Chapter 12

100 Years of Earth System Model Development

DAVID A. RANDALL,a CECILIA M. BITZ,b GOKHAN DANABASOGLU,c A. SCOTT DENNING,a PETER R. GENT;c ANDREW GETTELMAN,c STEPHEN M. GRIFFIES,d PETER LYNN,e HUGH MORRISON,c ROBERT PINCUS,f AND JOHN THUBURNg

a Department of Atmospheric Science, Colorado State University, Fort Collins, Colorado
b Department of Atmospheric Sciences, University of Washington, Seattle, Washington
c National Center for Atmospheric Research, Boulder, Colorado
d Geophysical Fluid Dynamics Laboratory, Princeton, New Jersey
e University College Dublin, Dublin, Ireland
f CIRES, University of Colorado Boulder, Boulder, Colorado
g University of Exeter, Exeter, United Kingdom

ABSTRACT

Today’s global Earth system models began as simple regional models of tropospheric weather systems. Over the past century, the physical realism of the models has steadily increased, while the scope of the models has broadened to include the global troposphere and stratosphere, the ocean, the vegetated land surface, and terrestrial ice sheets. This chapter gives an approximately chronological account of the many and profound conceptual and technological advances that made today’s models possible. For brevity, we omit any discussion of the roles of chemistry and biogeochemistry, and terrestrial ice sheets.

1. Introduction

The development of models for numerical simulation of the atmosphere and oceans was one of the great scientific triumphs of the twentieth century. The models have added enormously to our understanding of the diverse and complex processes at work in the Earth system, and to our ability to produce realistic simulations of both near-future weather and the longer-term future of the climate system. Understanding and simulation are the two broad goals of model development.

Today’s global atmospheric models are commonly coupled with ocean models, sea ice models, and land surface models that include representations of terrestrial vegetation and the carbon cycle. Because of the diversity of processes represented, it is becoming more common to refer to these large coupled models as “Earth system models (ESMs),” especially when the carbon cycle is included. In ESMs, the atmosphere, ocean, sea ice, and land surface models are included as submodels, which can be viewed as components of the larger coupled model. Some ESMs also include components representing atmospheric and marine chemistry, terrestrial ice sheets, ocean biology, and biogeochemistry, but we will not discuss those topics in this chapter. The atmosphere and ocean submodels of ESMs are often referred to as global circulation models (GCMs).

Each component of an ESM includes exchanges of mass, momentum, and energy with one or more of the other components. The atmosphere model is the only component of an ESM that carries out exchanges with all of the other components. The air, water, and ice are in constant motion. In the atmospheric component of an ESM, the adiabatic terms of the equation of motion, the thermodynamic equation, and the continuity equations for dry air, moisture, and chemical species are solved on a three-dimensional grid1 using what is called a “dynamical core.”2 The horizontal and vertical grid spacings determine the spatial “resolution”

1 Spectral models are discussed later.
2 “Dycore” for the enthusiasts.
of the model. This chapter includes an overview of the evolution of dynamical cores for global models of the atmosphere and ocean.

Atmospheric models also include parametric representations, called “parameterizations,” that are designed to incorporate the transports by radiation, precipitation, and the unresolved or “subgrid scale” motions of the air, as well as the phase changes of water, averaged up to the grid scale. This chapter includes a selective overview of the evolution of the parameterizations used in global atmospheric models. All of the parameterized processes are formulated in terms of the fields that are resolved by the model’s dynamical core. A fundamental issue in parameterization development is that the atmosphere and ocean contain eddies on all scales. Early studies aimed to choose the grid spacing so that it coincided with meteorologically inactive scales (e.g., Fiedler and Panofsky 1970), but it soon became clear that there is no such “spectral gap”; eddies exist on all scales (e.g., Nastrom et al. 1984), although of course some are more consequential than others.

The dynamical cores of ocean models are designed to cope with the complex geometry of the ocean basins. Numerical modeling of the ocean began somewhat later than numerical modeling of the atmosphere, but has today reached a comparable level of intellectual maturity. This chapter discusses the history of the hydrostatic primitive equation ocean models used as components of ESMs. Ocean models include parameterizations of the fluxes associated with unresolved motions of the water. We focus on dynamical and numerical aspects, and do not discuss regional and coastal ocean applications, biogeochemistry, or process modeling. Further discussion of ocean physics and dynamics is given in the chapter by Carl Wunsch and colleague, in this volume (Wunsch and Ferrari 2019).

Even sea ice and terrestrial ice sheet models can be said to have dynamical cores, in the sense that they include dynamical equations that govern the motions of the ice. Prior to 1950, there were no publications about sea ice models in English—possibly none at all—and few scientists had ever seen sea ice. Nevertheless, long before weather and climate models simulated the mass or momentum balance of sea ice, scientists recognized the importance for the climate system of the high albedo of sea ice. Early climate modelers used energy balance models with an ice–albedo feedback parameterized by raising the surface albedo when the surface temperature dropped below a critical value (Budyko 1969; Sellers 1969). When subjected to climate forcing, such as a reduction in the solar radiation, the energy balance models respond with cooling that is strongest in the high latitudes—a phenomenon now widely known as polar amplification.

Land surface models have no dynamical cores; in that sense, they are “all parameterization.” We humans live on the land surface, so it is hardly surprising that our science has spent a lot of effort to understand and predict conditions there. From the point of view of the atmosphere, the land is merely a lower boundary condition, but it is also where we grow most of our food, build our cities, and live our lives. Quantitative modeling of land surface processes goes back well over 100 years, primarily with applications to agriculture and water resources. The land surface is an important mediator in the flows of energy, water, carbon, and momentum. The albedo of the land surface is highly variable. Ordinarily, most net radiative energy absorbed by the surface is transferred to the atmosphere as turbulent fluxes of sensible and latent heat, with only a small residual driving changes of heat storage in the soil. These turbulent energy fluxes are important drivers of atmospheric energetics and circulation. Water from precipitation can infiltrate the surface or run off, and infiltrated water is stored and can be released later as vapor. The land surface is a strong sink of atmospheric momentum, and the friction arising from the land surface is felt throughout the atmospheric boundary layer and sometimes far beyond. The topography of the land surface has an enormous impact on the circulation of the atmosphere. Critically, much of the land surface is alive. It is inhabited by vegetation and by microbes in the soil, whose biological processes mediate the partition between turbulent fluxes of sensible and latent heat and regulate the ability of the atmosphere to extract water from beneath the ground. Vegetation is an important determinant of the surface albedo and surface friction. The responses of plants and soil microbes to changes in atmospheric conditions can dramatically affect the surface fluxes.

The purpose of this chapter is to give an account of the century or so of development work that led to today’s ESMs, starting from the early years of the twentieth century. Model development involves scientific analysis of how nature works, so that the model can work in the same way as far as possible. Some engineering is also involved, especially to achieve optimal performance on the available computers.

In writing this chapter, we have assumed that the readers have some familiarity with numerical modeling of the Earth system, but of course we have tried to avoid unnecessary technical details. Our chapter contains no equations. Applications of the models are briefly mentioned, but not emphasized. The story of the development of ESMs is huge and complicated, and our version of it is unavoidably incomplete. Space limitations make it impossible for us to mention all of the important contributions. We acknowledge our debt to earlier accounts, including those of Smagorinsky (1983), Wiin-Nielsen (1991), Arakawa (2000), Lynch (2006), Washington (2007), Edwards (2010, 2011),

The chapter is organized chronologically, to the extent possible. The first section deals with the gestational period from about 1900 to 1950. Then, starting with the 1950s, the sections are organized by decade, but with some exceptions to maintain narrative continuity. We tell the story of each decade using several subsections, some of which are focused on particular ESM components. We have attempted to interweave our accounts of the developments of numerical methods, radiative transfer, turbulence and cloud parameterizations, ocean and sea ice modeling, and land surface modeling, because of course that is the way it really happened. The all-important and rapidly evolving “supercomputers” needed to run the models are also mentioned in several places.

Our chapter inevitably infringes on the subject of numerical weather prediction, which is a major focus of a separate chapter in this volume by Stanley Benjamin and colleagues (Benjamin et al. 2019).

2. Before the beginning

a. Early work on dynamical cores

Concepts fundamental to Earth system modeling were developed in the early years of the twentieth century. Three visionary scientists played particularly central roles (Fig. 12-1). The great American meteorologist Cleveland Abbe recognized that meteorology is essentially the application of hydrodynamics and thermodynamics to the atmosphere (Abbe 1901), and he identified the system of mathematical equations that govern the evolution of the atmosphere (Willis and Hooke 2006). The Norwegian scientist Vilhelm Bjerknes undertook a more explicit analysis of the weather prediction problem from a scientific perspective (Bjerknes 1904). His stated goal was to make meteorology an exact science, a true physics of the atmosphere. He argued that it should be possible to predict changes in the weather by solving systems of partial differential equations, which is exactly what we do today.

The English Quaker mathematician, Lewis Fry Richardson, went further. He wanted a worked example for his book “Weather Prediction by Numerical Processes” (Richardson 1922). Partly to create such an example, he attempted what is now called numerical weather prediction (NWP): a direct (but approximate) solution of the equations of motion. The result was his famous “failed” numerical forecast (actually a hindcast) for 20 May 1910. He carried out the calculations by hand, in the intervals between driving for the Friends Ambulance Unit during the war in France (Ashford 1985; Lynch 2006). Although his results were not realistic, his achievement was heroic, and his book was remarkably prescient. His overall approach bears a striking resemblance to that used in modern weather and climate models, and he appreciated many of the issues that still preoccupy modelers today. In particular, he understood that the large-scale dynamics of the atmosphere would be resolved, while other processes, such as radiation, boundary layer turbulence, and cloud processes, would have to be parameterized. He used what we now call the quasi-static approximation. To obtain approximate solutions of the differential equations of the model, he proposed a method based on finite differences, a technique that he had devised and previously applied to stresses in a masonry dam (Richardson 1911). He discretized his domain on a longitude–latitude grid or “lattice” that covered part of western Europe, with five layers to represent the atmosphere’s vertical structure. He understood that a staggered arrangement of variables on the grid could improve the accuracy of finite differences, and he used
what we would now recognize as a pair of C grids (Lynch 2006; Arakawa and Lamb 1977). He also foresaw that his proposed grid would present difficulties in the polar regions. His model included equations for predicting soil moisture based on empirical work by hydrologists in the mid-nineteenth century. He knew that the weather is influenced by terrestrial vegetation, which had already been appreciated by Von Humboldt et al. (1859), and he understood the role of plant physiology in regulating the extraction of water from the soil (transpiration). Finally, he provided suggestions for initializing and integrating his model.

Richardson’s hindcast gave a totally unrealistic rate of change of the surface pressure: 145 hPa over a 6-h period. The full story of Richardson’s work, the reason for his disappointing numerical results, and a complete reconstruction of the forecast are described by Lynch (2006). When, in a later retrospective recreation, Richardson’s initial data were dynamically balanced, the initial tendency of surface pressure was reduced to a reasonable value of less than 1 hPa over 6 h, confirming that his unrealistic prediction was due to the dynamical imbalance of the initial data that he used. Details are presented in Lynch (2006).

Richardson’s forecasting scheme was quite impractical in the precomputer era, but he was undaunted, speculating that “some day in the dim future it will be possible to advance the computations faster than the weather advances.” In fact, developments on several fronts were necessary before NWP could be put into practice. First, a more complete understanding of atmospheric dynamics allowed the development of simplified but sufficiently general systems of equations. Advances in what used to be called “physical meteorology” pointed the way to useful statistical representations of the effects of unresolved physical processes on the resolved scales. Regular observations of the free atmosphere provided the initial conditions needed for numerical weather prediction; accurate and stable discretization schemes were developed. Finally, increasingly powerful digital computers provided a practical means of carrying out the prodigious calculations needed to forecast changes in the weather.

At the time of the First World War, computational weather forecasting was impractical for at least four reasons. First, the observations of the three-dimensional structure of the atmosphere were available only on a very occasional basis, with inadequate coverage, and never in real time. The registering balloons had to be recovered and the recordings analyzed to obtain the data, a process that took days or even weeks. Second, the numerical algorithms for solving the atmospheric equations were subject to instabilities that were not understood. Because of this, the numerical solution might bear little or no resemblance to the solution of the continuous equations. Third, the nearly balanced (e.g., nearly geostrophic) nature of atmospheric flow was not yet understood, and the imbalances arising from observational and analysis errors confounded Richardson’s forecast. Fourth, the massive volume of computation required to advance the numerical solution could not be carried out, even by a huge team of human computers. In reality, Richardson’s estimate that 64000 human computers would be needed to do the calculations for a useful forecast in real time, was a gross underestimate. It has been reckoned that closer to one million people would have been required for the task (Lynch 1993). It seems fair to say that what Richardson devised was a “method without a means.”

In the ensuing decades, a variety of key developments prepared the way for progress. Theoretical developments provided crucial understanding of atmospheric dynamics, in particular the approximate balance of the large-scale atmospheric state and the means of eliminating spurious high-frequency gravity waves. This led to the quasigeostrophic equations, which filter gravity waves and describe the large-scale motions of atmosphere away from the equator. Advances in numerical analysis led to the design of stable algorithms that faithfully replicated the true solution provided that certain restrictions on the size of the time step were respected. Timely observations of the three-dimensional atmosphere became available following the invention of the radiosonde in 1927. This provided real-time measurements of pressure, temperature, humidity and winds through a vertical column of the atmosphere. Finally, the development of digital computers provided a way of attacking the enormous computational task of weather forecasting.

The Electronic Numerical Integrator and Computer (ENIAC), an electronic computer commissioned by the U.S. Army for calculating the paths of projectiles, was completed in 1945. It was the first programmable electronic digital computer ever built. The gigantic machine used 18 000 thermionic tubes, filled a large room, and consumed 140 kW of power (Fig. 12-2). Both input and output were by means of punched cards. McCartney (1999) provides an absorbing account of the origins, design, development, and legacy of ENIAC.

In the late 1940s, the mathematician John von Neumann recognized that weather forecasting, a problem of both great economic and military importance, and strong intrinsic scientific interest, is an ideal application for a digital computer. He established a Meteorology Project at the Institute for Advanced Study in Princeton, and recruited meteorologist Jule Charney to lead it. The
project created a model in which the atmosphere was treated as a single layer, represented by conditions at the 500-hPa level. ENIAC was used to time step the barotropic vorticity equation, which expresses the conservation of absolute vorticity following the flow, and filters out gravity wave solutions. Centered-in-space finite differences were used to evaluate the vorticity transport, and leapfrog time differencing was used. A Poisson equation was solved to obtain the geopotential height from the predicted vorticity. Fortunately, Charney and his colleagues were aware of the work of Courant et al. (1928, 1967), which showed that in order for their explicit time-stepping method to be stable, the size of the time step cannot exceed the grid size divided by the signal speed, a constraint that we now call the Courant–Friedrichs–Lewey (CFL) criterion. With the barotropic vorticity equation, the relevant signal speed is the wind speed; in a system that permits gravity waves, the signal speed would be the much faster speed of wave propagation, and as a result the time step would have to be much smaller to satisfy the CFL criterion for stability. The initial data for the forecasts were prepared manually from standard operational 500 hPa analysis charts produced by the U.S. Weather Bureau. The heights were held constant on the outer boundaries of the domain, throughout each 24-h integration.

The resulting numerical predictions, carried out on ENIAC, were truly groundbreaking. Four 24-h forecasts were performed, and the results clearly showed that the large-scale features of the midtropospheric flow could be predicted numerically with a reasonable resemblance to reality. The forecasts were described in a pioneering paper by Jule Charney, Ragnar Fjørtoft, and John von Neumann (Charney et al. 1950). The success of the ENIAC forecasts had an electrifying effect on the meteorological community, worldwide. Several baroclinic (multilevel) models were developed in the following years. All of them were based on the filtered or quasi-geostrophic system of equations. Later, models using the more accurate primitive equations were introduced. Charney had anticipated this as a necessary step, and indeed André Robert later identified it as a key development in numerical weather prediction (see Lin et al. 1997).

Charney et al. (1950) noted that the computation time for a 24-h forecast was about 24 h. In other words, the team could just keep pace with the weather, provided that the ENIAC did not fail. The computation time included offline operations, such as the reading, punching, and interfiling of punch cards. Lynch and Lynch (2008) recreated the ENIAC integrations using a programmable cell phone, which they called the Portable Hand-Operated Numerical Integrator and Computer (PHONIAC). In this recreation, PHONIAC executed the main loop of the 24-h forecast in less than one second.

b. Early work on radiative transfer

Thanks to astronomers, methods that can be used for calculating radiative heating rates and fluxes in Earth’s atmosphere have been available since the first half of the twentieth century. Astronomers developed the two-stream methods used to compute the fluxes of radiation (Schuster 1905; Eddington 1916). The idea of collecting together parts of the spectrum with similar amounts of absorption, which forms the basis of the k-distribution technique now used in ESMs, was originally proposed by Ambartsumian (1936). The theory describing the scattering of light by round particles like cloud drops is usually attributed to Mie (1908).

