$$  

(7.63)

We now have two dependent variables, namely $h$ and $u$, hence

$$h = (1 - \alpha)h^* + \alpha h_e,$$

(7.64)

$$u = (1 - \alpha)u^* + \alpha u_e.$$  

(7.65)

Substituting for $u$ from (7.65) into (7.6) we get

$$\partial_t h = -H \partial_x u = (1 - \alpha)\partial_x u^* - H \partial_x u_e + H u^* \partial_x \alpha$$

(7.66)

The two first terms on the right-hand-side of (7.66) are the terms we would have obtained if we relax (7.5) directly assuming that (7.6) is valid for the interior as well as the exterior solution. However, since the relaxation parameter is a function of $x$ also a term appears containing the gradient of the relaxation parameter $\alpha$. Thus the volume (or mass) conservation expressed by
(7.6) is invalidated. To avoid the non-conservation of mass to impact our interior solution we must ensure that the relaxation parameter $\alpha$ is a slowly varying function. In turn this implies that the width of the FRS zone must be long enough for this to be realized.

Finally we notice that if the exterior solution is equal to or close to the true solution, then the friction term in (7.62) disappear as do the false divergence in (7.66). Under these circumstances the FRS is close to being a perfect open boundary or nesting condition. Thus the usefulness of the FRS depends to a certain extent on how good we are to “guess” the exterior solution. This is the reason why the FRS is mostly used as a one way nesting condition. When specifying the exterior solution to be the solution of the same governing equation, albeit for a coarser mesh, we ensure that the exterior solution is indeed close to the interior solution of the fine mesh model embedded in the coarser grid.

**Exercises**

1. Show that a one-sided, finite difference scheme in time and space of the radiation condition (7.1) can be written

\[
\phi_{B}^{n+1} = \begin{cases} 
\phi_{B}^{n} & ; \quad c_{\phi} > 0 \\
(1 + c_{\phi} \frac{\Delta t}{\Delta x}) \phi_{B}^{n} - c_{\phi} \frac{\Delta t}{\Delta x} \phi_{B+1}^{n} & ; \quad c_{\phi} \leq 0
\end{cases} \quad (7.67)
\]

The open boundary is to the left so that subscript $B$ denotes the values of the variables on the open boundary while subscript $B + 1$ indicates the values to the right of the open boundary.

2. Show by use of (7.67) that the radiation condition is “simply” an interpolation of values on the inside of the computational domain.
Chapter 8

GENERAL VERTICAL COORDINATES

Most modern models employed in the meteorological and oceanographic community replace the natural geopotential vertical coordinate (z-coordinate) with a new vertical coordinate. The reason for this is that the geopotential coordinate is quite cumbersome to work with in the presence of steep topography such as mountains in the atmosphere and shelf breaks and sea mountains in the ocean.

As early as the late 1940s Sutcliffe (1947) and Eliassen (1949), at the dawn of numerical weather prediction, suggested to use pressure surfaces to replace surfaces of geopotential height as the vertical coordinate in atmospheric models, a method successfully tested by Charney and Phillips (1953) using a quasi-geostrophic model (cf. Section 1.6 on page 9). The pressure coordinate has several advantages over ordinary geopotential height models. For instance it reduces the mass conservation equation to a diagnostic equation, which in turn eases the analysis of the large scale (hydrostatic) motions. The pressure coordinate, however, has certain computational disadvantages, in particular in the vicinity of mountains since the ground is not a pressure surface. To remedy this Phillips (1957) suggested to use terrain-following surfaces as the vertical coordinate. Such a coordinate system is now commonly referred to as the σ-coordinates, a coordinate system that has become quite popular in ocean models (Blumberg and Mellor, 1987; Haidvogel et al., 2008). Also other vertical coordinate systems are suggested. For instance it was suggested by Shapiro and Hastings (1973) to use surfaces of potential temperature as the vertical coordinate. This was tested in a quasi-geostrophic weather prediction model by Bleck (1973) and later in primitive equation models, so-called isentropic models, by Eliassen and Raustein (1968). In the ocean so called isopycnic models in which surfaces of potential density are used as vertical coordinates was explored in primitive equation models by Bleck and Smith (1990). In the recent decade it has also become quite common to explore the use of so called hybrid coordinate models in which the vertical coordinate changes from one to another throughout the height both in the atmosphere and in the ocean (e.g., Bleck, 2002). Finally it should be emphasized, as summarized in (Griffies, 2004, Chapter 6) that the various vertical coordinate systems all have their advantages and disadvantages.

In the following we will first show how we in general transform equations formulated in geopotential coordinates (Cartesian or z-coordinates) to a new general vertical coordinate, say \( s = s(x, y, z, t) \). To this end we follow the derivation made by Kasahara (1974). We then show
how the governing equations of a hydrostatic, non-Boussinesq fluid (cf. Section 1.3 on page 4) is affected by such a transformation. We end this chapter by showing an explicit example using the \( \sigma \)-coordinates as our example.

8.1 Transformation to a general vertical coordinate

In general we transform from one coordinate system of independent variables, say \((x, y, z, t)\), to another system, say \((x', y', s, t')\), by specifying how the independent variables in the transformed system depend on the independent variables of the original system. Here we will only replace the vertical height coordinate \(z\). Accordingly we define the transformation simply by

\[
x' = x, \quad y' = y, \quad s = s(x, y, z, t), \quad \text{and} \quad t' = t,
\]

where we have only replaced the normal geopotential height coordinate \(z\) with a general vertical coordinate \(s\) while the horizontal coordinates are left unchanged in three-dimensional space. To ensure that the transformation is unique we must require that \(s\) is a monotone function of height \(z\). Mathematically this means that the gradient of \(s\) with respect to \(z\) does not change sign within a fluid column, or

\[
\frac{\partial z}{\partial s} \geq 0, \quad \text{and} \quad \frac{\partial z}{\partial s} \neq 0. \quad (8.2)
\]

This is also a necessary condition to ensure that the inverse transformation \(z = z(x', y', s, t')\) exists as well.

Let \(\xi'\) denote any of the three independent variables \(x', y', \text{and} t'\) in the new coordinate system, and similarly let \(\xi\) denote any of the three independent variables \(x, y, \text{and} t\) in the original system. Then the transformation (8.1) gives

\[
\frac{\partial z}{\partial \xi'} = 0, \quad \frac{\partial z}{\partial x'} = \frac{\partial z}{\partial x} = \frac{\partial z}{\partial y'} = 0, \quad \text{and} \quad \frac{\partial z}{\partial t'} = \frac{\partial z}{\partial t} = 0, \quad (8.3)
\]

while

\[
\frac{\partial \xi'}{\partial x'} = \frac{\partial \xi'}{\partial y'} = \frac{\partial \xi'}{\partial t'} = 1. \quad (8.4)
\]

Similarly follows

\[
\frac{\partial s}{\partial \xi} = 0, \quad \frac{\partial s}{\partial x} = \frac{\partial s}{\partial y} = 0, \quad \frac{\partial s}{\partial x'} = \frac{\partial s}{\partial y'} = 0, \quad \text{and} \quad \frac{\partial s}{\partial t} = \frac{\partial s}{\partial t'} = 0, \quad (8.5)
\]

while

\[
\frac{\partial \xi}{\partial x} = \frac{\partial \xi}{\partial y} = \frac{\partial \xi}{\partial t} = 1. \quad (8.6)
\]

We emphasize that \(s\) is monotonic with respect to \(z\), which implies that \(\partial z s \neq 0\) and \(\partial z s \neq 0\). We also observe that if we transform \(z\) to \(z\), that is, let \(s = z\) then \(\partial z s = \partial z z = 1\).

Let \(\psi = \psi(x, y, z, t) = \psi(x', y', s, t')\) denote any scalar. Then the first property of the transformation is

\[
\frac{\partial z}{\partial s} s = \frac{\partial z}{\partial s} \frac{\partial s}{\partial \psi} \psi. \quad (8.7)
\]

If we take the derivative of \(\psi\) with respect to one of the independent variables in the coordinate system we transform to, say \(t'\), then we get

\[
\frac{\partial \psi}{\partial t'} = \frac{\partial \psi}{\partial t} \frac{\partial t}{\partial t'} + \frac{\partial \psi}{\partial x} \frac{\partial x}{\partial t'} + \frac{\partial \psi}{\partial y} \frac{\partial y}{\partial t'} + \frac{\partial \psi}{\partial z} \frac{\partial z}{\partial t'} = \frac{\partial \psi}{\partial t} + \frac{\partial \psi}{\partial s} \frac{\partial s}{\partial \psi} \frac{\partial s}{\partial t'}, \quad (8.8)
\]
where the last equal sign follows by utilizing (8.3) - (8.7). If we solve (8.8) with respect to $\partial_t \psi$ we further get

$$\partial_t \psi = \partial_t' \psi - \partial_z s \partial_s \psi \partial_t' z. \quad (8.9)$$

Similarly follows that

$$\partial_x \psi = \partial_x' \psi - \partial_z s \partial_s \psi \partial_x' z, \quad \text{and} \quad \partial_y \psi = \partial_y' \psi - \partial_z s \partial_s \psi \partial_y' z. \quad (8.10)$$

Let us define the horizontal gradient of $\psi$ in the new coordinate system by

$$\nabla_s \psi = i \partial_x' \psi + j \partial_y' \psi. \quad (8.11)$$

Then making use of (8.9) and (8.10) we obtain

$$\nabla_H \psi = \nabla_s \psi - \partial_z s \partial_s \psi \nabla_s z. \quad (8.12)$$

Furthermore we find that the horizontal divergence of any vector, say $\mathbf{a}$, transforms as

$$\nabla_H \cdot \mathbf{a} = \nabla_s \cdot \mathbf{a} + \partial_z s \partial_s \mathbf{a} \cdot \nabla_s z. \quad (8.13)$$

We note that all vectors project onto the horizontal geopotential surface. This is also true for the gradient (8.12). Thus the metric term associated with the vertical gradient of the surface $s$ in the geopotential coordinate system is eliminated.