Radiative transfer is fundamentally important for ESMs because radiation is (almost) the only mechanism by which Earth can exchange energy with the rest of the universe, and because motions of the atmosphere are fundamentally driven by spatial gradients in the electromagnetic radiation emitted by Earth, its atmosphere, and the sun. The same gradients also play a key role in determining the thermal structure of the atmosphere. The deep convective clouds of the tropics arise from a rough balance between destabilization by radiative cooling and the response of deep convection, for example, while the planetary-scale Hadley circulation is driven by the gradient in absorbed sunlight between the

---

3 More generally, a necessary condition for stability is that the domain of dependence of the numerical solution at any point should contain the corresponding physical domain of dependence.
equator and higher latitudes. Models of atmospheric motion therefore need to represent the flow of radiation through the atmosphere, especially the radiative flux divergences within the atmosphere that give rise to heating and cooling, and the fluxes of radiation that are absorbed (and emitted) by the surface. Models that are aimed at understanding climate (as opposed to weather) must accurately compute the net energy input at the top of the atmosphere.

The practical calculations needed to advance an atmospheric model are daunting, even today. The underlying reason is that the solution to the radiative transfer equation is nonlinear in the parameters used in the equation (optical depth $\tau$, single-scattering albedo $\omega_0$, and some measure of the scattering phase function, often the asymmetry parameter $g$). These parameters are quite variable in Earth’s atmosphere. For clear skies the primary problem is that for gases, the extinction, the differential value of $\tau$, varies by many orders of magnitude in very small spectral regions around each of the thousands to millions of absorption lines associated with each gas. Clouds present a different class of problem. Compared to the optical depths of gases, the optical depth of clouds varies far more smoothly with wavelength and by only three or four orders of magnitude overall, but much more rapidly in time and space.

The history of radiative transfer parameterizations for ESMs is about maximizing the utility of available computational power by focusing our scientific thinking on specific, motivating problems. One theme that emerges is that computational challenges have, over the last century, sparked useful insights and novel methods. A second is that the collective efforts to understand parameterization errors by comparison to reference line-by-line models have been instrumental in identifying the sources and magnitudes of those errors and pointing to possible solutions.

c. Where things stood in 1950

As the 1940s came to an end, new data sources were being used to carry out pioneering observational studies of the global circulation of the atmosphere, notably by Victor Starr’s group at the Massachusetts Institute of Technology (MIT; Starr 1948; Starr and White 1951), Eric Palmen and colleagues in Finland and at the University of Chicago (Palmén 1948; Palmén and Riehl 1957), and Jacob Bjerknes (the son of Vilhelm Bjerknes, who was mentioned earlier), and Yale Mintz at the University of California, Los Angeles (UCLA; Mintz and Bjerknes 1951; Bjerknes 1955). These observations proved to be both a motivation for and a basis for evaluation of the global atmospheric models that were soon to follow.

3. The 1950s

The 1950s saw some major advances in our understanding of the global circulation. For example, Edward Lorenz (1955) of MIT published the first of his most influential papers, which defined and analyzed available potential energy, and provided important insights into the atmospheric energy cycle. At the University of Chicago, David Fultz carried out rotating annulus experiments that reproduced some of the observed characteristics of the global circulation of the atmosphere (Fultz et al. 1959). Both of these studies (and many others) influenced the development of atmospheric numerical models during the 1950s.

a. Progress with dynamical cores

The landmark NWP success of Charney et al. (1950) was soon emulated in several other places around the world (e.g., Persson 2005b). As the 1950s unfolded, operational numerical weather prediction began in Sweden (1954; Bolin 1955), the United States (1955), and Japan (1959; Lynch 2006; Persson 2005a,b), though none of those early models were global or even hemispheric.

During this period, experiments began with three-dimensional models that could supplant the barotropic vorticity equation. At first, these continued to use filtered systems of equations that have no gravity wave solutions, but more accurate systems were needed. Early baroclinic models were developed by Charney and Phillips (1953), and experimental forecasts with the primitive equations were carried out by Hinkelmann (1951). Later, Charney (1962) experimented with both the primitive and balance equations. The forecasts produced using three-dimensional filtered models were not much better than those produced using the barotropic vorticity equation, and this motivated more work on hydrostatic primitive equation models (e.g., Shuman and Hovermale 1968; Bushby and Timpson 1967). Because the primitive equations support rapidly propagating gravity waves, a shorter time step is needed to ensure computational stability. In compensation, primitive equation models do not need the expensive elliptic solvers of the quasigeostrophic and balanced models.

Early model builders had to make some very basic choices that are still under discussion today. An example is the choice of how the different variables in the model should be arranged in the vertical. Charney and Phillips (1953) offset the thermodynamic variable, potential temperature $\theta$, relative to the horizontal wind components $u$ and $v$, because this arrangement is natural to capture hydrostatic and thermal wind balance. Lorenz (1960), on the other hand, placed $\theta$ at the same levels as
CHAPTER 12 RANDALL ET AL. 12.7

\[ \begin{align*}
\text{Charney–Phillips} & \quad \theta \\
& \quad u, v \\
& \quad u, v, \theta \\
\text{Lorenz} & \quad u, v, \theta
\end{align*} \]

**FIG. 12-3.** Schematic showing the vertical placement of the horizontal velocity components \( u \) and \( v \) and potential temperature \( \theta \) on the Charney–Phillips and Lorenz grids.

\( u \) and \( v \) (Fig. 12-3), because that arrangement is advantageous for conservation of total energy.

Subsequent applications of the Charney–Phillips and Lorenz vertical grids with more complete equation sets showed that the Charney–Phillips grid better captures wave motions that depend on buoyancy (e.g., Thuburn and Woollings 2005, and references therein). It also showed that the Lorenz grid possesses a computational mode—a pattern of perturbations in the model variables that is invisible to the numerical method and consequently behaves unphysically, for example, by failing to propagate (Tokioka 1978; Arakawa and Moorthi 1988). However, a satisfactory scheme for achieving energy conservation with a Charney–Phillips grid has proved elusive. For many years, Lorenz’s choice was almost universally adopted, but the relative merits of the Charney–Phillips and Lorenz grids were revisited several decades later (Arakawa and Moorthi 1988). Some recently developed models use the Charney–Phillips grid (e.g., Girard et al. 2014; Wood et al. 2014), while others use the Lorenz grid (e.g., Untch and Hortal 2004; Satoh et al. 2008; Skamarock et al. 2012; Zängl et al. 2015). The debate continues.

The dynamical simulation of climate using numerical models can be said to have started in 1956, when Norman Phillips carried out the first extended-range simulation of the global circulation of the atmosphere (Phillips 1956; Lewis 1998). The model predicted the winds at two vertical levels with, naturally, the Charney–Phillips vertical grid, which means that there was only one prognostic temperature, in the middle troposphere. The model was quasigeostrophic, on a beta-plane channel, with just \( 16 \times 17 \) grid columns. It was driven by a specified meridionally varying distribution of heating and cooling. Because the temperature was predicted at only one level, the static stability had to be prescribed; a smaller-than-observed value was used to mimic the effects of moist convection. Starting from a zonal flow with small random perturbations, a disturbance with a wavelength of 6000 km developed. It had the characteristic westward tilt with height of a developing baroclinic wave. Phillips examined the energy transformations associated with the developing wave, and found good qualitative agreement with observations of baroclinic systems in the atmosphere.

His simulation broke down after a few simulated weeks because of a previously unknown form of numerical instability (Phillips 1959). It was not of the sort of instability that results from violation of the CFL criterion; instead, it turned out to be an inherently nonlinear instability in which the spatial scale of nonlinear terms is misrepresented (aliased) by the finite-resolution grid, leading to feedback and the growth of small-scale noise (Phillips 1959). This type of instability can occur, in principle, even in a time-continuous model. Arakawa (1966) reasoned that if the Jacobian term could be computed in such a way as to conserve either energy or enstrophy then there would be “no room for nonlinear computational instability.” Moreover, conservation of both energy and enstrophy would prevent an unrealistic downscale cascade of energy. This motivated Arakawa to develop his energy- and enstrophy-conserving finite-difference Jacobian. The value of numerical methods that conserve physically important quantities emerged as a major theme in later work (e.g., Thuburn 2008).

von Neumann was tremendously impressed by Phillips’s work. To explore its implications, he arranged a conference at Princeton University in October 1955 on “Application of Numerical Integration Techniques to the Problem of the General Circulation.” The workshop had a galvanizing effect on the meteorological community. Within 10 years, there were several major research groups modeling the global circulation of the atmosphere. The first sign of these impending developments was Smagorinsky’s two-level model, formulated using a zonal channel on the sphere (Smagorinsky 1958).

In a further important advance, Norman Phillips proposed the use of the terrain-following \( \sigma \) coordinate (Phillips 1957a), which greatly simplifies the lower boundary conditions of atmospheric models. Variations of the \( \sigma \) coordinate are still very widely used today. Phillips’s invention of the \( \sigma \) coordinate marks the beginning of a multidecadal search for the optimal vertical coordinate systems for use in both atmosphere and ocean models. We return to that story later in this chapter.

**b. Early work on parameterizations of the boundary layer, the land surface, clouds, and cumulus convection**

The exchanges of momentum, sensible heat, and moisture between the atmosphere and the lower boundary are fundamental to understanding the Earth
system. In a key development of the 1950s, the Russian scientists Monin and Obukhov formulated a similarity theory for the “surface layer,” which is the lower portion of the atmospheric boundary layer (Monin and Obukhov 1954; Foken 2006). They showed how the surface fluxes of sensible heat and momentum are related to the near-surface profiles of temperature and wind. Later their ideas were extended to include the surface moisture flux over the oceans and other water surfaces. Two decades later the similarity functions described by Monin and Obukhov were measured in famous field experiments carried out in Kansas (Businger et al. 1971; Haugen et al. 1971) and Minnesota (Izumi and Caughey 1976). Today, Monin–Obukhov similarity theory is used to determine the surfaces fluxes of sensible heat, momentum, and moisture in virtually all atmospheric models. Further discussion is given in the chapter in this volume by Margaret LeMone and colleagues (LeMone et al. 2019).

The 1950s produced major advances in understanding the atmospheric boundary layer and cumulus clouds. Joanne Starr Malkus Simpson and colleagues carried out pioneering observations of turbulence and cumulus convection over the tropical and subtropical oceans, and developed simple and insightful theories of cumulus updrafts and downdrafts (Bunker et al. 1949; Malkus 1952; Starr Malkus 1954, 1955; Simpson et al. 1965; Simpson and Wiggert 1969). Their ideas played crucial roles in the subsequent development of parameterizations of the boundary layer and cumulus convection. It is an interesting fact that the concept of cumulus entrainment, which plays an important role in those parameterizations, was first discussed by oceanographer Henry Stommel (1951).

Riehl and Malkus (1958) used the (relatively meager) observations of their time to analyze the flows of energy through what we now call the intertropical convergence zone (ITCZ). They drew the fundamentally important conclusion that thunderstorms strongly transport energy upward through the depth of the tropical troposphere, and that at some levels the upward energy flux is against the gradient. Their study motivated the representation of cumulus updrafts as penetrative “hot towers” that act like express elevators, carrying energy and other quantities upward through the troposphere in an hour or less. As we will see, these ideas were widely used in cumulus parameterizations during the 1960s and later.

Cloud microphysics deals with cloud and precipitation particles, including their formation and the processes governing their evolution such as condensation, evaporation, melting and freezing. Since these processes act at the microscale (smaller than a micron to centimeters), they cannot be resolved and must be parameterized in all weather and climate models, now and for the foreseeable future. The parameterizations must describe the net effects of interactions between subgrid-scale microphysical processes and the gridscale temperature, water vapor, and winds. The parameterization of microphysics plays an essential role in quantitative precipitation forecasting, coupling with the model dynamics through latent heating and the condensate weight, radiative transfer, and coupling with aerosols and chemistry. While the roots of cloud microphysics extend back several centuries, quantitative understanding was not established until fairly recently. A rapid acceleration of microphysics research began abruptly around 1940, coinciding with growing military interest in cloud processes, the development of new observational techniques including radar, and a hope that it might be possible to modify precipitation production through artificial means (Pruppacher and Klett 1997). Cloud microphysics, and moist physics more generally, had a limited role in the early development of weather and climate models, because extreme simplicity was required. We will return to the subject of cloud physics later in this chapter. For a more thorough discussion of the history of cloud physics research, see the chapter in this volume by Sonia Kreidenweis and colleagues (Kreidenweis et al. 2019).

Modern land surface models also draw on important ideas from the 1950s. Soil temperature as a function of depth can be modeled as thermal diffusion of heat in the vertical, given estimates of heat capacity and thermal diffusivity (Lettau 1954). The vertical heat flux through the soil column is determined by the temperature difference between the air and the soil surface. Penman (1948) derived a simple parameterization for the rate of evaporation from a wet surface based on vapor pressure, wind speed, and net radiation. The chapter in this volume by Christa Peters-Lidard and colleagues summarizes 100 years of progress in hydrology, which is an important aspect of land surface modeling (Peters-Lidard et al. 2019).

c. Approaching 1960

As the 1950s drew to a close, the International Geophysical Year raised the profile of the Earth sciences (Sullivan 1961). Major technological innovations were also occurring. Digital computers were becoming more powerful, easier to program, and much more widely available. Beginning with Sputnik in 1957, artificial satellites were launched into orbit, soon to be followed by quantitative satellite-based observations of Earth. In the following decades, both of these new technologies had major impacts on the development and applications of ESMs.
4. Model development in the Age of Aquarius

The culturally, scientifically, and technologically tumultuous 1960s produced multiple landmark advances in the development of ESMs, including the creation of several now-legendary “ancestral” models, which were aimed mainly at climate simulation rather than weather prediction. In many cases, the earliest versions of the ancestral models were not truly global, and used simplified geography. They incorporated simple parameterizations of surface fluxes, radiation, cumulus convection, and stratiform or “large-scale” clouds, and they were coupled to very simple land surface models. With one important exception they used prescribed sea surface temperatures (SSTs), rather than coupling with an ocean model.

a. The GFDL model

Joseph Smagorinsky was the first director of the Geophysical Fluid Dynamics Laboratory (GFDL) of the National Oceanic and Atmospheric Administration. His vision was to recruit a team of scientists focused on the multidecadal task of using numerical models as an aid to understanding the global circulation of the atmosphere (Lewis 2008). GFDL’s atmosphere model was developed by Smagorinsky, Syukuro Manabe, and collaborators (Smagorinsky et al. 1965; Manabe and Smagorinsky 1967). Early versions covered only the Northern Hemisphere, with a stereographic map projection, and used idealized geography. The GFDL model used the $\sigma$ coordinate of Phillips (1957a). Some versions used “reduced grids” with fewer grid points around latitude circles near the poles (Kurihara 1965). By 1965, the GFDL model had relatively high vertical resolution for the time, with nine glorious layers.

During the 1960s, the GFDL modeling team achieved many important firsts, including a very influential parameterization for the horizontal diffusion of momentum (Smagorinsky 1963), the first radiation parameterization (Manabe and Möller 1961; Manabe and Strickler 1964), the first cumulus parameterization (Smagorinsky 1963; Smagorinsky et al. 1965; Manabe et al. 1965), and the first land surface model (Budyko and Zubenok 1961; Manabe 1969a). Figure 12-4 schematically summarizes the formulation of the early GFDL model (Manabe 1969b).

GFDL atmosphere model produced unrealistic results in humid regions with steep lapse rates. Manabe et al. (1965, p. 770) wrote that because of convective instability, intense grid-scale convection develops exponentially in the area where the lapse rate is unstable. . . . Therefore, it is desirable to design a scheme of convection such that the grid-scale convection does not develop. . . . We used a very simple scheme of convective adjustment depending upon both relative humidity and the lapse rate and successfully avoided the abnormal growth of grid-scale convection.

Moist convective adjustment was designed to remove convective instability by adjusting the lapse rate back to “moist neutral,” and limiting the relative humidity to 100% or less, while minimizing complexity. Moist convective adjustment couples neighboring layers of a model, pairwise. It does not try to represent the penetrative nature of deep convection, which had been emphasized by Riehl and Malkus (1958). Moist convective adjustment is still being used in some of today’s models.

Early results from the GFDL atmosphere model were published by Smagorinsky (1963) and Manabe et al. (1965). The primary application of the GFDL model was climate simulation, but Miyakoda et al. (1969) also used versions of the model in experimental NWP.

GFDL scientists also developed a simple but explicit representation of the surface energy balance over land. The bulk aerodynamic formula of Penman (1948) had been extended by Monteith (1965), who combined the constraints of surface energy balance with the conservation of water through turbulent transport. The combined Penman–Monteith equation includes the effect of surface or stomatal resistance to evapotranspiration. Stomata are microscopic pores on the undersides of the
leaves of plants through which water evaporates. The stomata provide the physiological mechanism for control of evapotranspiration. The Penman–Monteith equation has long been used by farmers and engineers to estimate the evapotranspiration, but the estimation of stomatal conductance remains entirely empirical. Budyko and Zubenok (1961) suggested that physiological control ratio of the ratio of actual evapotranspiration to the potential evapotranspiration could be usefully approximated by a linear ramp between 0 and 1 as soil moisture varied from the wilting point to field capacity. The linear ramp was adopted by Manabe (1969a), who represented soil hydrology by analogy to a bucket of water. Rainfall is added to the soil bucket, which has an arbitrarily set capacity of 15 cm. Additional rainfall when the bucket is full leads to runoff. Evapotranspiration removes water from the bucket at a rate \( \beta \) times the potential evapotranspiration rate, where \( \beta \) is just the ratio of the current contents of the bucket to its capacity. The linear ramp incorporated through \( \beta \) represents the well-known tendency of vegetation to take up less and less water as root-zone soil moisture is depleted.

The origins of numerical ocean circulation modeling can also be traced to GFDL. Smagorinsky recognized the importance of developing a World Ocean circulation model, and in 1960 he hired Kirk Bryan to lead GFDL’s ocean modeling project. It was a massive undertaking, believed by many to be a fool’s mission with extensive challenges. There were prominent naysayers in the community who felt that such efforts were ill-advised at best. Fortunately, Bryan was able to leverage from GFDL’s work on atmospheric numerical models. Mike Cox was an additional member of the ocean modeling team, whose pioneering scientific programming skills proved critical to the success of the project (Bryan 1991).

Bryan and Cox made assumptions to allow for efficient numerical integration using the computers available in the 1960s. One of these assumptions was that the upper boundary of the ocean is a rigid lid. Such a lid eliminates fast external gravity waves (effectively making their speed infinite), and converts a hyperbolic problem for surface gravity waves into an elliptic boundary value problem for the barotropic (depth integrated) streamfunction of the circulation. This innovation allowed for the use of relatively long time steps, thus enabling the century-long integrations needed for climate studies. An additional key element of the model was a momentum advection scheme based on the approach of Arakawa (1966) to remove nonlinear instabilities that had plagued models at that time (Bryan 1966). Bryan chose the Arakawa B grid (Arakawa and Lamb 1977) for staggering of tracer and velocity variables. This choice rendered a relatively accurate numerical calculation of geostrophically balanced motions using the coarse resolution allowed by computers of the day. Bryan and Cox completed their prototype World Ocean model in the mid-1960s (Bryan and Cox 1967). Their pioneering work has now been followed by nearly 50 years of enhancements and refinements. The further evolution of the Bryan–Cox code is discussed in section 5e.

Bryan’s ocean model was soon coupled to GFDL’s global atmosphere model to create the world’s first global coupled atmosphere–ocean model (Manabe and Bryan 1969), although with idealized geography. The fundamental importance of ocean–atmosphere interactions for climate makes it reasonable to say that the model of Manabe and Bryan (1969) was the first true climate model—a major milestone.

The GFDL ocean model was coupled with a sea ice model (Manabe 1969b; Bryan 1969a), which treated the sea ice as a slab of uniform thickness in each grid cell, with all-or-nothing coverage. The temperature profile of the sea ice was assumed to be linear and the effects of salt trapped in the sea ice were neglected. Sea ice less than 3 m thick was advected at the speed of ocean currents averaged over the upper 100 m, while thicker ice was assumed to be locked in place, so it could not converge indefinitely and build to excess. This method for treating sea ice motion came to be known as “free-drift with stoppage.”

The domain and geography of the GFDL model were gradually made more realistic. First results from a global version of the model, with realistic topography, were published by Holloway and Manabe (1971).

b. Leith’s model

Starting in 1960, the Livermore model (Leith 1965a,b, 1988; Michael 1996) was developed single-handedly by Cecil “Chuck” Leith of the Lawrence Radiation Laboratory. Leith’s model ran on the Livermore Automatic Research Calculator (LARC), which was one of the first computers to use transistors rather than vacuum tubes. At first the model represented only the Northern Hemisphere up to 60°N, but a later version was truly global. It used a spherical grid based on longitude and latitude, with a grid spacing of 5° in each direction, but with fewer grid points around latitude circles near the poles. It had five layers and used pressure as its vertical coordinate—the only numerical model of the atmosphere ever to do so, as far as we know. The surface

---

4 Which sounds like some sort of plumbing problem.
5 The Lawrence Radiation Laboratory was later renamed as the Lawrence Livermore National Laboratory.
pressure was predicted. The effects of mountains were not included. The model predicted water vapor, and included the warming effects of latent heat release, as well as precipitation, but had no parameterization of cumulus convection. It did have a parameterization of radiative transfer, including the diurnal cycle, but neglected the radiative effects of clouds. Leith’s dynamical core needed very strong damping to maintain numerical stability. His model had a relatively short lifetime, because his interests shifted toward two-dimensional turbulence. In 1968, he relocated to the National Center for Atmospheric Research (NCAR), which, as discussed below, had its own global modeling project. After moving to NCAR, Leith continued his studies of large-scale atmospheric turbulence, but he was only peripherally involved in the development of NCAR’s global atmospheric model.

c. The UCLA model

Beginning in 1961, the UCLA model (Arakawa et al. 1968; Langlois and Kwok 1969; Arakawa 1972) was developed by Akio Arakawa and collaborators, including Yale Mintz, at the University of California, Los Angeles. It was the only one of the four ancestral models to be developed at a university. A detailed first-person account of the project is given by Arakawa (2000). The early two-level version of the model, which was finished in 1963, did not predict water vapor, but it was global and had a realistic (but low-resolution) land–sea distribution and topography. Results from this version were published by Mintz (1968).

The UCLA model brought several important innovations. Its dynamical core used what are now called “finite volume” methods for both advection and the horizontal pressure-gradient force. It was designed with an emphasis on conservation of mass, energy (Arakawa 1966; Lilly 1997; Arakawa 1972) and other important quantities. These conservation properties were achieved through what are now called “mimetic” discretization methods (Hyman and Shashkov 1997). The model’s dynamical core was designed to optimally simulate the propagation of inertia-gravity waves, including the shortest waves that could be represented on the grid (Arakawa and Lamb 1977).

The cumulus parameterization of the early UCLA model made use of the entraining-plume ideas advocated by Stommel (1951), Riehl and Malkus (1958), and Simpson and Wiggert (1969). It allowed multiple “types” of cumulus clouds; the number of cloud types was determined by the number of layers used, which was three at the time. The UCLA model was the first to use the “mass flux” approach for parameterizing convection (Arakawa 1969), which has now been almost universally adopted. The closure used in the cumulus parameterization removed convective instability, but allowed a less-than-saturated relative humidity. The model’s radiative parameterization (Katayama 1967, 1972) included the diurnal cycle and the radiative effects of the predicted clouds.

d. The NCAR model

NCAR’s first global atmospheric model was developed by Akira Kasahara, Warren Washington, and David Williamson. The earliest version had two levels (Kasahara and Washington 1967; Washington and Kasahara 1970; Oliger et al. 1970). It had no orography, and water vapor was assumed to be at its saturation value throughout the atmosphere, so that latent heat was released wherever and whenever the air moved upward. The NCAR model was the first (and so far only) quasi-static global model to use constant-height surfaces as its vertical coordinate. Richardson’s equation was solved to determine the vertical velocity. Kasahara and Washington (1969), Kasahara and Washington (1971), and Washington and Williamson (1977) described a later six-level version of the model, which included the effects of mountains and predicted clouds. It was coupled to a simple land surface model. The radiation parameterization was developed by Sasamori (1968).

e. Additional advances during the 1960s

Radiation parameterizations for atmospheric models must account for heating and cooling by gases that vary in concentration within the atmosphere, notably water vapor and ozone. Early models focused on the impacts of individual gases (carbon dioxide, ozone, and especially water vapor) on radiation and heating rates within the atmosphere, exploiting the fact that each gas affects a different spectral region. Some approaches used gas amounts as a function of temperature to compute a broadband emissivity (Elsasser and Curdertson 1960) by fitting to observations and/or laboratory data. Emissivity could be used to compute heating rates from a spectrally integrated equation describing flux (e.g., Sasamori 1968). Others used band models (Curtis and Goody 1954, is one example) in which an assumed distribution of absorption line shapes, strengths, and relative positions determines the average transmission of a model layer as a function of absorber amount, temperature, and pressure within some finite spectral region. The absorption features of each gas were assumed to be spectrally independent so that the total transmission is the product of the transmission due to each gas. Total fluxes and heating rates can then be computed by adding up contributions from each spectral region. Longwave
cooling calculations were typically expressed as matrix problems, following Curtis (1956) and Rodgers and Walshaw (1966). This approach describes the exchange of radiation between pairs of not necessarily contiguous layers, and between each layer and the upper and lower boundary (space and the surface, respectively). Such a calculation scales as the number of layers squared times the number of spectral intervals, but the relatively coarse vertical and spectral resolutions of the time made it practical.

A desire to simulate the global circulation of the atmosphere on a more or less homogeneous grid motivated some early interest in quasi-uniform spherical grids, including overset grids (Phillips 1957b), icosahedral grids (Williamson 1968; Sadourny et al. 1968), and cubed spheres (Sadourny 1972). The first results with these methods were not very encouraging, however, and with the emergence of the spectral transform method around that time (Eliassen et al. 1970; Orszag 1970), interest waned until the 1990s. Both quasi-uniform spherical grids and the spectral method are discussed further later in this chapter.

The U.S. National Meteorological Center developed the first operational NWP model to incorporate precipitation and the effects of latent heating (Shuman and Hovermale 1968). The model predicted the precipitable water, that is, the total amount of water vapor in each atmospheric column, and instantaneously converted vapor to surface precipitation when the precipitable water in a grid cell exceeded 0.8 times the value that would occur if the air was saturated at all levels. An ad hoc approach was used to distribute the corresponding latent heating vertically, and there was no explicit representation of microphysical processes. This parameterization was used operationally starting in March 1967. A similar approach was used in many operational weather forecast models over the next two to three decades.

H. L. Kuo (1965) proposed the first cumulus parameterization to use an entraining plume to represent the cumulus updrafts (see also Kuo 1974). He assumed (incorrectly) that the environment of the cumulus clouds was warmed by outward diffusion of enthalpy from the updrafts rather than by convective fluxes. Kuo determined the intensity of convection based on the tendency of water vapor due to low-level convergence and surface evaporation. This moisture-convergence “closure” was widely used for many years (e.g., Anthes 1977; Tiedtke 1989), but later fell out of favor (Emanuel 1991; Arakawa 2004).

Cumulus convection was not the only important cloud type to receive close attention during the 1960s. Douglas Lilly published an elegant, insightful, and (ultimately) very influential analysis of marine subtropical stratocumulus clouds (Lilly 1968). He emphasized the importance of cloud-top processes, including radiative cooling, entrainment, and the evaporation of cloud water, for the evolution of stratiform cloud systems. Over a period of decades, Lilly’s 1968 paper has exerted a major influence on parameterizations of both clouds and boundary-layer turbulence.

During the 1960s and into the 1970s, global atmospheric models predicted water vapor distributions including the effects of precipitation and moist convection, but most specified a fixed distribution of clouds from observations for interaction with radiation (e.g., Manabe et al. 1965; Washington and Kasahara 1970). The UCLA model was an exception.

In the late 1950s and 1960s efforts were made to theoretically interpret cloud and precipitation observations. This work coincided in particular with the development and use of radar and other observational advances and was largely independent of weather and climate model developments at the time. Pioneering work in this area was conducted by Edwin Kessler. As a doctoral student at MIT and later as a researcher at the Weather Radar Branch at Great Blue Hill, Massachusetts, and director of the Atmospheric Physics Division at the Travelers Research Center in Hartford, Kessler recognized the utility of analyzing data using simplified water mass continuity equations. As he wrote (Kessler 1995, p. 121):

I worked with a strong sense for interactions among processes as discussed here, and in expectation that their study would be facilitated by simple means to portray microphysical processes. The first process to be considered was conversion of cloud to precipitation. How to portray it? I did little more than observe in the literature and with my own eyes that thin water clouds seem to be persistent, and that rain falls from dense clouds.

This behavior was captured by continuity equations for cloud water and rain mass that were developed and initially applied in a kinematic flow model (Kessler 1969). Conversion processes between cloud and rain were represented by “autoconversion” using a threshold cloud mass mixing ratio above which conversion occurred, and “accretion,” which represented the growth of existing raindrops by collection of cloud. Rain was allowed to evaporate and sediment and the precipitation rate was calculated explicitly from the predicted rain field. A diagram of the scheme is shown in Fig. 12-5a. It was a major advance, and it still provides a general framework for almost all bulk microphysics schemes used in weather and climate models up to the present day.
From the late 1950s to the early 1980s, the first attempts were made to simulate turbulence and cumulus clouds using high-resolution numerical models with relatively small domains (Malkus and Witt 1959; Lilly 1962; Ogura 1962; Deardorff 1964, 1972b, 1974, 1980; Moeng 1984). Today we speak of models for large-eddy simulation (LES), and cloud-resolving models. Such models are now extensively used for developing and testing global models, and for determining the numerical values of parameters used in subgrid-scale parameterizations for lower-resolution models.

Lorenz’s revolutionary paper on deterministic non-periodic flow (Lorenz 1963) transformed our understanding of the limits of deterministic weather prediction, and eventually led to ensemble forecasting (Lewis 2005). Motivated by Lorenz’s discovery, Charney (1966) used early versions of the Livermore, UCLA, and GFDL models to investigate the sensitivity of the atmospheric circulation to small perturbations. This work by Charney and colleagues could perhaps be viewed as the first model intercomparison study.

f. Where things stood at the end of the 1960s

It is interesting to list some of the ways in which the global modeling arena of the 1960s differed from today’s. First, all of the global atmosphere and ocean models of the 1960s were developed in the United States, although Japanese immigrants to the United States (Akio Arakawa, Akira Kasahara, and Syukuru Manabe) played key roles in the development of three of the models. All of the lead developers were men. The motivations for developing the models were purely academic, in the sense that the primary focus was improved understanding, rather than immediate practical applications. The funding that supported the modeling work was modest by today’s standards. The modeling teams were small and informally organized, in contrast to today’s much larger and more bureaucratic enterprises. All of the models used “gridpoint” methods with spherical (longitude–latitude) coordinates, and all of them used the quasi-static primitive equations. The atmospheric models simulated only the troposphere, with the exception of an early experiment by Manabe and Hunt (1968). Although, as discussed above, the 1960s did see some early work on models of the ocean, sea ice, and the land surface, by far the largest effort was aimed at developing atmospheric models. Finally, and importantly, the model-users of the 1960s were mostly the same as the model-developers, whereas today users vastly outnumber developers.

5. The 1970s

a. More modeling groups

During the 1970s, more global modeling projects started up, in various places around the world, including at the Met Office in Bracknell (Gilchrist et al. 1973; Rowntree 1976; Corby et al. 1977; Rowntree and Walker 1978), and the Laboratoire de Météorologie Dynamique (LMD) in Paris (Laval and Sadourny 1979; Laval et al. 1981b,a; Sadourny 1984). In the United States, the National Aeronautics and Space Administration (NASA) was motivated to enter the global modeling arena by a desire to maximize the meteorological utility of satellite data. Data assimilation is a process that combines new observations with preexisting information (often in the form of previous short-term forecasts), to provide an optimal estimate of the state of the atmosphere. Weather forecasts use data assimilation to create the “initial conditions” used to start a forecast. In work carried out at NASA’s Goddard Institute for Space Studies (GISS), in New York City, Charney et al. (1969) pointed to data assimilation, and especially the assimilation of satellite data, as an important new application of numerical models. To enable NASA’s work on data assimilation, a version of the UCLA model was provided to GISS in the early 1970s (Somerville et al. 1974). Data assimilation is
now key to operational NWP, and to the production of "reanalyses," which are discussed later in this chapter. See the chapter in this volume by Stanley Benjamin and colleagues for a more complete discussion of data assimilation (Benjamin et al. 2019).