We note that since the individual derivative$^1$ is independent of coordinate transformation we get

$$\frac{D \psi}{dt} = \partial_t' \psi + \mathbf{u} \cdot \nabla_H \psi + w \partial_z \psi \quad (8.14)$$

$$= \partial_t' \psi + \mathbf{u} \cdot \nabla_s \psi + \dot{s} \partial_s \psi, \quad (8.15)$$

where

$$\dot{s} = \frac{Ds}{dt} \quad (8.16)$$

is the speed of the surface $s$ in the direction of the three-dimensional velocity. Note that (8.14) is the derivative expressed in the geopotential coordinate system while (8.15) is the material derivative expressed in the new general vertical coordinate system. We now make use of (8.9) - (8.12) to replace the appropriate terms in (8.14). Then we get

$$\frac{D \psi}{dt} = \partial_t' \psi + \mathbf{u} \cdot \nabla_s \psi + (w - \partial_t' z - \mathbf{u} \cdot \nabla_s z) \partial_z s \partial_s \psi. \quad (8.17)$$

Furthermore, by use of (8.15) we get

$$\dot{s} = (w - \partial_t' z - \mathbf{u} \cdot \nabla_s z) \partial_z s \equiv \omega \partial_z s, \quad (8.18)$$

---

$^1$ Also by many authors referred to as the material derivative
where the identity in (8.18) defines the velocity $\omega$ by

$$\omega = w - (\partial_t z + \mathbf{u} \cdot \nabla_s z) = w - (\partial_t s + \mathbf{u} \cdot \nabla_H s) \partial_s z. \quad (8.19)$$

First we observe that if $s = z$ then $\omega = w$ as expected, and the velocity $\omega$ is interpreted as the speed by which the surface $s$ travels through the fixed $z$ levels of the geopotential coordinate system. Next we observe that if $s$ is a material surface the kinematic boundary condition is $w = (\partial_t s + \mathbf{u} \cdot \nabla_H s)$ Under these circumstance it follows that $\omega = 0$. This is again to be expected since a material surface is a surface that consists of the same fluid particles for all times. If the $s$ surface is not a material surface then $\omega \neq 0$ and is associated with the speed of the fluid particles through the surface $s$. The difference between $\omega$ and $w$ is hence associated with the speed of the surface $s$ in the geopotential coordinate system. Thus $\omega$ is interpreted as that part of the vertical movement of particles that remains when moving with the surface $s$.

### 8.2 Transformation of the governing equations

To give insight into how the transformation is applied, we apply it to a non-Boussinesq, hydrostatic fluid.

**The hydrostatic equation**

We start by transforming the hydrostatic equation

$$\partial_z p + \rho g = 0. \quad (8.20)$$

Using the transformation formulas of the previous section we get

$$\partial_s p + \rho g \partial_s z = 0. \quad (8.21)$$

We may use this equation to determine the metric factors $\partial_s z$ and $\partial_z s$ as follows

$$\partial_s z = -\frac{\partial_s p}{\rho g}, \text{ and } \partial_z s = -\frac{\rho g}{\partial_s p}. \quad (8.22)$$

We note in passing that if $s = p$ then (8.21) reduces to

$$1 + \rho g \partial_p z = 0. \text{ or } \partial_p z = -\frac{1}{\rho g}. \quad (8.23)$$

**Mass conservation**

Next, we transform the continuity equation

$$\partial_t \rho + \nabla \cdot (\mathbf{v} \rho) = 0. \quad (8.24)$$
We first rewrite this equation to yield
\[ \frac{1}{\rho} \frac{D\rho}{dt} + \nabla_H \cdot u + \partial_z w = 0. \] (8.25)

We then make use of the transformation formulas to obtain
\[ \frac{1}{\rho} \frac{D\rho}{dt} + \nabla_H \cdot u + \partial_z w = - \rho (\partial_s \alpha + u \cdot \nabla_s \alpha + \dot{s} \partial_s \alpha) + \partial_s s [\partial_s (\partial_s z) + \nabla_s \cdot (u \partial_s z) + \partial_s (\dot{s} \partial_s z)], \] (8.26)

where \( \alpha = 1/\rho \). To arrive at this result we have also solved (8.19) with respect to \( w \) to replace \( \partial_s w \). We may further develop (8.26) by making use of (8.22) to replace the metric term \( \partial_s z \).

Thus we get
\[ \frac{1}{\rho} \frac{D\rho}{dt} + \nabla_H \cdot u + \partial_z w = (\partial_s p)^{-1} \left[ \partial_s (\partial_s p) + \nabla_s \cdot (u \partial_s p) + \partial_s (\dot{s} \partial_s p) \right], \] (8.27)

and hence the transformed continuity equation reads
\[ \partial_s (\partial_s p) + \nabla_s \cdot (u \partial_s p) + \partial_s (\dot{s} \partial_s p) = 0. \] (8.28)

**Energy equation**

If we apply a similar procedure to the tracer equation (1.13) we get
\[ \partial_t C + u \cdot \nabla_s C + \dot{s} \partial_s C = \mathcal{F}_C + S_C \] (8.29)

where the right-hand side represents the transformed fluxes and source terms.

**The momentum equation**

We finally transform the horizontal component of the momentum equation for a non-Boussinesq, hydrostatic fluid (1.12) by first rewriting it to read
\[ \frac{D}{Dt} u + f k \times u = -\alpha \nabla_H p + \alpha \partial_z \tau + \nabla_H \cdot \mathcal{F}_M \] (8.30)

where \( \tau \) is the vertical mixing or flux vector, sometimes referred to as the the vertical shear stress. To transform this equation is a bit more complicated so we treat it term by term.

We first consider the pressure term, which is special. For a non-Boussinesq fluid we get
\[ \alpha \nabla_H p = \alpha \nabla_s p + g \nabla_s z = \nabla_s M - p \nabla_s \alpha \] (8.31)

where
\[ M = \alpha p + g z \] (8.32)
is the Montgomery potential (or stream function). In the case \( s = \rho \) the last term in (8.32) vanishes since then \( \nabla_s \alpha = 0 \). Under these circumstances the Montgomery potential becomes a true potential and is a streamfunction for the geostrophic velocity. For any other choice of \( s \), however, the last term in (8.32) must be retained. We finally note that the Montgomery potential appears because all vectors are projected onto the horizontal surface (with respect to gravity), even though all gradients are evaluated in the transformed \( x', y', s, t' \) system.

Next we consider the vertical shear stress term. In this we apply (8.7) and (8.22) to get

\[ \alpha \tau = \alpha \partial_z \sigma \partial_s \tau = -g \frac{\partial_p \tau}{\partial_p} = \partial_p \tau. \] (8.33)

Recalling that

\[ \frac{Du}{dt} = \partial_t u + u \cdot \nabla_s u + \dot{s} \partial_s u \] (8.34)

and that

\[ u \cdot \nabla_s u = \nabla_s \left( \frac{1}{2} u^2 \right) + \zeta k \times u \] (8.35)

where \( \zeta = k \cdot \nabla_s \times u \) is the relative vorticity relative to the new coordinate system. Hence the momentum equation becomes

\[ \partial_t u + \nabla_s \left( \frac{1}{2} u^2 \right) + (\zeta + f) k \times u + \dot{s} \partial_s u = -\nabla_s M + p \nabla_s \alpha - g \partial_p \tau + \nabla_s \cdot F_M^H. \] (8.36)

We may also write this equation in flux form. We then first recombine the second and third term on the left-hand side of (8.36) using (8.35). Next we multiply (8.36) by \( \partial_s p \) and then finally make use of the continuity equation in the form (8.28). We then get

\[ \partial_t (u \partial_s p) + \nabla_s \cdot (uu \partial_s p) + f k \times u \partial_s p + \partial_s (\dot{s} u \partial_s p) \]

\[ = -\partial_s p (\nabla_s M + p \nabla_s \alpha) - g \partial_p \tau + \partial_s p \nabla_s \cdot F_M^H. \] (8.37)

As alluded to earlier when treating the diffusion problem, we emphasize at this point that the mixing term or “diffusion” term is mostly added to prevent our numerical model from blowing up. Hence its exact transformation is of secondary importance.

### 8.3 Terrain-following coordinates

As an example we apply these transformation to transform the mass conservation equation to the so called \( \sigma \)-coordinate models. This particular coordinate system is defined by

\[ s = \sigma = \frac{z - \eta}{D} \quad \text{or} \quad z = \sigma D + \eta, \] (8.38)

where \( D = H + \eta \) is the total depth, \( \eta \) being the deviation of the upper surface from its equilibrium position and \( H \) is the equilibrium depth of the fluid columns. The terrain following coordinate
models are very popular in the oceanographic community, e.g., ROMS (Haïdvygel et al., 2008), and various versions of POM (Blumberg and Mellor, 1987; Engedahl, 1995a). It is also to some extent applied in numerical weather predictions models (Phillips, 1957; Kasahara, 1974).

First we note that the metric factor \( \partial_s s \) and \( \partial_s z \) using (8.22) becomes

\[
\partial_s s = \partial_z \sigma = \frac{1}{D} \quad \text{and} \quad \partial_s z = \partial_\sigma z = D, \tag{8.39}
\]

which allows us to rewrite the hydrostatic equation to

\[
\partial_\sigma p = -\rho g D. \tag{8.40}
\]

Furthermore we need to know the speed \( \omega \) trough the \( \sigma \) surfaces. Applying (8.38) we get

\[
\omega = w - \sigma \partial_s D - \partial_\tau \eta - \sigma \mathbf{u} \cdot \nabla_\sigma D - \mathbf{u} \cdot \nabla_\sigma \eta. \tag{8.41}
\]

Thus the mass conservation equation in the form (8.28) becomes

\[
\partial_t (\rho D) + \nabla_\sigma \cdot (\rho D \mathbf{u}) + \partial_\sigma (\rho D) = 0, \tag{8.42}
\]

Separating the effect of the density we get

\[
\frac{D \rho}{dt} + \frac{\rho}{D} [\partial_t D + \nabla_\sigma \cdot (D \mathbf{u}) + \partial_\sigma (\rho D)] = 0, \tag{8.43}
\]

We observe using (8.18) that \( \dot{\sigma} = \omega \partial_z \sigma = \omega D^{-1} \). Furthermore we note that \( \partial_t D = \partial_\tau \eta \). Substitution of these expressions into (8.43), and invoking the Boussinesq approximation (1.16) or \( \frac{D \rho}{dt} = 0 \), the continuity equation for a Boussinesq fluid in terrain-following coordinates is

\[
\partial_t \eta + \nabla_\sigma \cdot (D \mathbf{u}) + \partial_\sigma \omega = 0. \tag{8.44}
\]

We note that the remaining equations may be derived from their general expressions in a similar fashion. For instance using (8.40) the momentum equation in the flux form (8.37) becomes

\[
\partial_t (D \mathbf{u}) + \nabla_s \cdot (D \mathbf{u} \mathbf{u}) + f k \times D \mathbf{u} + \partial_\sigma (\omega \mathbf{u}) \\
= -D (\nabla_s M + p \nabla_s \alpha) - \frac{1}{\rho_0} \partial_\sigma \tau + D \nabla_s \cdot \mathbf{F}_M. \tag{8.45}
\]
Chapter 9

TWO-DIMENSIONAL PROBLEMS

Below we investigate the effect of including more than one-dimension in space on the numerical stability criterion. In particular we study the effect on the finite difference equations relating to the diffusion, advection and shallow water problems. Further expansion into three dimensions is then straightforward.