GISS was originally organized to study astronomical problems, in which radiative transfer is of course central. Radiative transfer studies at GISS were strongly influenced by methods that had been developed by the planetary atmospheres community, and these were adapted for use in global atmospheric models. For example, the adding method for computing the transport of radiation in scattering atmospheres is attributed by Lacis and Hansen (1974) to papers describing gamma-ray transfer, although the atmospheric formulation arose from a collaboration between James Hansen and Hendrik van de Hulst (A. Lacis 2017, personal communication). GISS was the first modeling center to use a $k$ distribution to model the spectral variation in optical depth (Somerville et al. 1974; Hansen et al. 1983). In a $k$ distribution, spectral regions with each band are ordered by extinction absorption coefficient, so that the integral over wavelength becomes smooth and just a few quadrature points provide high accuracy.

Notably, the 1970s saw the beginning of operational global numerical weather prediction (Stackpole 1978; Woods 2006), and the founding of the European Centre for Medium-Range Weather Forecasts (ECMWF; Woods 2006), which quickly established itself as the most skillful of the operational centers.

b. Atmospheric dynamical cores

1) THE SPECTRAL METHOD BECOMES POPULAR

During the 1970s and early 1980s, the global spectral method (Silberman 1954; Robert 1966; Baer 1972; Bourke 1974) became widely used in the dynamical cores of atmospheric models. In this approach, the horizontal distribution of model fields is represented by an expansion in spherical harmonics (Fig. 12-6). The spectral representation allows horizontal derivatives to be calculated very accurately and, with a triangular truncation of the expansion, gives homogeneous and isotropic resolution. Moreover, a spectral dynamical core that solves the barotropic vorticity equation conserves energy and enstrophy, as in the continuous system. The calculation of quadratic nonlinear terms directly from the spectral representation using interaction coefficients was prohibitively expensive, and for other types of nonlinearity even more so. This barrier to the use of the spectral method was removed with the introduction of the spectral transform method by Eliassen et al. (1970) and Orszag (1970). In the spectral transform method, the nonlinear advection terms, along with any terms based on physical parameterizations, are computed in grid space, and efficient transforms are used to go back and forth between grid space and the spectral representation (Jarraud and Simmons 1983).

A further important advantage of the spectral method is that it greatly facilitates the use of a semi-implicit time integration scheme. An implicit treatment of the terms responsible for fast gravity waves effectively enlarges the domain of dependence of the numerical solution, allowing the CFL criterion to be satisfied with larger time steps. The price to pay was the reappearance of an elliptic problem to be solved at each time step. Inspired by the work of Marchuk, semi-implicit schemes were proposed for both gridpoint models (Kwizak and Robert 1971; Robert et al. 1972) and spectral models (Robert 1969; Bourke 1974; Hoskins and Simmons 1975). A major advantage of the spectral method is that it allows fast (i.e., computationally inexpensive) solution of the semi-implicit elliptic problem that arises with semi-implicit time differencing. This, in turn, allows spectral models to use long time steps, which enhances their computational speed.

---

6To avoid confusion: The spectral method used in dynamical cores is a mathematical technique based on functional expansions, and has nothing to do with the electromagnetic spectrum that is dealt with by radiation parameterizations, or the apocryphal spectral energy gap mentioned earlier.

7As mentioned above, elliptic problems arise in the solution of the earlier quasigeostrophic and balanced models.
As a result of these strengths of the spectral method, it was soon adopted by GFDL, NCAR, and ECMWF, and it dominated atmospheric modeling efforts around the world for the next two decades (see the review by Williamson 2007). It is still used today at several major modeling centers.

2) IMPROVEMENTS TO GRIDPOINT MODELS

For the modeling centers that persevered with gridpoint methods, important progress was made along two lines. One was the understanding that, in order to adequately capture geostrophic balance, it is necessary to adequately simulate the adjustment toward balance that occurs through the radiation of gravity waves. Ideally, nonpropagating computational modes should be avoided and the entire wave spectrum should have group velocities of the correct sign. These properties depend crucially on the staggering of variables on the grid, and systematic study (Winninghoff 1968; Arakawa and Lamb 1977; Randall 1994) concluded that the B grid (for large $\Delta x/L_R$), C grid (for small $\Delta x/L_R$), and Z grid (for all $\Delta x/L_R$) horizontal staggerings perform best (Fig. 12-7). Here $\Delta x$ is the grid spacing and $L_R$ is a key dynamical length scale called the Rossby radius of deformation.

Another line of progress built upon Arakawa’s Jacobian work (Arakawa 1966) to develop schemes that conserve energy, enstrophy, or angular momentum for more complete equation sets. These developments involved both horizontal (Sadourny 1975; Burridge and Haseler 1977; Arakawa and Lamb 1981) and vertical (Arakawa and Lamb 1977; Simmons and Burridge 1981) discretizations.

The improved dynamical cores had to adapt to changing computer architectures. The most important architectural change during the 1970s was the introduction of “vector” computing, which became available to many scientists when a Cray-1 computer was delivered to NCAR in 1976. A vector computer can perform arithmetic on lists of numbers (called vectors) much faster than conventional machines.8 To take advantage of the increased speed of the vector hardware, the computer codes of the models had to be rewritten; this entailed a significant amount of programming work, but had many beneficial side effects in addition to the direct benefit of faster-running models.

c. Adding the stratosphere

During the 1970s, some global atmospheric models were extended upward to include the stratosphere. The earliest such model was described by Manabe and Hunt (1968). Later studies include those of Manabe and Mahlman (1976), Schlesinger (1976), and Schlesinger and Mintz (1979). With support from the Climate Impact Assessment Program (CIAP) of the U. S. Department of Transportation, some of the models were used to simulate the effects of supersonic airliners on stratospheric ozone (Johnston 1971; Groeber et al. 1974; National Research Council Climatic Impact Committee 1975; Morrisette 1989). This was the first time that agency funding was made available specifically for the application of global atmospheric models to investigate anthropogenic effects on the climate system. It can perhaps be viewed as a loss of innocence.

The temperature structure of the stratosphere is dominated by radiative processes, so including this layer...
motivated developments in radiative transfer parameterizations. At GFDL, Stephen Fels had an insight that pointed the way to more accurate calculations without large increases computational expense. As Green (1967) showed in a crisp two-page note, the radiative cooling calculation at any level in the atmosphere can be expressed as the sum of exchanges between the level in question and every other level, plus one more term representing the energy lost to the infinite heat sink of the rest of the universe—the cooling-to-space term. Temperatures throughout an atmospheric column, and hence the emitted longwave fluxes, vary by much less than the contrast between the atmosphere and outer space, so that when the cool-to-space term is nonzero it is usually much larger than the exchange term. Faced with the limited power of early computers, Fels and Schwarzkopf (1975) exploited this asymmetry in the simplified exchange approximation, the heart of which is a spectrally detailed, and hence more accurate, treatment of cooling to space, and a spectrally coarse treatment of regions in which intra-atmospheric exchanges dominate. The approach is one of the first in which a focus on a specific problem—for the simplified exchange approximation, the computation of heating rates within the atmosphere—allows for algorithms that save time by targeting that calculation. The parameterization was quickly adopted by the radiation community, and incorporated into GFDL’s SKYHI model in 1979.

d. Boundary layer and cloud parameterizations during the 1970s

1) BOUNDARY LAYER PARAMETERIZATIONS

The early ECMWF model used a surface-flux parameterization developed by Louis (1979). It was based on Monin–Obukhov similarity theory, but with some modifications to facilitate use in atmospheric models. The Louis parameterization is still very widely used today.

During the 1960s, a new approach to turbulence parameterization, called “higher-order closure,” emerged within the engineering community (Glushko 1966; Bradshaw et al. 1967; Beckwith and Bushnell 1968; Donaldson and Rosenbaum 1969). A few years later, an essay by Donaldson (1973) introduced higher-order closure to the atmospheric sciences. Soon thereafter, Mellor and Yamada (1974) proposed a detailed hierarchical approach for the application of higher-order closure to atmospheric modeling. Miyakoda and Sirutis (1977) were the first to test higher-order closure in a global atmospheric model. Higher-order closure has been of lasting and recently increasing importance for atmospheric modeling, so we devote some space to it here.

Higher-order closure uses the equations that govern selected “moments” of the subgrid-scale variables. The first moments are the gridcell-averaged values of the primary variables, which might include the liquid water potential temperature $\theta_l$, total water mixing ratio $q_{tot}$, and the three velocity components $u$, $v$, and $w$. These gridcell averages are directly predicted by the model’s dynamical core. Second moments (computed in terms of departures from the means) include variances and fluxes, for example, $(\theta_l)^2$ and $w \theta'_l$. Here a prime denotes a departure from a gridcell average. Third moments include fluxes of second moments, such as $w'w'\theta'$. A model that uses the equations for selected second moments but parameterizes the third moments is called a second-order closure model. A model that uses the equations for selected second and third moments but parameterizes the fourth moments is called a third-order closure model. Closures beyond third order are impractical.

Higher-order closure models need closures for four things:

1) the effects of higher moments that are not predicted, for example, as mentioned above, the third moments in a second-order closure model;
2) moments involving the pressure, which occur in the equations for the moments that involve velocity components;
3) dissipation terms, which are especially important in the equations governing variances; and
4) moments involving heating, precipitation, and other diabatic processes.

At first, it was hoped that second-order closure models would succeed in realistically representing the clear convective boundary layer. Experience showed, however, that second-order closures do not transport turbulence kinetic energy (TKE) realistically. As a result, the boundary layer deepens too slowly in second-order closure models Mellor and Yamada (e.g., 1974). This discovery motivated the development of third-order closure models (e.g., André et al. 1976), which more realistically transport TKE.

Ever since the 1970s, the literature on higher-order closure has been closely linked with the literature on cloud parameterizations, which were receiving greater attention in part because of more and better satellite observations of the atmosphere (e.g., Stowe et al. 1988; Schiffer and Rossow 1983). An important advance came when the equations of higher-order closure were applied to parameterize fractional cloudiness. Sommeria and Deardorff (1977) and Mellor (1977)
independently proposed combining higher-order closure with assumed probability density functions. See also Manton and Cotton (1977) and Chen (1991). The idea is that within the small grid cells of an LES, $\theta_i$ and $q_{i,\text{tot}}$ can be assumed to have a joint Gaussian distribution. Sommerv, Deardorff, and Mellor showed how this assumption can be used to diagnose the fractional cloudiness from the means, variances and covariance of $\theta_i$ and $q_{i,\text{tot}}$. Their approach was to predict the variances and covariance using second-order closure, and then use the predicted (first and) second moments to determine the parameters of the assumed joint Gaussian for each grid cell of a model. The joint distribution could then be used to diagnose the fractional cloudiness, and also the liquid water content of the clouds. As discussed in section 7e, more highly evolved parameterizations based on assumed distributions with higher-order closure are now increasingly being used in global atmospheric models.

2) CUMULUS PARAMETERIZATIONS

Cumulus parameterization underwent major theoretical advances during the 1970s, supported by new field observations. Arakawa and Schubert (1974, hereafter AS) proposed a very influential cumulus parameterization with several important new ideas. First, following Arakawa (1969), they allowed a spectrum of cumulus cloud sizes, distinguished by their fractional entrainment rates, and each with its own mass flux. Second, they determined the intensity of convective activity using the hypothesis of “quasiequilibrium,” which asserts that the cumulus clouds consume convective available potential energy (almost) as rapidly as it is generated by other processes. Third, they included a very simple but explicit representation of the interactions between the cumulus clouds and the subcloud boundary layer. Finally, AS allowed the cumulus updrafts to detrain liquid water and ice (and also water vapor) into the environment, thus providing a “hook” that can be used in a parameterization of convectively generated stratiform clouds. It is noteworthy that AS cited a total of nine papers that were authored or co-authored by Joanne Simpson.

Although AS appreciated that stratiform clouds often form in the outflows from cumulus clouds, modeling research at the time emphasized the role of convection, and tended to treat stratiform clouds as having significance only for their radiative effects. This paradigm was challenged by Houze (1977), who used an analysis of tropical field data to demonstrate that about 40% of the precipitation in a tropical convective system is stratiform in nature. Stratiform clouds received increased attention in subsequent model development efforts. Models also used the mass-flux approach to include convective momentum transport by both updrafts and downdrafts, with the simplifying assumption that horizontal momentum is conserved within updrafts and downdrafts except for the effects of entrainment.

AS also neglected the effects of convective downdrafts, which had been recognized in observational studies (Starr Malkus 1955). Johnson (1976) proposed a way to include downdrafts in a cumulus parameterization, and more such work followed (e.g., Emanuel 1981; Cheng and Arakawa 1997).

e. THE GFDL-BASED FAMILY OF OCEAN MODELS

Here we depart from the decade-by-decade organization of this chapter to describe a “family tree” of ocean models that sprang from the Bryan (1969b) model, which was developed during the 1960s. The tree began to grow during the 1970s, and continues to put out new branches in the twenty-first century.

1) DESCENDANTS OF THE BRYAN (1969B) OCEAN MODEL

As mentioned in section 4a, GFDL developed the Bryan–Cox ocean model during the 1960s. The model underwent extensive further development during the 1970s, and beyond. Figure 12-8 shows a flow diagram illustrating the lineage of ocean circulation models originating from the Bryan–Cox code. In addition to details offered in the extended figure caption, we highlight certain elements of the developments in the main text. The Bryan–Cox code was enhanced by Albert J. Semtner, Jr., who joined the global modeling group at UCLA after completing his Ph.D. at Princeton (Semtner and Mintz 1977). Semtner’s version of the code incorporated arbitrary land–sea masking (allowing for more realistic domains) and upgrades to the computational efficiency on vector machines (Semtner 1974). Semtner’s enhancements were incorporated into the Cox (1984) code, thus initiating a practice of sharing algorithmic upgrades among a community of developers. The Killworth et al. (1991) algorithm to include a free-surface option was also incorporated into the code. The Bryan–Cox–Semtner code was used for the first simulations of the global ocean at 1/2° resolution (Semtner and Chervin 1992). These simulations ushered in the era of global ocean models that admit transient mesoscale eddy activity (see Hecht et al. 2008 for a more recent compendium).

The Bryan–Cox–Semtner code was also used in the Parallel Ocean Climate Model (POCM) developed at NCAR during the 1990s. POCM was one of the first ocean models to make efficient use of the massively
parallel computer architectures that are now standard in the community.

In 1989, Mike Cox died at a relatively young age, at which point Ron Pacanowski, Keith Dixon, and Tony Rosati at GFDL took charge of the GFDL ocean model. Their efforts led to the first version of the Modular Ocean Model (MOM1), thus furthering the GFDL lineage that continues to this day with MOM6. MOM1 was the ocean component of many global climate models in the 1990s, such as the first climate model developed by the Met Office (Murphy 1995), and the second version of the Canadian Climate Center model (Flato et al. 2000). Climate models in Germany, Japan, and Australia also made use of MOM1.

2) SUPPORT FOR A COMMUNITY OF NUMERICAL MODELERS

The Semtner (1974) code and technical report were made available to other ocean modelers, which led to a much wider use of the GFDL code, especially in the United States and the United Kingdom. This idea of sharing code was then formalized in 1984 when Mike Cox made the GFDL ocean code freely available to the public (Cox 1984). The code could be configured to suit the scientific interests of the investigators. This promoted its use as an experimental tool for scientific investigation. Use of the Bryan–Cox–Semtner code thus spread through the ocean and climate modeling community worldwide. These efforts at community development are widespread in today’s world of open-source code development, but they were unique in the late 1970s and early 1980s. In addition to the FORTRAN code, Cox provided an updated technical manual

---

Fig. 12-8. Flow diagram showing relationships among numerical ocean codes originating from the methods of Bryan (1969b). The Bryan (1969b) algorithm was the basis for Cox’s code at GFDL and the starting point for extensions made by Semtner (1974) at UCLA. The Semtner (1974) branch on the left led to the Parallel Ocean Climate Model (POCM) used at NCAR and the Naval Postgraduate School (NPS). It also fostered the Parallel Ocean Program (POP) developed at the Los Alamos National Laboratory (LANL). The Cox (1984) code formed the basis of the Fine Resolution Antarctic Model (FRAM) developed in the United Kingdom by Peter Killworth and David Webb. The OCCAM project in the United Kingdom (Webb et al. 1998) was among the first global models with active mesoscale eddy variability (using resolutions as fine as 1/12°). On the right side of the diagram are Modular Ocean Model versions 1 and 2 (MOM1 and MOM2), representing the GFDL descendants of the Bryan–Cox code. On the far right are the NCAR efforts with NCOM (Gent et al. 1998, NCAR CSM Ocean Model) and CSM (Climate System Model). [Courtesy of Albert Semtner, Jr.]
describing the mathematical equations and numerical methods that formed the basis for the code. The Semtner (1974) technical report and the Cox (1984) manual proved extremely valuable in communicating the scientific and engineering rationales for various features of the model. As a result, the code was readily understandable by a broad community of oceanographers and numerical algorithm specialists.

These pioneering efforts at building a community of informed users paved a path toward enhancing the scientific integrity, transparency, and reproducibility of ocean model codes and the simulations produced with them (a formidable task to this day!). It also fostered several allied efforts to use the Bryan–Cox–Semtner code for a suite of scientific applications, and to enhance the physical parameterizations, numerical methods, and computational efficiency of the models.

3) OCEAN CODES INSPIRED BY MOM

The Parallel Ocean Program (POP) is a direct descendant of the Bryan–Cox–Semtner code. It was developed in the early 1990s for the Connection Machine by Rick Smith, John Dukowicz, and Bob Malone at Los Alamos National Laboratory (LANL; Smith et al. 1992). An implicit-free-surface formulation and other numerical improvements were added by Dukowicz and Smith (1994). Later, the capability for general orthogonal coordinates for the horizontal mesh was implemented (Smith et al. 1995). See also Murray (1996) for efforts with the Bryan–Cox–Semtner code and Madec et al. (1997) for efforts with the Ocean Parallélisé (OPA) model in France. In 2001, POP was adopted as the ocean component of the Community Climate System Model (CCSM) based at NCAR. Substantial efforts at both the LANL and NCAR have gone into adding various features to meet the needs of the CCSM (Smith et al. 2010; Danabasoglu et al. 2006, 2012). The POP code has been used as the ocean component of the CCSM, and versions 1 and 2 of the Community Earth System Model (Hurrell et al. 2013).

Upon the release of the Cox code in 1984, scientists around the world had access to the fruits of more than 20 years of focused efforts at GFDL. Nonetheless, as scientists are prone to do, many arrived at distinct ideas for how best to go about developing numerical models. One such effort is the ocean component of the Nucleus for European Modeling of the Ocean (NEMO). This code was developed from the OPA model, release 8.2; (Madec et al. 1997). The NEMO code has been used for a wide range of applications, both regional and global, as a forced ocean model and as a component of a climate model. In particular, it is used today in the global models of the Met Office, the European Centre for Medium-Range Weather Forecasts, and the French National Centre for Scientific Research.

The Max Planck Institute ocean model (MPIOM) is the ocean–sea ice component of the MPI Earth System Model. MPIOM is a primitive equation model (C grid, z coordinates, free surface) with the hydrostatic and Boussinesq assumptions. It includes a bottom boundary layer scheme for the flow across steep topography, and uses a curvilinear orthogonal grid, which allows for a variety of configurations. A description of MPIOM can be found in Marsland et al. (2003). A list of model development efforts that is current up to the year 2000 can be found in Griffies et al. (2000). Any list is incomplete, and we do not attempt an update here.

f. Sea ice advances during the 1970s and 1980s

The 1970s and 1980s were a golden age for the development of sea ice models, with major advances in the treatment of sea ice thermodynamics and the emergence of models that simulate sea ice dynamics, in which mechanical failure causes ridging and rafting among floes and also creates openings between floes known as leads. The regional jumble of sea ice caused by the interplay of deformation, growth, and melt results in a distribution of thicknesses that modelers wanted to simulate in order to capture the highly nonlinear thickness dependence on compressive stress and growth.

Observations and scientific understanding of sea ice had recently expanded as a result of the International Geophysical Year (IGY) in 1957–58. Norbert Untersteiner spent a year on the sea ice as chief scientist of an IGY field camp. In the decade that followed he published a series of papers that established the basic principles that govern a numerical model of sea ice thermodynamics. Together with graduate student Gary Maykut of University of Washington, he assembled a sea ice model that treated the surface energy budget and sea ice growth and melt with the unique dependence on sea ice brine pockets (Maykut and Untersteiner 1971). The concentration of brine in the pockets varies with heat stored in the sea ice. The temperature and brine concentration were simulated in 10-cm layers that absorbed sunlight and conducted heat between ocean and atmosphere. The physical interactions were so complex that their model was limited to just the vertical dimension, and because of its computational expense no climate or weather model adopted the Maykut and Untersteiner model of brine-pocket dynamics until the twenty-first century.

The RAND Corporation sought to create a simpler thermodynamic sea ice model to couple to early ocean
models. RAND commissioned Bert Semtner to reduce the complexity of the Maykut and Untersteiner model. Semtner did so by developing a very simple one-layer model of sea ice alone and an innovative three-layer model (two layers of sea ice and one of snow) with a reservoir of interior solar heating to mimic the effect of brine pockets and shift the timing of the surface melt season in a semirealistic way (Semtner 1976). These two reduced-complexity models by Semtner were the basis for sea ice thermodynamics in global climate models for decades.

In the fall of 1976, sea ice scientist William Hibler became the 25th visitor to GFDL. He was impressed by the practical issues of sea ice modeling in a global climate model. He learned how ocean models were formulated from his host Frank Bryan, who inspired him to simplify the sea ice model from the Arctic Ice Dynamics Joint Experiment (AIDJEX; Coon et al. 1974), which employed a constitutive law for plastic behavior to simulate the dependence of the stress tensor on the velocity field, allowing for material failure and deformation. Hibler (1979) greatly reduced the numerical complexity of the AIDJEX model by formulating a nonlinear viscous-plastic rheology for sea ice. He demonstrated the scheme in an 8-yr simulation of the Arctic basin. The AIDJEX model and Hibler’s viscous-plastic scheme remain the basis for the dynamics in most sea ice models used in climate models today, though many climate models, including GFDL’s, used highly idealized methods such as free drift with stoppage to model sea ice dynamics for several more decades as efforts continued to further reduce the computation demands of the viscous-plastic dynamics.

As Hibler and other sea ice modelers explored methods to simulate sea ice dynamics, the need for a subgrid-scale parameterization to simulate the distribution of sea ice thicknesses arose. An equation to describe the ice-thickness distribution was developed by Alan Thorndike with other colleagues at the University of Washington (Thorndike et al. 1975). Hibler soon implemented an ice-thickness distribution scheme in his Arctic basin model (Hibler 1980).

The advanced sea ice models developed during the 1980s were used only in experimental applications, occasionally coupled to an ocean model. They were not coupled to an atmosphere model for another decade. One reason is that climate modeling centers considered advanced sea ice models to be too computationally demanding. Another factor was that the focus of global climate modeling remained primarily on the tropics and northern midlatitudes until the twenty-first century.

### g. Simulations of global warming

Manabe and Möller (1961) demonstrated that radiation is roughly balanced by convection (Manabe and Strickler 1964).\(^{10}\) The GFDL team performed pioneering one-dimensional simulations of “radiative-convective equilibrium” (RCE; Manabe and Strickler 1964; Manabe and Wetherald 1967), an idealization that continues to be useful today (e.g., Wing et al. 2018). To mention one very important example, the study of Manabe and Wetherald (1967) pointed to the importance of the water vapor feedback for climate change. As noted by Manabe and Strickler (1964) in a paper describing single-column modeling of RCE, “one of the major purposes of our study is the construction of a model of radiative transfer simple enough to be incorporated into a general circulation model of the atmosphere.”

During the 1960s, Manabe and Wetherald (1967) had already studied the effects of increasing atmospheric carbon dioxide concentrations on the “climate” of a one-dimensional RCE model, and this work pointed to the importance of the water-vapor feedback on climate change. The first simulation of global warming with a true climate model was reported by Manabe and Wetherald (1975). Their model was idealized through the use of a limited computational domain, simplified topography, no energy transport by the oceans, no seasonal or diurnal cycles, and fixed cloudiness. It is remarkable that this first simulation with a simplified model, more than 40 years ago, predicted many changes that have now been observed in the real atmosphere, including a warming troposphere with greater warming near the pole, a cooling stratosphere, stronger precipitation, and increased atmospheric water vapor. The successful strategy of Manabe and Wetherald (1975), Manabe and Stouffer (1980), and Manabe and Wetherald (1980) was to explore the possibility of anthropogenic climate change using the relatively simple models available at the time, rather than waiting for the more complete models of the future.

### 6. The 1980s

#### a. Community modeling gets under way

In 1983, NCAR released the Community Climate Model (CCM) (Pitcher et al. 1983; Williamson 1983; Williamson et al. 1983; Kiehl et al. 1998). Initially, the CCM was essentially an atmosphere model coupled to a simple land surface model. It lacked a coupled ocean

---

\(^{10}\)That radiation and convection balance each other was appreciated before 1964, as is made clear by George Simpson’s comments on the paper of Guy Callendar (1938).
model, so calling it a “climate model” was a bit of an exaggeration. The CCM was widely used because it was freely available and fully documented.

During the 1980s, Washington and Meehl (1983), Washington and Meehl (1984) and Washington and Meehl (1989) used versions of the CCM to perform increasingly detailed simulations of anthropogenic greenhouse warming. In the late 1990s, Washington et al. (2000) developed the Parallel Climate Model (PCM). The atmosphere component was the CCM3 at T42 resolution, and the ocean component was the POP model at about 0.5° resolution. The PCM was one of the first models designed to run very efficiently on the parallel computers that were emerging at that time. The PCM was subsequently used to run ensembles of twentieth-century simulations forced by the individual climate forcings, such as greenhouse gases, aerosols and solar variability, rather than their combined effects. The interesting results are presented in Meehl et al. (2003). The PCM became very popular and well-known, particularly following Ramanathan’s and Hansen’s papers (Ramanathan et al. 1983; Hansen et al. 1983). We [GFDL] felt that to create proper energy balance, especially when doing climate calculations, you needed to have methane, nitrous oxide, and chlorofluorocarbons (interview with Ramaswamy, 28 November 2017).

As Stephens (1984) notes, the treatment of interactions between clouds and radiation in global models in the mid-1980s was fairly rudimentary. Cloud properties were almost universally prescribed, perhaps as a function of relative humidity but frequently as a function of location and season, with limited spectral information. Methods for more sophisticated treatments were already in place, with insights from work on planetary atmosphere (Hansen and Travis 1974) informing complete parameterizations in both spectral regions (Stephens 1978; Roach and Slingo 1979; Slingo and Schrecker 1982). The delta-scaling method for treating the sharp forward peaks in the scattering phase function of clouds had been developed (Potter 1970; Joseph et al. 1976) and the variety of proposed two-stream methods had been unified by Meador and Weaver (1980) and Zdunkowski et al. (1980). These tools were in place as models began to make more detailed calculations of cloud properties (section 6d), although treatments of “cloud overlap,” that is, how radiation was partitioned between clear and cloudy skies in the vertical, remained simple.

The move to better prediction of cloud properties was partly motivated by recognition of the role of cloud-radiation interactions in shaping the large-scale circulation including tropical convection. One example came from a small group working with the UCLA/Goddard Laboratory for Atmospheres (GLA) GCM, who demonstrated that predicting cloud properties and allowing those variable properties to influence the radiation field (Harschvvardhan et al. 1987) had tremendous impacts on the global distribution of cloudiness and the resulting gains, the method was soon adopted at a number of operational centers (Ritchie 1991; Ritchie et al. 1995).

c. Radiative transfer work in the 1980s

Stephens (1984) provides a useful view of the state of radiation parameterization in the early 1980s. Although more parameterizations were available than during the 1960s and 1970s (Stephens describes six), the ideas can all be traced back to the earliest treatments. One significant change was the addition of a wider range of gases, especially in models devoted to longer climate simulations as opposed to short-term weather forecasts. As Ramaswamy explained,

in the eighties, the so-called OTGs, the other trace gases became very popular and well-known, particular following Ramanathan’s and Hansen’s papers (Ramanathan et al. 1983; Hansen et al. 1983).
energy budget (Randall et al. 1989; Harshvardhan et al. 1989).

The importance of radiation for short-term weather forecasts was also becoming clear. Jean-Jacques Morcrette, now retired from ECMWF, recalled how more accurate forecasts was also becoming clear. Jean-Jacques Morcrette, (but more computationally expensive) radiation calculations helped to increase forecast skill:

I got the feeling that at least part of the problem might be in the clear-sky longwave cooling of the descending branch of the Hadley circulation (Now it is easy to state it so simply, at the time it was not so clear-cut.) The temperature dependence of the longwave absorption is very different in my band models (Morcrette 1990, 1991) from that in the then-current ECMWF scheme (essentially from Geleyn and Hollingsworth 1979). It took me till April 1989 to convince people that some revised version of the Lille codes [from which Morcrette’s codes have been adapted] were better overall even if they were less computer-efficient. The main impact is a much more stable maintenance of the Hadley circulation, which previously tended to weaken with the length of the forecast, and an increased geographical contrast in cloud forcing (J.-J. Morcrette 2017, personal communication).

The many independent spectral calculations required for broadband calculations, and the usual desire to compute fluxes with and without clouds to help understand the role of clouds in Earth’s energy budget, make radiation a computationally large burden. In most models, tendencies due to radiation are computed less frequently in time than other physical processes, but the need for computational efficiency has motivated other interesting compromises. Two approaches at NASA GISS do not seem to have been reported in the literature although they have been used since the original models described by Hansen et al. (1983).

The first is that the sampling in time is often done at intervals that are not divisors of the daily cycle, so that the entire diurnal cycle is eventually sampled. Andrew Lacis described how this arose:

[when sampling the diurnal cycle regularly] you get beat frequencies in there—pressure waves building up and stuff like that. When you make that odd fraction it would eliminate some of that type of noise. The other thing we did to speed up the radiation was sampling—we did every other grid box. So with sampling every 2 and a half hours and every other gridbox I think radiation might’ve been taking maybe 25 percent of the computing time (interview with Andrew Lacis, 26 October 2017).

The second is that fractional cloudiness is treated by sampling in time:

so if this cloud has a 50 percent chance of being there we’ll call it clear and smaller than a half we’ll call it cloudy. That idea might’ve come from Larry Travis, at least according to Bill Rossow. The rationale for doing that came from Charney, who basically said that random errors in climate model don’t matter that much but systematic errors do (interview with Andrew Lacis, 26 October 2017).

The 1980s also saw the development of widely available reference models and data for making radiative transfer calculations. For problems in clear-sky radiative transfer, this included the publication of spectroscopic databases covering an increasingly broad set of gases [e.g., the high-resolution transmission molecular absorption database (HITRAN); Rothman et al. 1987] and the development of line-by-line radiative transfer models (e.g., LBLRTM; Clough et al. 1992) capable of computing optical depths given the state of the atmosphere. Calculations involving clouds were made more tractable by the widespread availability of codes for doing Mie calculations (Wiscombe 1980) for single-scattering properties and discrete-ordinates calculations (Stamnes et al. 1988) to obtain the angularly resolved radiation field.

These codes provided an opportunity to test parameterizations against reference results. The first was the Intercomparison of Radiation Codes Used in Climate Models (ICRCCM; see Ellingson and Fouquart 1991; Ellingson et al. 1991; Fouquart et al. 1991). ICRCCM argued for the use of reference models, rather than direct observations, as the standard for radiation intercomparisons, given the difficulties in making simultaneous comparable measurements at the bottom and top of the atmosphere. The broad lesson from the intercomparison effort (Fig. 12-9) was that line-by-line models agreed to within a few percent (unsurprisingly, given that many use the same underlying spectroscopic information), but that parameterizations used in weather and climate models could be substantially in error, especially with respect to radiative forcing, that is, the sensitivity of changes in fluxes to changes in composition. The profiles used in ICRCCM were quite idealized, however, making it difficult to estimate the magnitude of errors in weather forecasting or climate projection applications, a problem that persists in more recent assessments (Oreopoulos et al. 2012; Pincus et al. 2015).

d. Boundary layer and cloud parameterizations during the 1980s

Deardorff (1972a) emphasized the importance and highly variable nature of the depth of the boundary layer, especially over land. He proposed a boundary layer parameterization in which the depth of the boundary layer is an explicit prognostic (i.e., time stepped)
variable of the model. Deardorff’s idea was implemented in the UCLA model by Randall (1976) and Suarez et al. (1983). A parameterization of stratocumulus clouds was included following Lilly (1968), and the boundary layer parameterization was coupled with the cumulus parameterization of Arakawa and Schubert (1974) by allowing the cumulus clouds to remove mass from the boundary layer. The model’s vertical coordinate system was modified to make the boundary layer top an internal coordinate surface. With this approach, the model layers below the boundary layer top comprised the boundary layer, and the depth of the boundary layer could change in response to mass fluxes across the boundary layer top because of entrainment and cumulus convection. Randall et al. (1985) analyzed change simulations with global models (Charney et al. 1974) by allowing the cumulus clouds to remove mass from the boundary layer. The model’s vertical coordinate system was modified to make the boundary layer top an internal coordinate surface. With this approach, the model layers below the boundary layer top comprised the boundary layer, and the depth of the boundary layer could change in response to mass fluxes across the boundary layer top because of entrainment and cumulus convection. Randall et al. (1985) analyzed change simulations with the model, and reported the results of some numerical experiments, including one in which the boundary layer depth was artificially held constant, and another in which the diurnal cycle of solar radiation was replaced by daily mean insolation. The results showed the importance of variations of the boundary layer depth for precipitation over land and for determining the amount of low-level clouds.