9.1 Diffusion equation

We start by expanding the simple one-dimensional diffusion equation to two dimensions in space. Thus we consider the continuous equation

$$\frac{\partial t}{\partial t} \theta = \kappa (\frac{\partial^2 \theta}{\partial x^2} + \frac{\partial^2 \theta}{\partial y^2}).$$  \hspace{1cm} (9.1)

As in the one-dimensional case (cf. Chapter 4) we employ the forward in time, centered in space scheme (FTCS). Thus we get

$$\theta_{jk}^{n+1} = \theta_{jk}^n + \kappa \frac{\Delta t}{\Delta x^2} (\theta_{j-1k}^n - 2\theta_{jk}^n + \theta_{j+1k}^n) + \kappa \frac{\Delta t}{\Delta y^2} (\theta_{jk-1}^n - 2\theta_{jk}^n + \theta_{jk+1}^n),$$  \hspace{1cm} (9.2)

where we have used the notation given in Section 2.9 of Chapter 2 (cf. equation 2.48 on page 22), and where $\Delta y$ is the space increment along the $y$ axis. To investigate the numerical stability of this scheme we still use von Neumann’s method. As in the one-dimensional case we first substitute $\theta_{jk}^n$ by its individual Fourier components. Since we now must allow for waves propagating in both of the two horizontal directions, the Fourier component must include waves propagating in the $x$ direction as well as in the $y$ direction, that is,

$$\theta_{jk}^n = \Theta_n e^{\alpha j \Delta x} e^{\beta k \Delta y},$$  \hspace{1cm} (9.3)

where $\alpha$ and $\beta$ are wavenumbers in the $x$ and $y$ directions, respectively. Next we insert the discrete Fourier component into (9.2) and solve for the growth factor. We then get

$$G = 1 - 2\kappa \frac{\Delta t}{\Delta x^2} (1 - \cos \alpha \Delta x) - 2\kappa \frac{\Delta t}{\Delta y^2} (1 - \cos \beta \Delta y).$$  \hspace{1cm} (9.4)
We observe that this expression is comparable to the one derived for the one-dimensional case, that is (4.31) on page 43, except that we have an additional term due to the two-dimensionality of (9.1). Applying von Neumann’s criterion for numerical stability (4.28) we still must require $|G| \leq 1$. Hence

$$-1 \leq 1 - 2\kappa \frac{\Delta t}{\Delta x^2} (1 - \cos \alpha \Delta x) - 2\kappa \frac{\Delta t}{\Delta y^2} (1 - \cos \beta \Delta y) \leq 1.$$  \hspace{1cm} (9.5)

As in the one-dimensional case we note that the right-hand inequality is trivially satisfied. To satisfy the left-hand inequality we must require

$$\kappa \frac{\Delta t}{\Delta x^2} (1 - \cos \alpha \Delta x) + \kappa \frac{\Delta t}{\Delta y^2} (1 - \cos \beta \Delta y) \leq 1$$  \hspace{1cm} (9.6)

Since the left-hand side of (9.6) is maximum when $\cos \alpha \Delta x = \cos \beta \Delta y = -1$ we note that if

$$\kappa \frac{\Delta t}{\Delta x^2} + \kappa \frac{\Delta t}{\Delta y^2} \leq \frac{1}{2}$$  \hspace{1cm} (9.7)

then the stability criterion is satisfied for all possible values of the wavenumbers $\alpha$ and $\beta$. Thus the time step is limited according to

$$\Delta t \leq \frac{\Delta x^2 \Delta y^2}{2\kappa \Delta x^2 + \Delta y^2}.$$  \hspace{1cm} (9.8)

In the special case when the grid is regular (a square grid), that is $\Delta x = \Delta y = \Delta s$, we get

$$\Delta t \leq \frac{\Delta s^2}{4\kappa}.$$  \hspace{1cm} (9.9)

If we compare (9.9) with the one dimensional case (4.34) we notice that the allowed time step is reduced by a factor of two. Thus the inclusion of more than one dimension implies that the criterion for numerical stability becomes more stringent. In fact this is a general result which applies to all problems.

### 9.2 Advection equation

Next we expand the simple one-dimensional advection equation to two dimensions in space. We then get

$$\partial_t \theta + u_0 \partial_x \theta + v_0 \partial_y \theta = 0,$$  \hspace{1cm} (9.10)

where $u_0$ and $v_0$ are constant speeds in the $x$ and $y$ direction respectively. To solve (9.10) numerically let us employ the well known second order accurate CTCS (leapfrog) scheme that worked well for the one-dimensional case. Thus

$$\frac{\theta^{n+1}_{jk} - \theta^{n-1}_{jk}}{2\Delta t} + u_0 \frac{\theta^n_{j+1k} - \theta^n_{j-1k}}{2\Delta x} + v_0 \frac{\theta^n_{jk+1} - \theta^n_{jk-1}}{2\Delta y} = 0.$$  \hspace{1cm} (9.11)
To investigate the numerical stability we again employ von Neumann’s method. Hence we insert the discrete Fourier component (9.3) into (9.11). We then get
\[ \Theta_{n+1} = \Theta_{n-1} - 2i \Theta_n \left( u_0 \frac{\Delta t}{\Delta x} \sin \alpha \Delta x + v_0 \frac{\Delta t}{\Delta y} \sin \beta \Delta y \right) \] (9.12)

Thus the equation for the growth factor becomes
\[ G^2 + 2i \lambda G - 1 = 0 ; \quad \lambda = u_0 \frac{\Delta t}{\Delta x} \sin \alpha \Delta x + v_0 \frac{\Delta t}{\Delta y} \sin \beta \Delta y, \] (9.13)
giving the two solutions
\[ G_{1,2} = i\lambda \pm \sqrt{1 - \lambda^2} \] (9.14)

If we compare with the one-dimensional problem we observe that the only difference is in the definition of \( \lambda \). As earlier we must require that the radical is a positive definite quantity and hence that
\[ |u_0 \frac{\Delta t}{\Delta x} \sin \alpha \Delta x + v_0 \frac{\Delta t}{\Delta y} \sin \beta \Delta y| \leq 1 \] (9.15)

For this to be valid for all possible choice of wavenumbers \( \alpha \) and \( \beta \) we must require
\[ |u_0| \frac{\Delta t}{\Delta x} + |v_0| \frac{\Delta t}{\Delta y} \leq 1 \quad \text{or} \quad \Delta t \leq \frac{\Delta x \Delta y}{|u_0| \Delta y + |v_0| \Delta x}. \] (9.16)

If we let the grid be regular, that is, let \( \Delta x = \Delta y = \Delta s \) and \( u_0 = v_0 = c_0 \) we get
\[ \Delta t \leq \frac{\Delta s}{2|c_0|}. \] (9.17)

Thus we observe, as in the diffusion problem, that increasing the dimension from one to two leads to a more stringent stability condition. This is to be expected since the physical interpretation of the CFL condition says that the characteristic must be within the cone of influence from time level \( n \).

### 9.3 Shallow water equations

#### Analytic solutions

As we did for the one-dimensional case it is worthwhile to first analyze the various wave motions supported by the two-dimensional problem. The two dimensional, linear rotating shallow water equation are given by (6.17) and (6.18). For convenience we repeat them her in scalar form. Thus in two dimensions we get
\[ \partial_t u + \partial_x \partial_x u + \partial_y \partial_y u - f v + \partial_x \phi = 0, \] (9.18)
\[ \partial_t v + \partial_x \partial_x v + \partial_y \partial_y v + f u + \partial_y \phi = 0, \] (9.19)
\[ \partial_t \phi + \phi (\partial_x u + \partial_y v) = 0. \] (9.20)
To analyze the possible wave motions we assume that all the variables are two dimensional waves of frequency \( \omega \). Thus we assume that the solution is
\[
x = x_0 e^{-i\omega t} e^{i(\alpha x + \beta y)},
\]
where \( \alpha \) and \( \beta \) are wave numbers in the \( x \)- and \( y \)-direction, respectively. We note that the frequency and the two wave numbers are all assumed to be real quantities. The dependent variable are contained in the vector \( x \), that is,
\[
x = \begin{bmatrix} u \\ v \\ \phi \end{bmatrix} \quad \text{and} \quad x_0 = \begin{bmatrix} u_0 \\ v_0 \\ \phi_0 \end{bmatrix},
\]
which in turn may be formulated as the homogeneous linear equation,
\[
\mathbf{A} \cdot x = 0,
\]
where \( \mathbf{A} \) is
\[
\mathbf{A} = \begin{bmatrix} i(\alpha \bar{u} + \beta \bar{v} - \omega) & -f & i\alpha \\ f & i(\alpha \bar{u} + \beta \bar{v} - \omega) & i\beta \\ i\alpha \bar{\phi} & i\beta \bar{\phi} & i(\alpha \bar{u} + \beta \bar{v} - \omega) \end{bmatrix}.
\]
For non-trivial solutions to exists, the determinant of the tensor \( \mathbf{A} \) must be zero, which gives
\[
i(\alpha \bar{u} + \beta \bar{v} - \omega) [i(\alpha \bar{u} + \beta \bar{v} - \omega)^2 - \bar{\phi}(\alpha^2 + \beta^2) + f^2] = 0.
\]
As in the one-dimensional case we find that we get three solutions for the frequency \( \omega \), namely
\[
\omega_1 = \bar{u} \alpha + \bar{v} \beta, \\
\omega_2 = \bar{u} \alpha + \bar{v} \beta + \sqrt{c_0^2(\alpha^2 + \beta^2) + f^2}, \\
\omega_3 = \bar{u} \alpha + \bar{v} \beta - \sqrt{c_0^2(\alpha^2 + \beta^2) + f^2},
\]
where
\[
c_0 = \sqrt{\bar{\phi}} = \sqrt{gH}
\]
is the wave phase speed. The first solution is simply the geostrophic balance as displayed in (1.40) on page 9 with \( \phi = gh \), that is,
\[
u = \frac{1}{f} \mathbf{k} \times \nabla_H \phi.
\]
We easily derive this interpretation by substituting $\omega_1$ from (9.29) into (9.23) and (9.24), respectively, that is,

\begin{align*}
0 - f v + i\alpha \phi &= 0, \quad (9.34) \\
0 + f u + i\alpha \phi &= 0, \quad (9.35)
\end{align*}

which gives

\begin{align*}
u &= -\frac{1}{f} i\beta \phi \quad \text{and} \quad v = \frac{1}{f} i\alpha \phi \Rightarrow u = -\frac{1}{f} \partial_y \phi \quad \text{and} \quad v = \frac{1}{f} \partial_x \phi. \quad (9.36)
\end{align*}

The last implication follows by using the Fourier solution backwards and shows that the geostrophic balance (9.33) is recovered.