We mention four advances in cumulus parameterization during the 1980s. Emanuel developed a similarity theory of convective downdrafts (Emanuel 1981). Raymond and Blyth (1986) proposed that mixed parcels created through entrainment migrate to their levels of neutral buoyancy. This idea, called “buoyancy sorting,” has been very influential. The Betts–Miller parameterization (Betts 1986; Betts and Miller 1986), developed at ECMWF, used an adjustment to empirically determined soundings of both temperature and water vapor. Finally, the Tiedtke convection parameterization (Tiedtke 1989) was implemented in the ECMWF model. It used the moisture-convergence closure developed by Kuo (1965), but a later version used by the Max Planck Institute for Meteorology was modified to use a buoyancy closure (Nordeng 1994).

Cloud microphysical parameterizations began to appear in global atmospheric models during the 1980s. The earliest low-resolution mesoscale models developed in the 1970s used the gridscale saturation removal method for calculating surface precipitation, similar to what was done in early operational NWP models. However, cloud-scale models around this time quickly adopted a Kessler-like approach with separate equations for cloud and rain mass (e.g., Klemp and Wilhelmson 1978). By the early to mid-1980s many mesoscale models had also adopted this type of approach (e.g., Hsie et al. 1984). Around this time both mesoscale and cloud models also began incorporating ice microphysics; commonly used schemes included those from Lin et al. (1983) and Rutledge and Hobbs (1984). These schemes generally assumed two or three ice categories (cloud or small ice, snow, and graupel or hail) and included conversion processes between the categories analogous to the Kessler approach for liquid microphysics. Beginning in the late 1960s and 1970s detailed bin microphysical schemes that explicitly evolved the particle size or mass distributions by predicting the total water mass in size or mass bins were also developed (e.g., Bleck 1970; Berry and Reinhardt 1974), but computational cost precluded their wider use in models until the 1990s and 2000s.

As noted above, larger-scale NWP models at operational centers through the 1970s and 1980s continued to convert water vapor to surface precipitation when the water vapor mixing ratio exceeded some threshold value. Microphysical processes were not considered in this approach. By the late 1990s, the operational Eta Model at the U.S. National Centers for Environmental Prediction (NCEP) adopted a prognostic cloud scheme that explicitly included evolution equations for cloud condensate and a diagnostic treatment of precipitation from the predicted cloud fields (Zhao et al. 1997). Both the cloud fraction and predicted cloud water content were accounted for in the radiation parameterization. Some forecast models with prognostic cloud condensate included more detailed representations of subgrid-scale condensation for both stratiform and convective clouds (Sundqvist et al. 1989) as well as prognostic cloud fraction (Tiedtke 1993). These models typically partitioned condensate as liquid or ice according to temperature. These were important advances for operational forecast models, but the representation of microphysics was still highly simplified compared to contemporaneous finer mesh mesoscale and cloud models that employed several prognostic categories of cloud and precipitation water.

The representation of the hydrologic cycle in global climate models in the 1970s through the 1990s was generally at a similar level of complexity to that of operational NWP at the time, but with more sophisticated diagnostic parameterizations to represent the cloud fraction and optical properties. The diagnostic schemes of the 1980s were used to predict the occurrence and radiative properties of clouds based on the relative humidity, vertical velocity, temperature, and/or precipitation rate (e.g., Slingo 1985; Wetherald and Manabe 1988).

The 1980s brought an increased emphasis on the radiative effects of clouds, for which greatly improved observations were becoming available (Schiffer and Rossow 1983; Barkstrom 1984; Ramanathan et al. 1989). Cloud feedback became an important focus of climate change simulations with global models (Charney et al.
Fig. 12-9. The treatments of radiation in global models have been compared to benchmark “line-by-line” calculations for many decades. (a) Results for longwave fluxes at the surface and the tropopause in a single idealized but quasi-realistic atmosphere. The line-by-line models, shown as plus signs, agree with each other quite well, partly because they share the same spectroscopic data; by comparison both narrow- and
wide-band models (‘‘NBMs’’ and ‘‘WBMs’’ respectively) show significantly more variation. (b) Published 15 years later, focuses on forcing (i.e., the change in flux caused by a change in composition) here the impacts of methane and nitrous oxide. Line-by-line calculations are indistinguishable from one another while the GCMs show variation of 25% or more of the signal in the longwave while entirely ignoring the impact in the shortwave. (c) Appearing almost a decade later still, shows that errors in GCM parameterizations (circles) still swamp those from line-by-line models (squares) in calculations of the forcing by quadrupled carbon dioxide concentrations. Panel (a) is Fig. 15 of Ellingson et al. (1991); (b) is Fig. 6 of Collins et al. (2006b); (c) is redrawn from Fig. 3 and supplemental material in Pincus et al. (2015).
properties, and soil hydraulic properties were prescribed as global maps derived from many disparate sources.

Turbulent fluxes in these models were represented using a network of nodes and resistors (Fig. 12-10) using an "electrical analogy" to Ohm’s law, which was first introduced by Richardson (1922). Temperature and water vapor are treated as potentials, the fluxes of sensible and latent heat among model components as resulting currents proportional to the difference in potentials, and the proportionality coefficients as variable resistors. Ohm’s law thus amounts to diffusion, with diffusivity at the molecular scale for transpiration through plant stomata and at turbulent scales elsewhere in the model. BATS (Dickinson et al. 1986) used a single plant canopy layer, whereas SiB (Sellers et al. 1986) introduced a subcanopy or “understory” of grass or shrubs beneath a taller tree canopy.

Simulation experiments revealed important modes of interaction between the vegetated land surface and the atmosphere that can affect climate. Charney (1975) showed that land clearing and overgrazing could lead to drought through a feedback between surface albedo and enhanced atmospheric subsidence. Shukla and Mintz (1982) demonstrated that evapotranspiration from land exerted a profound influence on Earth’s hydrologic cycle and climate. Dickinson and Henderson-Sellers (1988) used a coupled land–atmosphere GCM to explore the consequences of tropical deforestation. They manipulated model surface parameters to simulate the conversion of the Amazon rain forest to grassland, which resulted in substantial surface warming due primarily to changes in albedo and roughness.

Land surface modelers recognized the very substantial mismatch in spatial scales between microscopic stomata, macroscopic plant canopies, and GCM grid cells that were hundreds of km across. Jarvis and McNaughton (1986) recognized that under low-wind conditions evapotranspiration from vegetation acted to
humidify the air near the surface, reducing the vapor pressure gradient and thereby shifting the energy balance toward sensible heat. They introduced the idea of estimating surface energy fluxes at landscape scales as a continuum between physiological control by stomates and environmental control by radiation, atmospheric humidity, and wind speed. They introduced a coupling coefficient to represent this continuum, and showed that at larger scales transpiration is influenced less by stomata and more by radiation.

Besides the leaf-to-canopy conundrum, scaling local fluxes to GCM grid cells was also problematic because the coupling may include dynamical processes above heterogeneous surfaces that are distributed at subgrid scale. These are very common, arising from juxtaposed farms and cities, forests and pastures, or even locations with and without antecedent rain. Anthes (1984) showed that mesoscale circulations induced by strong gradients in temperature and sensible heat fluxes above wet and dry patches in semiarid regions acted like inland sea breezes to enhance convection along the boundaries between patches. These circulations can interact with both the atmospheric and land states to create mesoscale energy and water fluxes that cannot be obtained by linear averaging of surface fluxes in isolation (Avissar and Pielke 1989; Pinty et al. 1989; Pielke et al. 1991). Pielke et al. (1991) used limited area coupled models and found that they could be significant under conditions of large patches and low mean wind speeds.

Another key development in the 1980s was global mapping of vegetation properties using satellite imagery. Chlorophyll and other pigments absorb strongly in visible wavelengths to drive photosynthesis, but they are highly reflective in the near-infrared part of the solar spectrum (Tucker 1979). A series of polar-orbiting weather satellites maintained by the National Oceanic and Atmospheric Administration to monitor cloud properties produced near daily global coverage of a quantity called the normalized difference vegetation index (NDVI). The NDVI was shown to be highly correlated with plant growth and CO₂ uptake (Tucker et al. 1986; Fung et al. 1987). Algorithms were developed to derive self-consistent vegetation parameters for land surface models from satellite imagery (Sellers 1985). These began to replace the ad hoc and often inconsistent sets of global parameter maps.

g. Reanalysis

Improved global observations of the atmosphere and Earth’s surface, especially from satellites, made global weather analyses a realistic possibility (Bengtsson et al. 1982). Global models and their associated data assimilation systems were essential for the production of these analyses. “Reanalysis” of historical data using the best available global model was advocated by Bengtsson et al. (1982), and soon became a reality (Kalnay et al. 1996; Uppala et al. 2005; Saha et al. 2010; Onogi et al. 2007; Schubert et al. 1993; Rienecker et al. 2011; Gibson et al. 1997; Dee et al. 2011). Reanalyses use a fixed but up-to-date forecast model and data assimilation system to process historical observations over a long record. Fixing the systems avoids some of the temporal discontinuities that occur in a series of routine operational analyses, though not discontinuities caused by large changes in the available observations. Reanalyses have now become essential for atmospheric science research. The application of global ocean models to the reanalysis of ocean observations is at an earlier stage of development (Schiller et al. 2008; Balmaseda et al. 2013), but reanalyses of the coupled ocean–atmosphere system are beginning to appear (Laloyaux et al. 2018).

h. Global warming becomes a societal concern

During the 1960s, climate scientists were aware of the possibility of anthropogenic climate warming due to increasing greenhouse gas concentrations, but the issue had not yet reached the public consciousness. This changed during the 1980s, as observations showed continuing increases in CO₂ concentrations, and the U.S. Congress and other governmental bodies began to take an interest (Shabecoff 1988; Weart 2008). More simulations of anthropogenic climate change were appearing in the literature (e.g., Hansen et al. 1981; Washington and Meehl 1989). Today, ESM development is increasingly driven by the global warming issue.

7. The 1990s

a. New models and new interactions among modeling groups

During the 1990s, important new models were created, and modeling groups began interacting in important new ways.

In 1990 the Met Office Hadley Centre was opened (Folland et al. 2004), creating a dedicated center for research on Earth’s climate (e.g., Senior and Mitchell 2000; Mitchell et al. 1995b). The Hadley Centre’s Unified Model (Cullen 1993; Cullen et al. 1997; Davies et al. 1998) is designed for use in both operational NWP and climate simulation. This has the advantage that operational NWP is an excellent way to test a climate model (e.g., Palmer et al. 2008; Senior et al. 2010).

A version of the ECMWF forecast model was modified to create ECHAM (Roeckner et al. 1989; Simmons et al. 1989; Stevens et al. 2013), a climate model in use at the Max Planck Institute for Meteorology in Hamburg,
Germany. More recently, the center is using a new global atmosphere model called the Icosahedral Non-hydrostatic GCM (ICON), which is based on a geodesic grid and which has been developed in a partnership with the German Weather Service (Wan et al. 2013; Giorgetta et al. 2018). Here again we see a single model being used for both operational NWP and climate simulation.

As mentioned in section 6d, Cess et al. (1989) organized an intercomparison of results from many modeling groups. Additional intercomparisons proliferated during the 1990s. An important example is the Atmospheric Model Intercomparison Project (AMIP; Gates 1992). AMIP was presaged by the study of Lau (1985), who showed that the atmosphere responds strongly and predictably to prescribed observed interannual changes in sea surface temperatures. An AMIP simulation uses an atmospheric model (coupled to a land surface model) with prescribed observed sea surface temperatures for a sequence of real years. An AMIP simulation can be used to test the ability of a global atmospheric model to respond realistically to interannual variability of sea surface temperatures such as that associated with El Niño. The experimental design is similar to that developed by Lau (1985), but follows a formal protocol. AMIP simulations continue to be a valuable and widely used method to test global atmospheric models (e.g., Eyring et al. 2016). Intercomparisons have also been crucial for the work of the Intergovernmental Panel on Climate Change (IPCC), which issued its first assessment report in 1990 (IPCC 1990) and continues its work today (e.g., Stocker et al. 2013). The IPCC is a truly historic enterprise that is strongly reliant on results from ESMs. The Coupled Model Intercomparison Project (CMIP) has been particularly central to the work of the IPCC (Meehl et al. 2000; Covey et al. 2003; Eyring et al. 2016).

Operational seasonal prediction with coupled ocean–atmosphere models began during the 1990s, and has gradually been maturing since (Palmer et al. 2000; Kanamitsu et al. 2002; Woods 2006; Kirtman and Pirani 2009).

b. Atmospheric dynamical cores

During the 1990s spectral semi-implicit semi-Lagrangian models were well established. It was shown that the east–west density of grid points could be reduced near the poles (giving a reduced grid) with negligible loss in accuracy (Hortal and Simmons 1991; Courtier and Naughton 1994). Also, since the advective nonlinearity was now handled by the semi-Lagrangian advection, the extra grid resolution needed to avoid aliasing of quadratic nonlinear terms was no longer necessary, and a coarser “linear grid” could be used (Côté and Staniforth 1988; Williamson 1997). Both of these ideas led to further significant efficiency gains. An additional motivation for the adoption of semi-Lagrangian advection was that spectral advection of water vapor proved to be problematic (Williamson and Rasch 1994).

Around this time, interest was growing in global nonhydrostatic models. This stemmed partly from a desire for unified modeling systems that could operate either globally or at nonhydrostatic scales (Cullen et al. 1997) and partly from an ambition for global modeling that resolved nonhydrostatic scales, which growing computer power would soon permit (Qian et al. 1998; Yeh et al. 2002; Satoh et al. 2008; Matsuno 2016). The fully compressible equations support acoustic waves, and the CFL criterion for an explicit treatment of their vertical propagation would be very restrictive. Therefore, inspired by earlier work on small-scale non-hydrostatic models, one of four options was generally adopted. The first was a fully three-dimensional semi-implicit treatment of acoustic waves (Tapp and White 1976; Cullen et al. 1997; Qian et al. 1998; Yeh et al. 2002); a variety of solution methods for the resulting elliptic problem were tried. The second was a split-time-step method in which acoustic wave propagation was treated via shorter substeps. The third approach is to use a horizontally explicit but vertically implicit (HEVI) time-differencing scheme (Klemp and Wilhelmson 1978; Satoh et al. 2008; Weller et al. 2013). This is motivated by the fact that vertical grid spacing is typically much finer than horizontal grid spacing, so that it is the vertically propagating sound waves that place the most severe limit on a model’s time step. The fourth approach is to use a sufficiently accurate set of equations that filters vertically propagating sound waves (Arakawa and Konor 2009), thus eliminating the problem before discretization begins.

Semi-Lagrangian schemes proved to be very effective for weather forecasting, but for longer climate simulations they were limited by their lack of conservation. This prompted the development of several conservative large-time-step advection schemes (e.g., Harris et al. 2011, and references therein), some of which were eventually incorporated into operational models (Lauritzen and Nair 2008; Wood et al. 2014).

c. The evolution of the vertical coordinates used in atmosphere and ocean models

The vertical coordinate is a fundamental algorithmic choice for atmosphere and ocean models. It determines how subgrid-scale parameterizations manifest and what parameterizations are appropriate. A comprehensive discussion of vertical coordinate systems for hydrostatic
atmosphere models was published by Kasahara (1974). For ocean models, the vertical coordinate determines representations of the upper ocean and bottom boundary layers and the stratified ocean interior as well as their interactions with each other and with the solid Earth (Griffies et al. 2000). Vertical coordinates for numerical modeling of the atmosphere and ocean have been under study at least since the 1950s, and work continues today. We choose to summarize the topic in this section of our chapter because many important ideas emerged during the 1990s.

1) QUASI-EULERIAN VERTICAL COORDINATES

As mentioned earlier, both pressure and height were used as vertical coordinates in early global atmospheric models. These choices are both problematic (especially pressure), in part because the coordinate surfaces intersect the lower boundary. The terrain-following $\sigma$ coordinate of Phillips (1957a) solves that problem by conforming to the lower boundary, but it leads to difficulty in the accurate computation of the horizontal pressure-gradient force above steep topography (Smagorinsky et al. 1967; Kurihara 1968; Sundqvist 1975). The problem arises because with the sigma coordinate the horizontal pressure-gradient force is the sum of two terms. Over steep topography, these two terms are individually large and of opposite sign, and eliminates the problem with the horizontal pressure gradient force near steep terrain by introducing “step mountains” that come in discrete sizes, like off-the-rack clothing. The sizes of the mountains are chosen to match the specified thicknesses of the model layers. For about 25 years, the $\eta$ coordinate was used in the operational Eta Model used for regional prediction by the U.S. National Centers for Environmental Prediction (Janjić 1994); this was mentioned in section 6d. Simmons and Burridge (1981) suggested a different way to address the problem with the horizontal pressure-gradient force near steep terrain, through the use of a hybrid vertical coordinate that behaves like $\sigma$ near the lower boundary but like pressure aloft. Their hybrid approach has been very widely used.

For most applications of large-scale basin and global ocean circulation modeling, the vertical coordinate is based on geopotential (or depth). Similar but more flexible “quasi-Eulerian” approaches have been developed (Adcroft and Hallberg 2006). They can be used with ocean models that retain the traditional Bryan (1969b) algorithmic architecture, in which the vertical motion crossing coordinate surfaces (i.e., vertical velocity) is diagnosed through mass continuity in non-Boussinesq models or volume continuity in Boussinesq models. Of particular note for global climate efforts is the rescaled geopotential coordinate, $z^*$, which allows more flexibility with realistic undulations of the ocean free surface than the traditional geopotential ($z$) coordinate models (Stacey et al. 1995). The $z^*$ coordinate was first implemented in the MITgcm by Adcroft and Campin (2004), and has been used for climate applications with MOM4.1 and MOM5 (Griffies et al. 2005).

Another advance was made by Marshall et al. (2004), who made use of an isomorphism between pressure coordinate non-Boussinesq (compressible) fluids and geopotential coordinate Boussinesq (incompressible) fluids. This made it straightforward to incorporate compressible dynamics into formerly incompressible ocean models. Doing so allows ocean models to include a full representation of oceanographic processes impacting the model’s sea level, including the global thermosteric effects that are missing from Boussinesq models (Griffies and Greatbatch 2012).

The $\sigma$ coordinate has also been adapted for use in ocean models, for example, by Lemarié et al. (2012). Their work, implemented in the Regional Ocean Modeling System (ROMS; Shchepetkin and McWilliams 2005), bridges the gap between regional and global modeling applications. The Russian Institute of Numerical Mathematics Ocean Model (INMOM; Volodin et al. 2010) also uses a global, $\sigma$ coordinate model as the ocean component of an ESM.

2) QUASI-LAGRANGIAN METHODS

Quasi-Lagrangian methods have also been used in both atmosphere and ocean models. In these methods, the vertical “layers” of a model are bounded by surfaces that move with the fluid, as nearly as possible, so that little or no mass crosses layer edges. For atmospheric models, one approach is to use potential temperature $\theta$ as a vertical coordinate; in the absence of heating, the “vertical velocity” vanishes with the $\theta$ coordinate. For ocean models, the corresponding isopycnal approach is to use potential density as the vertical coordinate, so that the vertical velocity vanishes in the absence of diapycnal diffusion (Adcroft and Hallberg 2006). Isopycnal ocean models have the advantage of naturally including advection along potential density surfaces in the quasi-adiabatic ocean interior below the strongly mixed upper ocean. However, they are at a disadvantage at high latitudes where their vertical resolution declines because of the very small density difference between the surface and deep ocean.
The utility of $\theta$ coordinates for observational analyses was appreciated very early, by Rossby (1937) and Starr (1945). It was further developed by Johnson (1989), and by Hoskins et al. (1985), who emphasized the dynamical importance of the isentropic potential vorticity. Early atmospheric models based on $\theta$ coordinates were developed by Eliassen and Raustein (1968) and Bleck (1973). More recently, the merits of models based on $\theta$ coordinates have been discussed by Hsu and Arakawa (1990), Johnson (1997) and Benjamin et al. (2004), among others.

An issue with the use of $\theta$ coordinates in models is that $\theta$ surfaces intersect Earth’s surface. This has motivated numerous proposals for hybrid $\sigma-\theta$ coordinates (e.g., Johnson and Uccellini 1983; Zhu et al. 1992; Bleck and Benjamin 1993; Zapotocny et al. 1994; Konor and Arakawa 1997; Benjamin et al. 2004; Bleck et al. 2010).

An alternative is the arbitrary Lagrangian–Eulerian (ALE) approach (Bleck and Benjamin 1993; Bleck et al. 2010), which allows deviations from strict $\theta$ coordinates on the basis of a set of “rules.” For example, a rule might enforce a minimum pressure difference across a model layer (Toy and Randall 2009), or it might periodically “remap” the edges of quasi-Eulerian layers to prespecified target values of the $\sigma$ coordinate (Lin 2004). The ALE method allows for a mapping to an arbitrary vertical surface, such as geopotential, $\sigma$, potential temperature (or potential density), or even coordinates with no explicit mathematical definition.

Isopycnal ocean models were pioneered by Rainer Bleck at the University of Miami with the Miami Isopycnic Coordinate Ocean Model (MICOM), a well-used community layered model (Bleck and Boudra 1986; Sun and Bleck 2001). A version of MICOM has been developed by Helge Drange and colleagues at the University of Bergen, and is the ocean component of the Norwegian Earth System Model (Bentsen et al. 2013). Dunne et al. (2012) used a layered isopycnal model developed at GFDL for use in climate [General Ocean Layer Dynamics (GOLD)].

The quasi-Lagrangian approach used in isopycnal models provides a useful starting point for efforts to implement the ALE methods (Donea et al. 2004) in ocean models (Bleck 2002). The ALE method provides a natural framework for wetting and drying, such as for studies of coastal inundation and moving ice shelf grounding lines (Goldberg et al. 2012). Hybrid Coordinate Ocean Model (HYCOM) made use of ALE to blend an isopycnal coordinate in the deeper ocean with a depth coordinate in the strongly mixed upper ocean, with terrain-following coordinates along the shelves. HYCOM became the ocean component of the Goddard Institute for Space Sciences climate model (Sun and Bleck 2006).

The ALE approach is now spreading throughout the ocean modeling community to codes such as MPAS-O (Ringler et al. 2013) and MOM6 (Adcroft and Hallberg 2006). Similar methods are also becoming more fully realized in the atmospheric modeling community (Bleck et al. 2015, 2010; Sun et al. 2018).

d. Radiative transfer modeling in the 1990s: Unification

The errors evident in many radiation codes used in weather and climate models in the early 1990s prompted the development of new codes with a close link to reference models. Mlawer et al. (2016) describes the development of one such code: Rapid Radiative Transfer Model (RRTM) (Mlawer et al. 1997). RRTM implements a correlated $k$ distribution (Goody et al. 1989; Lacis and Oinas 1991; Fu and Liou 1992), an extension of the original $k$-distribution technique to vertically inhomogeneous atmospheres. The code was originally developed as an offline column model aimed at reproducing line-by-line calculations (themselves tightly constrained by a new wealth of observations, as Mlawer et al. (2016) describes), but soon included an offline with reduced spectral resolution (RRTMG) for use in atmospheric models.

The Met Office undertook a similar effort aimed at their new Unified Model (Cullen 1993), the first to be used to make both routine weather forecasts and climate projections. The Edwards–Slingo code (Edwards and Slingo 1996) stressed flexibility: the correlated $k$-distribution is specified at run time, allowing for different spectral resolutions and computational costs for different applications and for easy integration of new spectroscopic knowledge. Each $k$-distribution can be traced back to and assessed against a line-by-line calculation. Clouds are treated consistently across the longwave and shortwave spectrum; this includes treating scattering by clouds in longwave calculations. John Edwards attributes many of these design decisions to Tony Slingo:

Tony Slingo was the one who had the vision for it, and a couple of things that he wanted very strongly. One was, because it was to be used in climate, we really wanted to get the forcings right, and so having the ability to run at different spectral resolutions and have, as far as possible, traceability from precise comparisons with line-by-line, was seen as very important. Another thing was that we wanted the same cloud overlap assumptions in longwave and short-wave so we’d be doing cloud radiative effect consistently between the two spectral regions (interview with John Edwards, 20 October 2017).
As discussed in section 5d(1), Sommerria and Deardorff (1977) used higher-order closure with an assumed bivariate Gaussian distribution for (roughly speaking) temperature and moisture to determine the fractional cloudiness and liquid water mixing ratio. This approach can be called assumed distributions with higher-order closure (ADHOC). The intended application of Sommerria and Deardorff (1977) was large-eddy simulation, with grid cells less than 100 m across. Much later, Lewellen and Yoh (1993) suggested using a pair of joint Gaussians instead of one. This approach is more appropriate for larger grid cells that contain many clouds. In such larger grid cells, one of the Gaussians can represent the cloudy part of the domain, while the other represents the clear spaces between the clouds.

Randall (1987), Randall et al. (1992), and Lappen and Randall (2001) added vertical velocity to the mix, so that vertical fluxes of temperature and moisture could be computed from the parameters of the resulting trivariate distribution. Following the mass flux approach, they used a pair of delta functions, one representing turbulent updrafts and the other representing downdrafts. Finally, Golaz et al. (2002) combined the approaches of Lappen and Randall (2001) and Lewellen and Yoh (1993), resulting in a pair of trivariate Gaussians. This method has been used by Bogenschutz and Krueger (1993), resulting in a pair of trivariate Gaussians. Two-moment schemes were developed and applied in a few cloud-scale models in the 1980s (e.g., Ziegler 1985), but came into widespread use for cloud and mesoscale modeling in the mid-1990s through the 2000s (e.g., Schoenberg Ferrier 1994; Cohard and Pinty 2000; Seifert and Beheng 2001).

Starting in the 1990s, the development of aerosol representations for use in global climate models was motivated by a need to study the direct effects of aerosols on radiative forcing (e.g., Kiehl and Briegleb 1993; Taylor and Penner 1994; Mitchell et al. 1995a; Haywood et al. 1997). During this time, climate models also began to simulate the indirect effects of aerosols on radiation through their influence on clouds, by diagnostically relating droplet number to aerosol properties (e.g., Boucher and Lohmann 1995).

With the availability of fully coupled global land–atmosphere models and the widespread recognition of the problems of scale, a series of ambitious field experiments were undertaken to evaluate models by quantifying regional land–atmosphere interactions in nature. These included the First ISLSCP Field Experiment (FIFE) over the Kansas prairie (Hall and Sellers 1995), the Hydrologic Atmospheric Pilot Experiment (HAPEX) in the African Sahel (Prince et al. 1995), the Hydrologic Atmospheric Pilot Experiment (HAPEX) in the African Sahel (Prince et al. 1995), the Large-Scale Biosphere–Atmosphere Study (LBA) in Brazil (Keller et al. 2004). Each of these experiments involved simultaneous measurements of both atmospheric and surface conditions at a range of spatial scales from individual leaves and soil probes to regional footprints meant to represent entire GCM grid cells. The pioneering field experiments made extensive use of new satellite datasets and provided a huge resource for both testing models derived from local relationships and especially for learning how scales of land–atmosphere interaction worked in nature.

During the 1990s, many studies used coupled models to analyze land–atmosphere interactions in nature (e.g., Betts et al. 1996). In particular, comparisons of models and observations showed that soil moisture could act as a long-memory component in the climate system to amplify or extend the duration of droughts and rainy
periods (Oglesby and Erickson 1989; Lean and Rowntree 1993; Dirmeyer 1994; Milly and Dunne 1994; Brubaker and Entekhabi 1996; Diedhiou and Mahfouf 1996; Trenberth and Guillemot 1996; Eltaher 1998; Fennessy and Shukla 1999; Douville et al. 2001). Precipitation recycling of water through evapotranspiration was recognized as a major process at regional scales (Trenberth 1999). By this time, land–atmosphere coupling had also been adopted in numerical weather forecasting (Viterbo and Beljaars 1995). Interactive land–atmosphere models were used to analyze the role of the land surface in amplifying or extending the duration of droughts and rainy periods. Beljaars et al. (1996) used coupled models to analyze the effect of anomalies in soil moisture on persistent atmospheric circulation patterns associated with major drought and floods. They found that forecasts of the summer U.S. drought in 1988 and the Mississippi River floods in 1993 were dramatically improved when they initialized their coupled model with realistic soil moisture.

An innovative approach to the problem of subgrid-scale heterogeneity at the land surface was developed by Koster and Suarez (1992), in which many instances of the land parameterization are coupled to a single overlying atmosphere. The separate instances, or “tiles” have different properties such as assemblages of vegetation or soils, or may represent separate hydrologic catchments within a larger GCM grid cell. Separate calculations of prognostic soil temperature and moisture are done for each tile, and then the energy fluxes of each are weighted by their subgrid-scale fractional area before being passed to the atmospheric component. This approach has since been widely adopted to represent heterogeneity. Unlike the mesoscale flux experiments discussed above with limited area models (e.g., Pielke et al. 1991), tiling is tractable in global models because the computational expense of multiple instances of the land model is modest.

Plant physiologists worked with climate modelers to improve the biological realism of parameterized stomatal resistance. Rather than the simple empirical functions relating stomatal aperture to radiation, humidity, and temperature (Jarvis 1976; Dickinson et al. 1986), a new generation of models coupled stomatal function with photosynthesis. The new approach recognized that stomatal conductance solves an optimization problem in which plants evolved physiological mechanisms to maximize carbon gain under the constraint of minimizing water loss. Sellers et al. (1992) introduced the calculation of photosynthetic carbon assimilation using enzyme kinetic relationships previously studied in the laboratory (Farquhar et al. 1980). High rates of photosynthesis require highly conductive (open) stomates, which also allow transpiration. This simultaneously depletes CO$_2$ and enhances vapor pressure at the leaf surface, which feeds back on both photosynthesis and stomatal conductance (Ball 1988). An additional node was inserted between stomatal pores and the canopy air space to the resistance network in previous models. This laminar boundary layer at the leaf surface may be only a few millimeters thick, but maintains higher vapor pressure in immediate contact with stomatal pores and retards the upward flow of water vapor by turbulent exchange. Adding this extra resistance largely solved the coupling problem previously highlighted by Jarvis and McNaughton and allowed a greater degree of biophysical realism (Collatz et al. 1991). The models were iterated to solve simultaneously for the stomatal conductance and the rates of photosynthesis and transpiration.

Research continued into the critical problem of scaling physiological processes from stomates to grid cells. Sellers et al. (1992) showed that a simultaneous solution for canopy-scale transpiration could be obtained from leaf-level parameters by assuming that 1) the progressive downward attenuation of solar radiation through vegetation canopies followed an exponential decay with cumulative leaf area (Beer’s law), and that 2) plants have evolved to redistribute scarce resources (primarily nitrogen) according to the time-mean vertical distribution of light. These two assumptions allowed leaf-level equations for stomatal conductance and the rates of photosynthesis and transpiration to be integrated vertically in closed form.

Leaf area index is the area of leaves in a canopy per unit area of ground, and Sellers et al. (1992) used the cumulative leaf area index above a point in the canopy as a vertical coordinate. Integrating assimilation (photosynthesis) rate from the top of the canopy to the ground, and assuming that photosynthesis decreases exponentially along with light, they obtained an equation for the fraction of photosynthetically active radiation (FPAR) absorbed by the canopy. Importantly, the FPAR is related to the remotely sensed NDVI, which was mentioned in section 6d. Retrievals of NDVI from space allow global estimates of canopy-scale stomatal conductance and the rates of photosynthesis and transpiration based on leaf-level physiology and the FPAR relationship. Coupling of photosynthesis and transpiration with canopy integration from remote sensing was used to construct a new coupled GCM (Sellers et al. 1996a; Randall et al. 1996a), and a complete suite of satellite-derived parameters for land surface modeling (Sellers et al. 1996b; Los et al. 2000). Within a few years, many groups around the world also developed global land surface models based on integrated photosynthesis and transpiration, which were coupled to GCMs (Friend...