The two other solutions represented by $\pm \sqrt{\bar{\phi}(\alpha^2 + \beta^2) + f^2}$ are combined inertia and gravity waves. The inertia part is associated with frequencies $\omega$ proportional to $f$, so called inertial oscillation, while gravity waves are associated with frequencies $\omega = \sqrt{c_0^2(\alpha^2 + \beta^2)}$. Recalling that the dispersion relation for gravity waves is $\omega = c_0 \sqrt{\alpha^2 + \beta^2}$ which shows that $c_0$ is the phase speed of gravity waves.

Note that to construct the analytic solution () - () for any given initial and boundary conditions we just expand the solution into a two-dimensional Fourier series, that is,

\begin{align*}
x = \sum_{\alpha=-\infty}^{\infty} \sum_{\beta=-\infty}^{\infty} x_0(\alpha, \beta) e^{i(\alpha x + \beta y - \omega t)}, \quad (9.37)
\end{align*}

where we observe that $x_0(\alpha, \beta)$ is the amplitude of each Fourier component.

**Finite difference equation**

To solve (9.18) - (9.20) by numerical means we employ the CTCS (leapfrog) scheme. Hence

\begin{align*}
\frac{u_{jk}^{n+1} - u_{jk}^{n-1}}{2\Delta t} + \bar{u} \frac{u_{j+1,k}^{n} - u_{j-1,k}^{n}}{2\Delta x} + \bar{v} \frac{u_{jk-1}^{n} - u_{jk+1}^{n}}{2\Delta y} - f v_{jk}^{n} &= -\frac{\phi_{j+1,k}^{n} - \phi_{j-1,k}^{n}}{2\Delta x}, \quad (9.38) \\
\frac{v_{jk}^{n+1} - v_{jk}^{n-1}}{2\Delta t} + \bar{u} \frac{v_{j+1,k}^{n} - v_{j-1,k}^{n}}{2\Delta x} + \bar{v} \frac{v_{jk-1}^{n} - v_{jk+1}^{n}}{2\Delta y} + f u_{jk}^{n} &= -\frac{\phi_{j+1,k}^{n} - \phi_{j-1,k}^{n}}{2\Delta y}, \quad (9.39) \\
\frac{\phi_{jk}^{n+1} - \phi_{jk}^{n-1}}{2\Delta t} + c_0^2 \frac{u_{jk+1}^{n} - u_{jk-1}^{n}}{2\Delta x} + c_0^2 \frac{v_{jk+1}^{n} - v_{jk-1}^{n}}{2\Delta y} &= 0. \quad (9.40)
\end{align*}

To investigate the numerical stability of the two-dimensional shallow water problem we are only interested in that part of the solution that contain the waves. Thus we neglect all other momentum and volume sources, as well as any steady state solution upon which the waves may
ride. Accordingly we let \( \bar{u} = \bar{v} = 0 \) in which case (9.38) - (9.40) reduces to

\[
\frac{u_{jk}^{n+1} - u_{jk}^{n-1}}{2\Delta t} - f v_{jk}^{n} + \frac{\phi_{jk+1k}^{n} - \phi_{jk-1k}^{n}}{2\Delta x} = 0, \tag{9.41}
\]

\[
\frac{v_{jk}^{n+1} - v_{jk}^{n-1}}{2\Delta t} + f u_{jk}^{n} + \frac{\phi_{jkk+1}^{n} - \phi_{jkk-1}^{n}}{2\Delta y} = 0, \tag{9.42}
\]

\[
\frac{\phi_{jk}^{n+1} - \phi_{jk}^{n-1}}{2\Delta t} + c_{0}^{2} \frac{u_{jk+1k}^{n} - u_{jk-1k}^{n}}{2\Delta x} + c_{0}^{2} \frac{v_{jkk+1}^{n} - v_{jkk-1}^{n}}{2\Delta y} = 0. \tag{9.43}
\]

Again we make use of von Neumann’s method and insert the discrete Fourier component. In this case the discrete Fourier component is

\[
x_{jk}^{n} = X_{n} e^{i(\alpha j \Delta x + \beta k \Delta y)}, \tag{9.44}
\]

where the transpose of the vector \( x_{jk}^{n} \) is \( x_{jk}^{nT} = [u^{n}_{jk}, v^{n}_{jk}, \phi_{jk}^{n}] \) and the transpose of the vector \( X_{n} \) is \( X_{n}^{T} = [U_{n}, V_{n}, \Phi_{n}] \). Insertion into (9.41) - (9.43) then gives

\[
U_{n+1} - U_{n-1} = 2f \Delta t V_{n} - 2i \Phi_{n} \frac{\Delta t}{\Delta x} \sin \alpha \Delta x, \tag{9.45}
\]

\[
V_{n+1} - V_{n-1} = -2f \Delta t U_{n} - 2i \Phi_{n} \frac{\Delta t}{\Delta y} \sin \beta \Delta y, \tag{9.46}
\]

\[
\Phi_{n+1} - \Phi_{n-1} = -2ic_{0}^{2} U_{n} \frac{\Delta t}{\Delta x} \sin \alpha \Delta x - 2ic_{0}^{2} V_{n} \frac{\Delta t}{\Delta y} \sin \beta \Delta y. \tag{9.47}
\]

To find an equation for the growth factor we first eliminate \( V_{n} \) and \( U_{n} \). We do this by first replacing \( n \) by \( n + 1 \) in (9.45) and (9.46) followed by a replacement of \( n \) by \( n - 1 \). By subtracting the results we get

\[
U_{n+2} - 2U_{n} + U_{n+2} = 2f \Delta t (V_{n+1} - V_{n-1})
- 2i(\Phi_{n+1} - \Phi_{n-1}) \frac{\Delta t}{\Delta x} \sin \alpha \Delta x, \tag{9.48}
\]

\[
V_{n+2} - 2V_{n} + V_{n+2} = -2f \Delta t (U_{n+1} - U_{n-1})
- 2i(\Phi_{n+1} - \Phi_{n-1}) \frac{\Delta t}{\Delta y} \sin \beta \Delta y. \tag{9.49}
\]

Substituting for \( V_{n+1} - V_{n-1} \) from (9.46) in the first equation and \( U_{n+1} - U_{n-1} \) from (9.45) in the second equation we get

\[
U_{n+2} - 2(1 - 2f^{2} \Delta t^{2}) U_{n} + U_{n+2} = -4f \Phi_{n} \frac{\Delta t^{2}}{\Delta y} \sin \beta \Delta y
- 2i(\Phi_{n+1} - \Phi_{n-1}) \frac{\Delta t}{\Delta x} \sin \alpha \Delta x, \tag{9.50}
\]

\[
V_{n+2} - 2(1 - 2f^{2} \Delta t^{2}) V_{n} + V_{n+2} = 4if \Phi_{n} \frac{\Delta t^{2}}{\Delta x} \sin \alpha \Delta x
- 2i(\Phi_{n+1} - \Phi_{n-1}) \frac{\Delta t}{\Delta y} \sin \beta \Delta y. \tag{9.51}
\]
TWO-DIMENSIONAL PROBLEMS

9.3 Shallow water equations

We are now in a position to eliminate \( U_n \) and \( V_n \) from (9.47). To this end we first replace \( n \) by \( n + 2 \) in (9.47), and then replace \( n \) by \( n - 2 \). Adding the results and subtracting (9.47) multiplied by \( 2(1 - f^2 \Delta t^2) \) we get

\[
\Phi_{n+3} - \Phi_{n+1} - 2(1 - f^2 \Delta t^2)(\Phi_{n+1} - \Phi_{n-1}) + \Phi_{n+1} - \Phi_{n-3} = \\
-2i c_0^2 \left[ U_{n+2} - 2(1 - f^2 \Delta t^2)U_n + U_{n-2} \right] \frac{\Delta t}{\Delta x} \sin \alpha \Delta x \\
-2i c_0^2 \left[ V_{n+2} - 2(1 - f^2 \Delta t^2)V_n + V_{n-2} \right] \frac{\Delta t}{\Delta y} \sin \beta \Delta y.
\] (9.52)

Thus substituting from (9.50) and (9.51) we finally get

\[
\Phi_{n+3} - (1 + 2\lambda)\Phi_{n+1} + (1 + 2\lambda)\Phi_{n-1} - \Phi_{n-3} = 0
\] (9.53)

where

\[
\lambda = 1 - 2f^2 \Delta t^2 - 2 \left( \frac{c_0 \Delta t}{\Delta x} \right)^2 \sin^2 \alpha \Delta x - 2 \left( \frac{c_0 \Delta t}{\Delta y} \right)^2 \sin^2 \beta \Delta y
\] (9.54)

Defining the growth factor by \( \Phi_{n+2} = G \Phi_n \) we get a third order equation for the growth factor, that is,

\[
G^3 - (1 + 2\lambda)G^2 + (1 + 2\lambda)G - 1 = 0
\] (9.55)

One solution is \( G_1 = 1 \), while the two other solutions are solutions to the second order equation

\[
G^2 - 2\lambda G + 1 = 0.
\] (9.56)

Thus

\[
G_{2,3} = \lambda \pm i\sqrt{1 - \lambda^2}.
\] (9.57)

As so many times before we have to require that the radical is real in which case \( |G_{1,2,3}| = 1 \) and the scheme is neutrally stable. The condition for this to be true is hence that

\[
|\lambda| \leq 1
\] (9.58)

or

\[
-1 \leq 1 - 2f^2 \Delta t^2 - 2 \left( \frac{c_0 \Delta t}{\Delta x} \right)^2 \sin^2 \alpha \Delta x - 2 \left( \frac{c_0 \Delta t}{\Delta y} \right)^2 \sin^2 \beta \Delta y \leq 1.
\] (9.59)

The right-hand inequality is no problem. The left-hand inequality requires

\[
\left( \frac{c_0 \Delta t}{\Delta x} \right)^2 \sin^2 \alpha \Delta x + \left( \frac{c_0 \Delta t}{\Delta y} \right)^2 \sin^2 \beta \Delta y \leq 1 - f^2 \Delta t^2
\] (9.60)

and hence that

\[
\Delta t \leq \frac{\Delta x}{c_0} \sqrt{\frac{\Delta y^2}{\Delta x^2 + \Delta y^2} (1 - f^2 \Delta t^2)}.
\] (9.61)
In practice $f^2 \Delta t^2 \ll 1$. Thus

$$\Delta t < \frac{\Delta x}{c_0} \sqrt{\frac{\Delta y^2}{\Delta x^2 + \Delta y^2}}.$$  \hspace{1cm} (9.62)

If we let $\Delta x = \Delta y = \Delta s$ then we get

$$\Delta t < \frac{\Delta s}{c_0 \sqrt{2}}.$$  \hspace{1cm} (9.63)

We are now in a position to compare (9.63) with the similar conditions for the one-dimensional case as displayed in (6.53). We observe, as for the diffusion and advection problem, that including more than one dimension in space results in a more stringent condition because of the appearance of the factor $\sqrt{2}$ in the denominator. This result therefore appears to be general and is.
Chapter 10

ADVANCED TOPICS

The purpose of this chapter is to present material to indicate how to expand on our knowledge acquired through the previous chapters. One such example is how to construct higher order schemes. Another is how to expand to more dimensions in space, and what to do when we combine advection and diffusion, that is, the common advection-diffusion problem. We also use this opportunity to give some more details on the nature of the numerical, nonlinear instability, and to introduce the spectral method.

10.1 Higher order advection schemes

As alluded to in Section 2.6 we may construct schemes with higher order accuracy using Taylor series expansion. As an example let us consider how to construct a fourth order accurate scheme for the advection equation (Haidvogel and Beckmann, 1999)

$$\partial_t \theta + u \partial_x \theta = 0. \tag{10.1}$$

First we recall from Section 2.6 that (2.25) and (2.27) (cf. page 19) in our notation becomes

$$\theta^n_{j\pm 1} = \theta^n_j \pm \partial_x \theta^n_{j\pm \Delta x} \pm \frac{1}{2} \partial^2_x \theta^n_{j\pm \Delta x} \Delta x^2 \pm \frac{1}{6} \partial^3_x \theta^n_{j\pm \Delta x} \Delta x^3 + O(\Delta x^4). \tag{10.2}$$

Thus follows that

$$\partial_x \theta^n_j = \frac{\theta^n_{j+1} - \theta^n_{j-1}}{2\Delta x} - \frac{1}{6} [\partial^3_x \theta] \Delta x^2 + O(\Delta x^4). \tag{10.3}$$

Note that we have on purpose kept terms of $O(\Delta x^2)$ which we previously dropped when constructing a second order approximation to the spatial derivative in (10.1). We also notice that to we constructed (10.3) using the points adjacent to $\theta^n_j$, that is, the points $\pm \Delta x$ away. Suppose we used points located $\pm 2\Delta x$ away instead. Then the Taylor series (10.2) becomes

$$\theta^n_{j\pm 2} = \theta^n_j \pm \partial_x \theta^n_{j\pm 2\Delta x} \pm \frac{1}{2} \partial^2_x \theta^n_{j\pm 2\Delta x} (2\Delta x)^2 \pm \frac{1}{6} \partial^3_x \theta^n_{j\pm 2\Delta x} (2\Delta x)^3 + O(\Delta x^4), \tag{10.4}$$
that is, we simply replaced $\Delta x$ by $2\Delta x$. Hence solving with respect to $\partial_x \theta_j^n$ we get

$$
\partial_x \theta_j^n = \frac{\theta_j^{n+2} - \theta_j^{n-2}}{4\Delta x} - \frac{2}{3} \partial_x^2 \theta_j \Delta x^2 + \mathcal{O}(\Delta x^4).
$$

(10.5)

Use of (10.5) to construct a second order finite difference approximation to $\partial_x \theta_j^n$ by neglecting all terms $\mathcal{O}(\Delta x^2)$ and higher is of course as valid as (10.3). In the limit $\Delta x \to 0$ they both tend to $\partial_x \theta$, that is, they are both numerically consistent. We may therefore combine them linearly to give

$$
a \frac{\theta_j^{n+1} - \theta_j^{n-1}}{2\Delta x} + b \frac{\theta_j^{n+2} - \theta_j^{n-2}}{4\Delta x} = (a + b) \partial_x \theta_j^n + \frac{1}{6} (a + 4b) \partial_x^3 \theta_j^n \Delta x^2 + \mathcal{O}(\Delta x^4),
$$

(10.6)

requiring that $a + b = 1$, where $a$ and $b$ are linear weights yet to be found. Solving (10.6) with respect to $\partial_x \theta_j^n$ and using the requirement that $a + b = 1$ we get

$$
\partial_x \theta_j^n = a \frac{\theta_j^{n+1} - \theta_j^{n-1}}{2\Delta x} + b \frac{\theta_j^{n+2} - \theta_j^{n-2}}{4\Delta x} - \frac{1}{6} (a + 4b) \partial_x^3 \theta_j^n \Delta x^2 + \mathcal{O}(\Delta x^4).
$$

(10.7)

We may now use (10.7) to construct a fourth order finite difference approximation to $\partial_x \theta_j^n$ by choosing the numbers $a$ and $b$ so that in addition to requiring $a + b = 1$ we also require $a + 4b = 0$. This eliminates the second order term on the right-hand side of (10.7) and the leading truncation error becomes $\mathcal{O}(\Delta x^4)$. Thus we have two equations to solve for the two unknown weights $a$ and $b$ which gives $a = \frac{4}{5}$ and $b = -\frac{1}{5}$. Hence a numerically consistent, fourth order in space and second order in time scheme for the advection equation is

$$
\theta_j^{n+1} - \theta_j^n = \frac{4}{3} \frac{\Delta t}{\Delta x} \left\{ \frac{4}{3} \frac{\theta_j^{n+1} - \theta_j^{n-1}}{2\Delta x} - \frac{1}{3} \frac{\theta_j^{n+2} - \theta_j^{n-2}}{4\Delta x} \right\} + u \left[ \frac{\Delta t}{\Delta x} \left( \theta_{j+1}^{n} - \theta_{j-1}^{n} - \frac{1}{8} (\theta_{j+2}^{n} - \theta_{j-2}^{n}) \right) \right].
$$

(10.8)

or

$$
\theta_j^{n+1} = \theta_j^n - \frac{4}{3} \frac{\Delta t}{\Delta x} \left\{ \theta_j^{n+1} - \theta_j^{n-1} - \frac{1}{8} (\theta_{j+2}^{n} - \theta_{j-2}^{n}) \right\}.
$$

(10.9)

Since the forth order scheme (10.9) is based on Taylor series we know that it is consistent. We also suspect it to be conditionally stable, but what about the condition? Will it be more restrictive or more tolerant? To consider the stability of (10.9) we use von Neumann’s method. Thus letting $\theta_j^n = \Theta_n e^{\alpha j \Delta x}$ we first get

$$
\theta_j^{n+1} - \theta_j^{n-1} = 2i \Theta_n e^{\alpha j \Delta x} \sin \alpha \Delta x
$$

(10.10)

and

$$
\theta_j^{n+2} - \theta_j^{n-2} = 2i \Theta_n e^{\alpha j \Delta x} \sin 2\alpha \Delta x = 4i \Theta_n e^{\alpha j \Delta x} \sin \alpha \Delta x \cos \alpha \Delta x.
$$

(10.11)

The equation for the growth factor $G = \Theta_{n+1}/\Theta_n$ then becomes

$$
G^2 + 2i \lambda G - 1 = 0 \quad ; \quad \lambda = \frac{\Delta t}{3 \Delta x} \sin \alpha \Delta x (4 - \cos \alpha \Delta x),
$$

(10.12)
and hence that

\[ G_{1,2} = -i\lambda \pm \sqrt{1 - \lambda^2}. \]  

(10.13)

As expected the CTCS schemes for the fourth order scheme return a growth factor whose absolute value does equal one even though its expression is slightly more complex than for the second order scheme. Thus the fourth order scheme is neutrally stable under the condition that the radical in (10.13) is a positive definite quantity, that is,

\[ \frac{1}{3} C |\sin \alpha \Delta x|(4 - \cos \alpha \Delta x) \leq 1 \]

(10.14)

where \( C = |u|\Delta t/\Delta x \) is the Courant number. Since the maximum value of \((4 - \cos \alpha \Delta x)\) is five and the maximum value of \(|\sin \alpha \Delta x|\) is one we finally get

\[ C \leq \frac{3}{5} \text{ or } \Delta t \leq \frac{3\Delta x}{5|u|}, \]

(10.15)

which indeed more stringent than the condition \( C \leq 1, \text{ or } \Delta t \leq \Delta x/|u| \), that we obtained for the second order, leapfrog scheme.

Finally we may ask whether the numerical dispersion property is affected. To this end we follow the procedure given in Section 5.13 on page 79. Thus we start by decomposing \( \theta^n_j \) into its Fourier components in time and space,

\[ \theta^n_j = \Theta_\theta e^{i(j\Delta x - cn\Delta t)}. \]

(10.16)

We then substitute (10.16) into (10.8) and solve with respect to \( c \), which gives

\[ c = \frac{1}{\alpha \Delta t} \arcsin \left\{ u\alpha \Delta t \left[ \frac{4}{3} \left( \frac{\sin \alpha \Delta x}{\alpha \Delta x} \right) - \frac{1}{3} \left( \frac{\sin 2\alpha \Delta x}{2\alpha \Delta x} \right) \right] \right\}. \]

(10.17)

To leading order in \( \alpha \Delta x \) we then obtain\(^1\)

\[ c \approx u \left\{ 1 - \frac{4}{5!} (\alpha \Delta x)^4 + \cdots \right\}. \]

(10.18)

Recalling (see Section 5.13 on page 79) that the second order in space leapfrog scheme phase speed was

\[ c = \frac{1}{\alpha \Delta t} \arcsin \left( \frac{\sin \alpha \Delta x}{\alpha \Delta x} \right) \approx u \left\{ 1 - \frac{1}{3!} (\alpha \Delta x)^2 + \cdots \right\}, \]

(10.19)

we observe that (remember that \( 0 \leq \alpha \Delta x \leq \pi \)) the fourth order scheme is nearly always superior to the second order scheme. This process of constructing higher order finite difference approximations may be continued. For example we note that the scheme

\[ \frac{\theta^n_{j+1} - \theta^n_{j-1}}{2\Delta t} + u \left\{ \frac{3}{2} \frac{\theta^n_{j+1} - \theta^n_{j-1}}{2\Delta x} - \frac{3}{5} \frac{\theta^n_{j+2} - \theta^n_{j-2}}{4\Delta x} + \frac{1}{10} \frac{\theta^n_{j+3} - \theta^n_{j-3}}{4\Delta x} \right\} = 0. \]

(10.20)

\(^1\)Note that \( \sin z = 1 - z^2/6 + \cdots \) while \( \arcsin z = 1 + z^2/6 + \cdots \) for \(|z| < 1\).
10.2 Combined advection-diffusion

In Chapter 4 we learned that the diffusion equation was unstable when applying a centered in time, centered in space scheme (leapfrog scheme), while we in Chapter 5 learned that a forward in time, centered in space scheme (Euler scheme) was unstable for the advection equation. As underscored in Chapter 3 most of the problems encountered regarding evolution of tracers in the atmosphere and oceans contain both advection and diffusion in one and the same equation. The question is therefore what scheme should we employ when solving equations which is a combination of the two processes, that is, when solving the so-called advection-diffusion equation?

We investigate this by seeking finite difference approximations to the continuous combined advection-diffusion equation (3.1). We start by using the parameterization given by (3.3) for the advective flux and (3.4) for the diffusive flux (cf. page 31). Moreover, we simplify the problem by assuming that the problem is one-dimensional in space and that the velocity is constant, that is, \( v = u_0 i \) where \( u_0 \) is a constant. Thus the advection-diffusion equation becomes

\[
\frac{\partial \theta}{\partial t} + u_0 \frac{\partial \theta}{\partial x} = \kappa \frac{\partial^2 \theta}{\partial x^2}.
\] (10.23)

To obtain a stable scheme we must ensure that the diffusive part is forward in time and the advective part is centered in time. We may for instance make use of the approximation

\[
\theta_j^{n+1} = \theta_j^{n-1} - \frac{u_0 \Delta t}{\Delta x} (\theta_j^{n+1} - \theta_j^{n-1}) + \frac{2\kappa \Delta t}{\Delta x^2} (\theta_j^{n+1} - 2\theta_j^{n-1} + \theta_{j-1}^{n-1}).
\] (10.24)
We take note that the diffusive part is taken at time level \(n-1\) and thus becomes forward in time with a time step of \(2\Delta t\). In contrast the advective part is evaluated at time step \(n\) and is thus centered in time with a time step of \(\Delta t\). Hence each part is stable in itself. If \(\kappa = 0\) then the advective part is stable if the Courant number \(C \equiv |u_0|\frac{\Delta t}{\Delta x} \leq 1\). If \(u_0 = 0\) the diffusive part is stable under the condition \(\kappa \Delta t \Delta x \leq \frac{1}{4}\). The factor \(\frac{1}{4}\) arises because of the \(2\Delta t\) time step used for the diffusive part, and replaces the \(\frac{1}{2}\) factor in (4.34) on page 44.

In the general case with \(u_0 \neq 0\) and \(\kappa \neq 0\) we get a modified condition. We again use von Neumann’s method and substitute the single Fourier component into (10.24) to get an equation for the growth factor. The algebra is left to the reader (cf. Exercise 1 at the end of this Chapter) and gives

\[
G^2 + 2i\lambda G - \lambda_2 = 0
\]  
(10.25)

where

\[
\lambda = \frac{u_0 \Delta t}{\Delta x} \sin \alpha \Delta x \quad \text{and} \quad \lambda_2 = 1 - 4\frac{\kappa \Delta t}{\Delta x^2} (1 - \cos \alpha \Delta x)
\]  
(10.26)

are two real numbers. We note that since \(1 - \cos \alpha \Delta x \geq 0\) it follows that \(\lambda_2 \leq 1\). In accord with (10.25) the growth factor has two solutions given by

\[
G_{1,2} = -i\lambda \pm \sqrt{\lambda_2 - \lambda^2}.
\]  
(10.27)

To ensure that the two complex solutions have a real part we require that the radical is positive, that is, \(\lambda^2 \leq \lambda_2\). As a corollary we note that this also implies that \(\lambda_2 \geq 0\). The two roots are then complex conjugates and hence

\[
|G| = \sqrt{GG^*} = \sqrt{\lambda^2 + \lambda_2 - \lambda^2} = \sqrt{\lambda_2}.
\]  
(10.28)

The solution is thus conditionally stable because \(0 \leq \lambda_2 \leq 1\). Moreover the condition \(\lambda^2 \leq \lambda_2\) gives

\[
C^2 \sin^2 \alpha \Delta x \leq 1 - 4K (1 - \cos \alpha \Delta x)
\]  
(10.29)

where \(C = |u_0|\frac{\Delta t}{\Delta x}\) is the Courant number and \(K = \kappa \Delta t / \Delta x^2\). The condition (10.29) may be rewritten to give

\[
(C^2 + 4K) \sin^2 \alpha \Delta x = 1 + 4K \cos \alpha \Delta x (1 - \cos \alpha \Delta x)
\]  
(10.30)

We therefore conclude that the sufficient condition for stability of the combined advection-diffusion scheme (10.24) is

\[
C^2 + 4K \leq 1, \quad \text{or} \quad \frac{(u_0 \Delta t)^2 + 4K \Delta t}{\Delta x^2} \leq 1.
\]  
(10.31)

We note that that for either \(u_0 = 0\) or \(\kappa = 0\), the stability condition for the individual advective and diffusive schemes are recovered. We also note that imposing each condition is not a sufficient condition. We therefore obtain the somewhat surprising results that adding explicit diffusion in the advection equation actually reduces the maximum time step allowed for advection. What (10.31) says is that by adding diffusion we arrive at a more restrictive condition. This is
inside of the rectangle is where both individual conditions are met
region of numerical stability is inside parabola and is smaller

Figure 10.1: The diagram illustrates the region of stability for the combined advection-diffusion equation approximated in (10.24). This corresponds to the area inside of the parabola (hatched area). The area inside the rectangular is where both the advection and the diffusion are stable individually. We notice that we obtain a more stringent stability condition to the advection equation when we are adding diffusion.

visualized in Fig. (10.1). For most cases in oceanography and meteorology this is not a serious problem since commonly

$$K \ll C^2.$$  \hspace{1cm} (10.32)

We mentioned earlier (cf. Section 4.6) that it is common to add a diffusion term to avoid nonlinear problems to become numerically unstable by so called nonlinear instabilities as discussed in the next section (Section 10.3). The diffusion term is therefore not part of the physics we are solving for, but rather an artificial term added to make the numerical solution stable. Under these circumstances we may use Dufort-Frankel scheme (cf. Section 4.6 on page 48) to approximate the diffusion term, even though it is inconsistent. This is fine as long as the remaining terms in our governing equations are treated by consistent schemes. Thus we proceed by making the following FDA of (10.23),

$$\theta^{n+1}_j = \theta^{n-1}_j - \frac{u_0 \Delta t}{\Delta x} (\theta^{n}_j - \theta^{n}_j) + 2K(\theta^{n+1}_j - \theta^{n+1}_j - \theta^{n-1}_j + \theta^{n-1}_j), \hspace{1cm} (10.33)$$

in which we have combined a consistent conditionally stable scheme for advection with an unconditionally stable, inconsistent scheme for diffusion. The growth factor then follows the equation

$$(1 + 2K)G^2 - 2\lambda G - (1 - 2K) = 0 \hspace{1cm} (10.34)$$

where

$$\lambda = 2K \cos \alpha \Delta x - i \frac{u_0 \Delta t}{\Delta x} \sin \alpha \Delta x \hspace{1cm} (10.35)$$
and thus the growth factor has two solutions given by

\[ G_{1,2} = \frac{1}{1 + 2K} \left( \lambda \pm \sqrt{4K^2 + \lambda^2 - 1} \right) \]  

(10.36)

It can be shown that for the one-dimensional case it is sufficient to satisfy the CFL condition \( C \leq 1 \). In the more general case for instance for a two-dimensional case a more stringent condition has to be applied (Cushman-Roisin, 1984).

Many authors (e.g. Clancy, 1981) suggest to use the unstable forward in time, centered in space (FTCS) scheme when combining advection and diffusion. The approximation to (10.23) then becomes

\[ \theta_{n+1}^j = \theta_n^j - \frac{u_0 \Delta t}{2 \Delta x} (\theta_{j+1}^n - \theta_{j-1}^n) + \frac{\kappa \Delta t}{\Delta x^2} (\theta_{j+1}^n - 2 \theta_j^n + \theta_{j-1}^n). \]  

(10.37)

The amplification or growth factor then follows the equation

\[ G = 1 - \frac{u_0 \Delta t}{2 \Delta x} \sin \alpha \Delta x - 2 \frac{\kappa \Delta t}{\Delta x^2} \left( 1 - \cos \alpha \Delta x \right). \]  

(10.38)

As shown by Clancy (1981) the scheme is stable provided the two conditions

\[ \frac{\kappa \Delta t}{\Delta x^2} \leq \frac{1}{2}, \quad \text{and} \quad \frac{|u_0| \Delta t}{\kappa} \leq 1 \]  

(10.39)

are both satisfied at the same time. Despite the enthusiasm of several authors we do not recommend the use of the FTCS scheme. Rather we advocate to use the more conservative schemes (10.24) and (10.33).

### 10.3 Non-linear instability

Towards the end of Section 3.2 we mentioned that every non-linear solution of a problem of hyperbolic nature in which friction is neglected will eventually become numerically unstable. This is independent of the time step chosen and is associated with the energy cascade towards smaller and smaller scales that is the nature of non-linear problems. Hence it is not sufficient, for instance in the case of solving the non-linear advection equation (5.1), to satisfy the linear CFL criterion.

To satisfy ourselves that this is indeed true it is enough to solve a simple non-linear advection problem like (5.1). Sooner or later disturbances of wavelengths in the range \( 2\Delta x \) to \( 4\Delta x \) crops up. These disturbances are at first small in amplitude but growing. At some stage into the calculation the solution falls short of satisfying the linear CFL condition and the solution blows up, that is, becomes linearly, numerically unstable. The solution is then useless. It is common to credit Phillips (1959) to be the first to demonstrate this phenomenon by analytic means. Richtmyer (1963) provided another example which is reproduced below. Robert et al. (1970) generalized the previous example.

Before entering into details we notice:
1. All good functions may be expanded in terms of a discrete set of waves or exponentials

2. In a linear system waves of different wavelengths exist independent from each other

3. In a non-linear system the latter is no longer true and waves of different wave numbers will interact and sometime generate waves of new periods

4. Given a finite grid of size $\Delta x$ we have a band-limited wavenumber space, that is, only a finite number of discrete waves can exist

The first point is well known. It simply tells us that all good functions $\Psi(x)$ of period $2L$ may be formulated into a Fourier series, that is, for $x \in [-L,L]$ the function $\Psi(x)$ is written as

$$\Psi(x) = a_0 + \sum_{m=1}^{\infty} a_m \sin(\alpha_m x) + b_m \cos(\alpha_m x)$$  \hspace{1cm} (10.40)

where $\alpha_m = m\pi/L$ is the discrete wavenumber, $a_m$ and $b_m$ is the amplitude or energy associated with the wavenumber $\alpha_m$ and $a_0$ is the mean or average value of $\Psi$ for $x \in [-L,L]$.

The second point tells us that if the system is linear there is no exchange of energy between them. Thus two wave trains of different amplitude, wavelength and direction will pass each other without changing neither of them.

The third point emphasizes the fact that it is the non-linearity that causes exchange to happen. To illustrate this suppose we have a solution, say wind or current $u(x,t)$, given by

$$u(x,t) = \sum_n u_n(t) \sin(\alpha_n x)$$ \hspace{1cm} (10.41)

Then nonlinear products will give rise to terms having wavenumbers which are the sum of and difference of the two original wavenumbers, e.g.,

$$\sin(\alpha_1 x) \sin(\alpha_2 x) = \frac{1}{2} [\cos(\alpha_1 - \alpha_2)x - \cos(\alpha_1 + \alpha_2)x]$$ \hspace{1cm} (10.42)

Thus in a non-linear case the two wave trains will be different after the passage, that is, they will experience a change in either wavelength, as illustrated by (10.42), amplitude or direction.

The fourth and last point tells us that when we formulate the function $\Psi(x)$ as a sum of discrete waves on a grid of size $\Delta x$ we have a band-limited wavenumber space in which the shortest wave that can possibly be resolved is $2\Delta x$. Thus our wavenumber space is limited to wavenumbers $\pi/L \leq \alpha_m \leq \pi/\Delta x$. For a non-linear problem in which the various waves interact to produce waves of wavenumber $\alpha > \pi/\Delta x$, that is waves of wavelengths shorter than $2\Delta x$, they are unresolved by our grid. Unfortunately these unresolved waves are folded into some low wavenumber. In fact as displayed in Figure 10.2 a wave of wavelength $\frac{4}{3}\Delta x$ is indistinguishable from a wave of wavelength $4\Delta x$. Let us arbitrarily call $\alpha < \pi/2\Delta x$ low wavenumbers and $\pi/\Delta x < \alpha < \pi/2\Delta x$ high wavenumbers. The latter are then waves of wavelengths between $2\Delta x$ and $4\Delta x$ and corresponds to the shortest waves that are resolved by our grid of size $\Delta x$. 

150
10.3 Non-linear instability

Figure 10.2: Displayed are the two waves of wavelength $4\Delta x$ (solid curve) and $\frac{3}{4}\Delta x$ (dashed curve), in a grid of grid size $\Delta x$. Note that our grid cannot distinguish between the unresolved wave of wavelength $\frac{3}{4}\Delta x$ and the resolved wave of wavelength $4\Delta x$. Thus the energy contained in the unresolved wave will be folded into the low wavenumber space represented by the $4\Delta x$ wave.

Points three and four above make us expect a priori that even though all initial energy is low wavenumber (long waves), non-linear interactions will eventually provide variance (or energy) at high wavenumbers (short waves). This is easily verified by investigating the model problem we use below, which is a simple non-linear advection equation in one dimension, that is,

$$\partial_t u + u \partial_x u = 0. \quad (10.43)$$

The difference between (10.43) and the earlier advection equation, e.g., (5.1) on page 59, is the appearance of the nonlinear term $u \partial_x u$. Suppose we have a solution at a particular time level $n$ that is a monochromatic wave of wavelength $2\pi/\alpha$ where $\alpha$ is the wavenumber, that is, $u^n(x) = u_0 \sin \alpha x$. Then from (10.43), using a scheme that is centered in time, we get

$$u^{n+1} - u^{n-1} = -u_0^2 \sin \alpha x \partial_x \sin \alpha x = -u_0^2 \sin \alpha x \cos \alpha x = -\frac{1}{2} u_0^2 \sin 2\alpha x. \quad (10.44)$$

Hence the solution at the next time level is a wave of wavelength $2\pi/2\alpha$, that is, a wavelength half of that of the original wavelength at time level $n$. Thus, we observe, as expected, that all the energy originally contained at low wavenumbers (long waves) end up at high wavenumbers unresolved by our grid. Due to the folding of the energy contained in the unresolved waves, the energy contained in the shortest wave resolved by our grid, that is, waves of wavelength $4\Delta x$, accumulates. Thus after a sufficient time period the numerical model blows up due to ordinary numerical, linear instability, even though the linear problem is numerically stable.

\footnote{Note that (10.43) is the acceleration term in the momentum equation for a one-dimensional problem. Hence non-linearity is ubiquitous in all realistic atmospheric and oceanographic models.}

151
To inspect the non-linear instability in some more detail we use the example of Richtmyer (1963). We start with the assumption the problem has variance at wavelengths $\infty$ (zero wavenumber), $4\Delta x$ and $2\Delta x$, and that the model problem is the simple non-linear advection equation in one dimension (10.43). Let us as Richtmyer (1963) approximate (10.43) using the classic leapfrog (CTCS) scheme, that is

$$u_{j}^{n+1} = u_{j}^{n-1} - \frac{\lambda}{2} [(u_{j+1}^{n})^{2} - (u_{j-1}^{n})^{2}]$$

(10.45)

where $\lambda = \Delta t / \Delta x$. We will not concern ourselves with boundary conditions, but as in the von Neumann analysis of Section 4.3 on page 42 perform a local analysis. Then making use of (10.40) with $\Psi = u$ and $a_0 = V$ we can show that an exact, formal solution to (10.45) is

$$u_{j}^{n} = C_{n} \cos \left( \frac{\pi j}{2} \right) + S_{n} \sin \left( \frac{\pi j}{2} \right) + U_{n} \cos (\pi j) + V.$$ 

(10.46)

We can identify the amplitudes $C_{n}, S_{n}$ as the amplitudes of a wave with length $4\Delta x$, $U_{n}$ as the amplitude of a wave of length $2\Delta x$, and $V$ as a wave of low wavenumber ($\alpha < \pi/2\Delta x$) with zero wavenumber (infinite wavelength). When we substitute (10.46) into (10.45) and notice that $(u_{j+1}^{n})^{2} - (u_{j-1}^{n})^{2} = (u_{j+1}^{n} - u_{j-1}^{n})(u_{j+1}^{n} + u_{j-1}^{n})$ we obtain relationships among the amplitudes,

$$C_{n+1} - C_{n-1} = 2\lambda S_{n}(U_{n} - V)$$

$$S_{n+1} - S_{n-1} = 2\lambda C_{n}(U_{n} + V)$$

$$U_{n+1} = U_{n-1}.$$ 

(10.47)

The last equation in (10.47) says that $U_{n}$ may take on different values for the odd and even time steps, say $A$ for odd time steps ($n = 1, 3, 5, \ldots$) and $B$ for the even time steps ($n = 2, 4, 6, \ldots$), that is, $U_{2m} = A$ and $U_{2m+1} = B$ for $m = 1, 2, 3, \ldots$. By eliminating $S_{n}$ from the first equation in (10.47), we obtain

$$C_{n+2} - 2C_{n} + C_{n-2} = 4\lambda^{2}(A + V)(B - V)C_{n}.$$ 

(10.48)

The question then arise. Is this solution stable in the von Neumann sense? As in the simple linear case, using the von Neumann method, we define a growth factor associated with the $4\Delta x$ wave. For the $4\Delta x$ wave to be stable the growth factor has to be less than or equal to one. Thus we first define the growth factor by letting $G \equiv C_{n+2}/C_{n}$. Substituting this into (10.48) we derive

$$G^{2} - 2\gamma G + 1 = 0,$$ 

(10.49)

where $\gamma$ is a real number given by

$$\gamma = 1 + 2\lambda^{2}(A + V)(B - V).$$ 

(10.50)

The roots of (10.49) are

$$G_{1,2} = \gamma \pm i \sqrt{1 - \gamma^{2}}.$$ 

(10.51)
We notice that as long as the radical is real then

$$|G_{1,2}| = \sqrt{\gamma^2 + 1 - \gamma^2} = 1$$  (10.52)

The $4\Delta x$ wave is therefore neutrally stable provided

$$1 - \gamma^2 \geq 0 \quad \text{or} \quad -1 \leq \gamma \leq 1.$$  (10.53)

As is obvious it is only possible to satisfy (10.50) if the amplitude of the $2\Delta x$ wave is such that $|A| < V$ and/or $|B| < V$. This is violated when the amplitude of the $2\Delta x$ wave is large in comparison with the energy contained in the longer waves (low wavenumbers). In this case the $4\Delta x$ will grow exponentially and the scheme is unstable.

### 10.4 Smoothing and filtering

In numerical models of the atmosphere or ocean, we have learned that it is important to damp out the smallest space scales to control for instance non-linear instability. We also notice that this may be done by adding explicit eddy viscosity of momentum diffusion as described in Section 3.2 and Section 4.6. Here we focus on another method, namely employing filtering techniques to control spurious growth of short waves due to numerical errors and computational instabilities that would otherwise obscure a good forecast. In fact even if a catastrophic instability does not occur we still may want to remove the noise in the shortest wavenumber band for aesthetic reasons. Sometimes we apply such smoothing to the final product only.

The simplest form of smoothing is to apply a so called one-dimensional three-point operator or filter (Shapiro, 1970, 1975). Such a filter is defined by

$$\tilde{u}_j^n = (1 - \mu)u_j^n + \frac{1}{2}\mu(u_{j+1}^n + u_{j-1}^n) ,$$  (10.54)

where $\mu$ is a constant. If the solution is a monochromatic wave, say $u_j = U_ne^{i\alpha j\Delta x}$, then the filtered solution is

$$\tilde{u}_j^n = Ru_j^n$$  (10.55)

where

$$R = 1 - \mu(1 - \cos\alpha\Delta x) = 1 - 2\mu\sin^2\left(\frac{\alpha\Delta x}{2}\right)$$  (10.56)

is the response function associated with the filter. Thus the filter does not affect the wave length nor the phase speed (provided $R \geq 0$). Furthermore if $R<1$ then the wave is damped. Moreover, for the particular wave number $\alpha = 2\pi/2\Delta x$, that is, for the shortest wave resolved in our grid, we get

$$R = 1 - 2\mu.$$  (10.57)

For the particular choice $\mu = 1/2$ we then get $R = 0$, and hence the waves of wavelength $2\Delta x$, the two gridlength waves, are completely removed by the filter.
We furthermore observe that the filter may be rewritten to yield

$$\bar{u}_j^n = u_j^n + \frac{1}{2}\mu (u_{j+1}^n - 2u_j^n + u_{j-1}^n).$$  \hfill (10.58)

Making use of Taylor series we recognize the last term on the right-hand side of (10.58) as the finite difference approximation of the second order derivative in space with a truncation error of $O(\Delta x^2)$. Thus we get

$$\bar{u}_j^n = u_j^n + \frac{1}{2}\mu \Delta x^2 [\partial_x^2 u_j^n],$$  \hfill (10.59)

which shows that the filter acts similar to diffusion with a mixing coefficient given by $\kappa = \frac{1}{2}\mu \Delta x^2$.

For more in depth details the reader is referred to \textit{(Haltiner and Williams, 1980, Chapter 11-8, page 392 and onward)}.

\section*{10.5 The spectral method}

When we applied the various approximation to the advection equation above, we only consider grid-point values of the dependent variables. We did not make any assumption about how the variables behaved between grid points other than assuming that they are good functions.

An alternative approach is to expand the dependent variables in terms of finite series of orthogonal functions (cf. Section 2.10 on page 24). The problem is then reduced to solving a set of ordinary differential equations which determine the behavior in time of the expansion coefficients. Following this approach is known as the \textit{spectral method}.

The spectral method is particularly suitable for global atmospheric models where the dependent variables are zonally cyclic function and hence easy to expand. The method is therefore commonly applied in modern global atmospheric models, for instance in the global model used at the European Centre for Medium range Weather Forecast (ECMWF). We note that the method is a bit more cumbersome to apply in non-global models, and also in global oceanographic models since the latter has to deal with the continental land boundaries.

\subsection*{Application to the one-dimensional linear advection equation}

We will demonstrate the spectral method for the one-dimensional advection equation on the globe, i.e., along a latitude. Under these circumstances the natural boundary condition is the periodic or cyclic boundary condition (cf. Section 2.5 on page 16).

We recall from Section 5.1 on page 61 that the one-dimensional advection equation is

$$\partial_t \phi = -u_0 \partial_x \phi, \quad \text{for} \quad x \in [0, L] \quad \text{and} \quad t > 0$$  \hfill (10.60)

where $L$ is the length of the circumference at a particular latitude. We recall from Section 2.10 that (10.60) is just a special case of the general equation (2.59) on page 25. Hence the linear
operator of Section 2.10 is $\mathcal{H} = -u_0 \partial_x$. Since we will solve (10.60) along a latitude we first conveniently transform to a coordinate system in which

$$2\pi x = \xi L \quad \text{or} \quad \xi = \frac{2\pi x}{L} \quad (10.61)$$

where $\xi \in (-\infty, 2\pi)$ is the new (dimensionless) zonal coordinate. Since

$$\partial_x \phi = \partial_\xi \phi \partial_\xi \xi \quad (10.62)$$

(10.60) then transforms to

$$\partial_t \phi = -\gamma \partial_\xi \phi \quad (10.63)$$

where

$$\gamma = \frac{2\pi u_0}{L} \quad (10.64)$$

is the angular velocity (in units one per second). The cyclic boundary condition is written

$$\phi(\xi, t) = \phi(\xi + 2\pi m) \quad m = 1, 2, 3, \ldots \quad (10.65)$$

where $m$ describes the number of times you have traveled around the world at that latitude. We further let the initial condition be described by the good function $f(\xi)$, and hence

$$\phi(\xi, 0) = f(\xi). \quad (10.66)$$

As outlined at the beginning of Section 5.1 on page 59 the true solution to (10.63) is then

$$\phi = f(\xi - \gamma t). \quad (10.67)$$

Solving (10.63) using expansions in terms of orthogonal functions requires us to choose a suitable set of expansion functions. The obvious choice in our case following (10.40) on page 150 above is to choose complex exponentials (sine and cosine functions), since these are eigenfunctions of the differential operator $\mathcal{H} = -u_0 \partial_x$. For a continuous function we get

$$\phi = \sum_\alpha \phi_\alpha(t) e^{i\alpha \xi}, \quad (10.68)$$

where $\alpha$ is the wave number and the summation is for all possible wavenumbers from $-\infty$ to $+\infty$. Solving (10.63) using numerical methods implies that we are band-limited in wavenumber space and hence we must use a truncated version of (10.68), that is,

$$\phi(\xi, t) = \sum_{l=-l_{\text{max}}}^{l=l_{\text{max}}} \phi_l(t) e^{i\alpha_l \xi}. \quad (10.69)$$

where $l = l_{\text{max}}$ gives the maximum wave number $\alpha_{l_{\text{max}}}$ resolved on the grid, that is the shortest wavelength resolved by our choice of grid size (here $2\Delta \xi$). Since $\phi_{-l} = \phi_l^*$, we need only be concerned with $0 \leq l \leq l_{\text{max}}$, rather than the full set of expansion coefficients.
We now substitute (10.69) into (10.63) and equate coefficients of the expansion functions. Thus
\[
\partial_t \phi_l = -i \alpha_l \gamma \phi_l \quad \text{for} \quad l = 0(1)l_{\text{max}}
\]
giving \(2l_{\text{max}} + 1\) equations for the expansion coefficients \(\phi_l\)'s. For this particular case (10.70) can be integrated exactly for each wavenumber \(\alpha_l\) separately to give
\[
\phi_l(t, \alpha_l) = \phi_l(0, \alpha_l)e^{-i \alpha_l \gamma t}
\]
where \(\phi_l(0, \alpha_l)\) is the initial condition associated with the wavenumber \(\alpha_l\). If we expand the good function \(f(\xi)\) in terms of a truncated Fourier series, that is,
\[
f(\xi) = \sum_l a_l e^{i \alpha_l \xi} = \sum_l \phi_l(0, \alpha_l)e^{i \alpha_l \xi},
\]
we get that \(\phi_l(0, \alpha_l) = a_l\), and hence that the complete solution to (10.63) is
\[
\phi(\xi, t) = \sum_{l=0}^{l_{\text{max}}} a_l e^{i \alpha_l (\xi - \gamma t)},
\]
which is the same as the true solution. Hence there is no dispersion due to the space discretization, unlike in the finite difference approximation above. This fact is due to the space derivatives being computed analytically while they were approximated in the finite difference method. We recall due to the orthogonality property of the expansion functions, in our case \(e^{i \alpha_l \xi}\), that
\[
\phi_l = \int_{0}^{2\pi} \sum_m \phi_m e^{i \alpha_m \xi} e^{-i \alpha_l \xi} d\xi.
\]
Thus by multiplying (10.69) by the complex conjugate of the expansion functions and integrating in space we get
\[
\phi_l(t, \alpha_l) = A_l \int_{0}^{2\pi} \phi(\xi, t)e^{-i \alpha_l \xi} d\xi,
\]
where \(A_l\) are the normalization factors. Note that (10.75) is the so called direct Fourier transform. The normalization coefficients are determined from the initial condition, or by use of (10.75), that
\[
A_l = \frac{a_l}{\int_{0}^{2\pi} \phi(\xi, 0)e^{-i \alpha_l \xi} d\xi}.
\]
In practice we have at our disposal the grid points values of \(\xi\) rather than a continuous function in space. Thus we know \(\xi\) at \(J + 1\) points \(\Delta \xi\) apart such that \(\xi_j = j \Delta \xi\) where \(j = 0, 1, 2, \cdots, J - 1, J\) and where \(\xi_J = 2\pi\). In this case we think of the truncated Fourier series of \(\phi\) as given in (10.69) as representing an interpolating function which exactly fits the values of \(\phi\) at the \(J + 1\) grid points. We then write (10.75) as a discrete direct Fourier transform,
\[
\phi_l(t, \alpha_l) = A_l^* \sum_{j=1}^{J} \phi(\xi_j, t)e^{-i \alpha_l \xi_j},
\]
where the normalization coefficients are found by discretization (10.76),

\[ A'_l = \frac{a_l}{\sum_{j=1}^{J} \phi(\xi_j, 0) e^{-i\alpha_t \xi_j}}. \]  (10.78)

The corresponding discrete inverse Fourier transform is then

\[ \phi(\xi_j, t) = \sum_{l=-l_{\text{max}}}^{l_{\text{max}}} \phi_l(t, \alpha_l) e^{i\alpha_l \xi_j}. \]  (10.79)

Both (10.77) and (10.79) can be computed with the Fast Fourier Transform (FFT) algorithm. It can be shown that starting from the set \( \phi_l(t, \alpha_l) \) going to the set \( \phi(\xi_j, t) \) with \( j = 0, 1, 2, \ldots, J - 1, J \) and returning to the set \( \phi_l(t, \alpha_l) \) we recover exactly the original values provided the number of grid points \( J \) are such that \( J > 2l_{\text{max}} + 1 \). Recall that \( l_{\text{max}} \) is the number of waves used to compute the direct Fourier transform in (10.79). In addition we must require that the points \( \xi_j \) are equally spaced or that \( \Delta \xi \) is a constant.

It remains to find the expansion coefficients \( \phi_l(t, \alpha_l) \) at an arbitrary time given their initial values \( \phi_{0l} \). We do this by a time stepping procedure, for instance applying a centered in time scheme to (10.70),

\[ \phi_{l}^{n+1} = \phi_{l}^{n-1} + 2i\alpha_l \gamma \Delta t \phi_{l}^{n} \quad l = 1, 2, 3, \ldots, LM \]  (10.80)

for each wavenumber \( \alpha_l \). Using von Neumann’s method we show that numerical stability is ensured provided

\[ |\alpha_l \gamma \Delta t| \leq 1; \quad \forall l. \]  (10.81)

Since the maximum wavenumber is \( \alpha_{LM} \) we require \( |\alpha_{LM} \gamma \Delta t| \leq 1 \). Moreover, since the maximum dimensionless wavenumber\(^3\) is \( \alpha_{LM} = L/2\Delta x \) it follows that the stability condition in terms of the Courant number \( C = u_0 \Delta t / \Delta x \) is

\[ C \leq \frac{1}{\pi}. \]  (10.82)

which is actually more stringent than the one derived for the finite difference approximation. Although being more restrictive the spectral method and scheme has the great advantage that it is nearly non-dispersive, and that the dispersiveness is very small even for the shortest waves of two grid lengths.

**Exercises**

1. Use von Neumann’s method (Chapter 4.3) to show that the expression (10.25) is indeed the correct expression for the growth factor when using the scheme given in (10.24).

2. Show that the condition (10.31) is a sufficient condition for numerical stability.

\(^3\)The maximum dimensional wavenumber is \( 2\pi / 2\Delta x \).
Bibliography


Mesinger, F., and A. Arakawa (1976), Numerical methods used in atmospheric models, GARP Publication Series No. 17, 64 p., World Meteorological Organization, Geneva, Switzerland.