Although Sellers et al. (1992) intended the FPAR to represent the continuous attenuation of light in vegetation canopies, Bonan (1996) showed that this quantity can also be interpreted as the sunlit (as opposed to shaded) leaf area index. In this interpretation, only sunlit leaves are integrated in the scheme of Sellers et al. (1992), meaning that photosynthesis and transpiration are likely underestimated because of the presence of shaded leaves illuminated by diffuse radiation from the sky. Pury and Farquhar (1997) developed a simple scheme to separate plant canopies into sunlit and shaded fractions with different temperatures, stomatal conductance, and rates of photosynthesis and transpiration. Although the photosynthesis rate of shaded leaves is less than that of sunlit leaves, they use light more efficiently because diffuse light penetrates more deeply into dense canopies. This “two big leaf” approach has since been adopted by most land models (Dai et al. 2004).

g. Sea ice advances during the 1990s

Eventually sea ice modelers tried to simplify the methods used to predict the motion of the sea ice, in order to make them practical for climate models. First Flato and Hibler (1992) simplified the viscous-plastic dynamics by treating sea ice as a cavitating fluid, which lacks shear strength. Several modeling centers implemented cavitating-fluid dynamics and so became the first to simulate sea ice with a constitutive law. But most of the centers abandoned cavitating-fluid dynamics when better options became available. Next, Los Alamos National Laboratory scientists Elizabeth Hunke and John Dukowicz developed a numerical approximation to the viscous-plastic dynamics to simulate sea ice as an elastic-viscous-plastic material (Hunke and Dukowicz 1997), a method that asymptotes to the full viscous-plastic solution but is more efficient and highly parallelizable. In the same year, Zhang and Hibler (1997) made the viscous-plastic numerics more efficient and parallelizable. The latter two dynamics schemes made possible major improvements in simulating sea ice in climate models. The elastic-viscous-plastic approach is widely used among climate models today in part because the code was made readily available for sharing, with high-quality documentation and regular updates maintained by Hunke et al. (2010) in a comprehensive model known as the Los Alamos sea ice model (CICE).

8. Into the twenty-first century

Our story now approaches the present day, which means that much of the work is still ongoing, and we lack a historical perspective. Selected current issues are highlighted, but we do not attempt a comprehensive overview.

a. Current issues in atmospheric dynamical cores

1) HORIZONTAL GRIDS IN ATMOSPHERE AND OCEAN MODELS

Evolving computer architectures are now having a significant effect on preferred numerical methods. From the mid-1990s onward, the performance of computing machines has increased mainly through increased numbers of processors rather than faster processors. The communication of data between processors is relatively slow, and is becoming a significant bottleneck to computational performance for both the spectral method, which requires global communication for the spectral transforms, and for gridpoint methods on the longitude–latitude grid because of the polar resolution clustering. The trend toward massively parallel hardware has pushed the modeling world toward higher horizontal resolution. One consequence of these developments has been renewed interest in the use of quasi-uniform grids.

It is now conventional to distinguish between “structured” and “unstructured” horizontal grids. A structured horizontal grid covers the sphere with quadrilateral cells, so that each cell in the grid can be identified by a pair of indices, such as (i, j), and its neighbors can be specified by adding or subtracting 1 to i and/or j. Early structured grids made use of spherical latitude–longitude coordinates to tile the sphere. However, such grids suffer from singularities at the poles.

Unstructured grids are more flexible. They cover the sphere with simple shapes, such as triangles, squares or hexagons. In contrast to structured methods, the unstructured approach does not rely on a fixed number of gridcell neighbors, nor does it insist on local coordinate orthogonality. Although the spatial pattern of the cells may be very orderly, unstructured grids require a pre-stored list of the neighbors of each cell.

Since the 1960s and 1970s the importance of numerical properties such as conservation, monotonicity, accurate wave dispersion and balance, and avoidance of computational modes, has been much better appreciated, and a wide range of methods giving acceptable performance on unstructured quasi-uniform grids has been developed (e.g., Masuda and Ohnishi 1986; Heikes and Randall 1995). A related point is that the relative cost, in time and energy, of data movement in and out of memory as well as between processors, compared to computation, has greatly increased. Consequently, computationally intensive methods such as high-order
Galerkin methods are no longer seen as prohibitively expensive.

Starting with the German Weather Service’s GME icosahedral grid model (Majewski et al. 2002), this second wave of development on unstructured quasi-uniform grids has led to a number of production or production-capable models for both NWP and climate prediction (e.g., McGregor and Dix 2008; Satoh et al. 2008; Putman and Lin 2007; Qaddouri and Lee 2011; Skamarock et al. 2012; Dennis et al. 2012; Zängl et al. 2015; Sun et al. 2018). This appears to be a trend.

Today, most ocean models make use of structured grids in the horizontal according to either the Arakawa B grid or C grid (Arakawa and Lamb 1977; Griffies et al. 2000). With spherical coordinates, ocean models encounter a singularity at the North Pole, but not at the South Pole, which lies in the middle of the Antarctic continent. To remove the north pole singularity while retaining a structured grid, ocean modelers today use alternative coordinates while retaining local orthogonality. A common approach for is the tripolar grid of Murray (1996) and Madec et al. (1997), whereby the North Pole singularity is split into two singularities safely “hidden” over land. An alternative is to displace the North Pole over land as in the displaced pole approach used by POP simulations (Smith et al. 1995). More specialized uses have also been considered, such as Marsland et al. (2007) who used the MPI ocean model to study ice shelves in a global model.

Recent advances in the use of unstructured horizontal grids for ocean modeling have been based on both finite-volume (Ringler et al. 2013; Korn 2017) and finite-element (Danilov 2013) methods. The Model for Prediction Across Scales Ocean (MPAS-O) has been developed at LANL (Ringler et al. 2013), and uses the ALE vertical discretization described earlier, as well as an unstructured horizontal grid based on finite-volume methods. This model is targeted toward global ocean circulation applications as well as coupled climate modeling. The Finite Element Sea ice/Ocean Model (FESOM) was developed at the Alfred Wegener Institute, with particular applications focusing on high-latitude ocean domains and global ocean climate simulations (Danilov 2013). The greater flexibility of unstructured grids makes it possible to more faithfully represent the complex horizontal geometry of the World Ocean. It also offers an elegant means to nest fine-resolution subdomains within a coarser global grid. The drawback is that unstructured approaches can be more computationally expensive than structured approaches.

2) COUPLERS

As suggested in the discussion above, the atmosphere, ocean, and land surface components of an ESM are often implemented on grids that have different shapes and different resolutions. For this reason, ESMs include “couplers” (e.g., Craig et al. 2012) that are designed to allow the components to exchange information via interpolation (particularly from coarser to finer grids) or averaging (from finer to coarser). These exchanges are formulated so as to respect important physical principles such as conservation of mass and energy. It is possible that future very high-resolution models will not need couplers.

b. Current issues in radiative transfer parameterization

With the widespread availability of accurate parameterizations of clear-sky radiative transfer, attention turned to clouds, and particularly problems introduced by subgrid-scale variability. A range of observations (Cahalan et al. 1994; Pincus et al. 1999; Rosson et al. 2002) had shown that clouds are substantially inhomogeneous on the 10–100-km scales of the day’s global models, and Cahalan et al. (1994) had used simple calculations to argue that ignoring this variability inevitably biased reflectivity calculations, especially in the shortwave. Awareness that similar biases were likely influencing calculations of precipitation (Pincus and Klein 2000; Rotstyn 2000) motivated the development of cloud schemes that explicitly predicted internal variability (e.g., Tompkins 2002; Golaz et al. 2002). Further discussion is given in section 8c.

Various solutions to the problem have been proposed. Barker (1996) and Oreopoulos and Barker (1999) developed a closed-form solution to the two-stream equations integrated over a specific distribution of optical depth. Cairns et al. (2000) and Petty (2002) proposed rescaling the optical properties of the clouds based on a measure of their variability, an idea related to methods for treating radiative transfer in random media. A more flexible solution was to do independent calculations over optimally chosen elements of the cloud optical depth distribution (Collins 2001; Neu et al. 2007; Shonk and Hogan 2008).

An intercomparison effort was key to identifying the intertwined roles of cloud overlap and internal variability in causing errors in cloudy-sky radiation calculations. ICRCCM-III (Barker et al. 2003) reported domain-averaged fluxes from a range of three- and one-dimensional radiative models applied to a high-resolution description of clouds obtained from finescale models.

The intercomparison highlighted the weakness of the analytic treatments of cloud overlap that had been used since the 1960s, which introduce errors on par with those caused by neglecting variability. The paper describes errors as arising “mostly because of inappropriate
cloud overlap assumptions, incorrect application of overlap assumptions, neglect of horizontal variability of cloud, and inappropriate assumptions about horizontal variability.”

ICRCCM-III highlighted the need for flexibility in computing radiative transfer in cloudy skies: accuracy required that calculations be able to adapt to a wide range of overlap specifications as well as complicated descriptions of internal variability. Any practical new method had to meet these accuracy requirements without substantially increasing computational cost, which was already high enough that models typically computed radiation less frequently in time, and possibly at lower spatial resolution, than other physical processes.

The Monte Carlo Independent Column Approximation (McICA; see Pincus et al. 2003) uses a different, randomly generated discrete sample from the distribution of all possible cloud states with each spectral quadrature point, essentially replacing a two-dimensional integral over wavelength and cloud state with a Monte Carlo estimate. The fluxes computed with McICA are unbiased but, if the states used in each calculation (location, time, etc.) are chosen independently, the error introduced is also random. Extensive experience (e.g., Barker et al. 2008) demonstrated that this random error does not impact the simulation, and the technique has been widely adopted.

c. New approaches to representing cloud processes in global atmospheric models

1) GLOBAL CLOUD-RESOLVING MODELS

The continuing increase in computer power (Habata et al. 2003) has made possible the development of global atmospheric models with grid spacings of just a few kilometers, so that they can crudely but explicitly simulate individual large clouds (Tomita and Satoh 2004; Tomita et al. 2005; Satoh et al. 2008, 2014; Putman and Suarez 2011). For now, these “global cloud-resolving models” are too expensive for simulation of century-scale climate change, but at present they can be used for simulations of about one year. Global cloud-resolving models are qualitatively different from lower-resolution models because they do not need parameterizations of deep convection, but they still need parameterizations of radiation, microphysics, and turbulence including small clouds.

2) SUPERPARAMETERIZATION

In 1999, NCAR scientists Wojciech Grabowski and Piotr Smolarkiewicz created a multiscale GCM in which the physical processes associated with clouds were represented by implementing a simple cloud-resolving model (CRM) within each grid column of a low-resolution global model (Grabowski and Smolarkiewicz 1999; Grabowski 2001, 2004). With the embedded CRM, parameterizations of radiation, cloud microphysics, and turbulence (including small clouds) are still needed, but larger clouds and some mesoscale processes are explicitly (though crudely) simulated. The GCM simulates the large-scale weather, while the CRMs simulate the small-scale convective response, which is fed back to the GCM. Grabowski and Smolarkiewicz found that their model produced interesting simulations of organized tropical convection, including systems that resembled the Madden–Julian oscillation (MJO; Madden and Julian 1971, 1972).

Inspired by the results of Grabowski and Smolarkiewicz, Khairoutdinov and Randall (2001) created a multiscale version of the CAM (Collins et al. 2006a). They replaced most of the parameterizations used by CAM with a simplified version of Khairoutdinov’s CRM (Khairoutdinov and Randall 2003). Parameterizations of radiation, microphysics, and turbulence are included in the CRM. One copy of the CRM runs in each grid column of the CAM. The CRM is two-dimensional (one horizontal dimension, plus the vertical), and uses periodic lateral boundary conditions. In the study of Khairoutdinov and Randall (2001), the CRM had a horizontal domain 64 grid columns wide, with a horizontal grid spacing of 4 km. Because the CRM is two-dimensional, it cannot produce realistic vertical fluxes of horizontal momentum. For this reason, the momentum feedback to the GCM was not included.

Khairoutinov and Randall dubbed the embedded CRM a “super-parameterization.” The combination of a GCM with a superparameterization is now called a Multiscale Modeling Framework (MMF), and the MMF based on the CAM is now called the SP-CAM. Several additional MMFs have since been created, each based on a different GCM. In a major step, Stan et al. (2010) coupled the SP-CAM to a low-resolution version of POP. As reported by Stan et al. (2010), the coupled model gives a more realistic simulation of the atmospheric circulation than the SP-CAM “right out of the box,” without any tuning, a somewhat surprising result in view of the earlier experiences of others (e.g., Sausen et al. 1988). Superparameterized GCMs are much less computationally expensive than global cloud-resolving models.

SP-CAM and the other MMFs have produced interesting simulations of the MJO, the diurnal cycle of precipitation, the Asian and African monsoons, and other phenomena, including anthropogenic climate change. Further discussion is given by Randall et al. (2016).
Super-parameterization has also been tested in an ocean model (Campin et al. 2011) to simulate the small but important regions where deep convection occurs. Though promising, the superparameterization technique has up to now been used less for the ocean than for the atmosphere.

3) CLOUD MICROPHYSICS AND AEROSOLS

With increased computing power and the related trend toward finer model resolution, more detailed representations of microphysics, including two-moment schemes, have recently been adopted for operational numerical weather prediction. Examples include the two-moment Milbrandt–Yau scheme in the High Resolution Deterministic Prediction System in Canada (Milbrandt et al. 2016), and the aerosol-aware Thompson scheme (Thompson and Eidhammer 2014) in the U.S. Rapid Refresh (RAP) and High Resolution Rapid Refresh (HRRR) models. A diagram of a typical two-moment, multi-ice-class scheme is shown in Fig. 12-5b. The recent development and operational use of high-resolution, convection-permitting kilometer-scale forecast models has in particular motivated the use of more sophisticated microphysics schemes, since convective and cloud scale motions are more directly coupled to the microphysics. Over the last decade, schemes have also been developed that have moved away from the traditional paradigm of using fixed categories representing different types of ice (e.g., Hashino and Tripoli 2007; Harrington et al. 2013; Morrison and Milbrandt 2015). These schemes evolve ice properties smoothly by predicting characteristics such as particle aspect ratio and density, and avoid some of the difficulties that arise with fixed ice categories. Although bin schemes are still too computationally expensive for operational modeling, they have been used for process studies of topics such as cloud-aerosol interactions (e.g., Feingold et al. 1996; Fridlind et al. 2004; Khain et al. 2005) and microphysical–dynamical interactions (e.g., Stevens et al. 1996; Ackerman et al. 2004). They have also been used to develop and test bulk-microphysics schemes for weather and climate models (e.g., Khairoutdinov and Kogan 2000).

Two-moment schemes have also been developed for climate models, but motivated more by a need to physically treat clouds and radiation by predicting cloud droplet number, and links to aerosols (e.g., Ghan et al. 1997; Lohmann et al. 1999; Ghan et al. 2001; Ming et al. 2007; Lohmann et al. 2007), which have important radiative as well as microphysical effects (Boucher et al. 2013). The models use parameterizations of cloud condensation nuclei activation, coupled with prognostic multispecies aerosol chemistry and transport schemes (e.g., Stier et al. 2005; Seland et al. 2008; Liu et al. 2012). Over the last decade, some climate models have also incorporated the effects on clouds of ice-nucleating aerosols (e.g., Lohmann and Hoose 2009; DeMott et al. 2010), and including mixed phase clouds (Hoosie et al. 2010; Gettelman and Morrison 2015). Current state-of-the-art ice nucleation parameterizations (e.g., Hoosie et al. 2010) can directly incorporate laboratory and field measurements of ice nucleating particles, but there are still large uncertainties in ice nucleation properties.

With higher resolution and increased computational resources, the microphysics schemes used in climate models now incorporate many features previously used in mesoscale models, including prognostic two-moment precipitation (Possett and Lohmann 2008; Gettelman and Morrison 2015). A critical issue when incorporating such schemes in larger-scale models is that the cloud-scale and mesoscale motions driving the microphysics are not resolved, and thus the microphysics must be coupled with “macrophysics” parameterizations of the driving dynamic and thermodynamic processes. A related issue is that subgrid-scale variability of cloud quantities, typically neglected in small-scale models, is critical in larger-scale models because microphysical process rates often depend nonlinearly on predicted cloud quantities (e.g., Pincus and Klein 2000; Larson et al. 2001; Rotstyn 2000). This has been dealt with in global climate models by ad hoc tuning of process rates (e.g., Golaz et al. 2011), or by integrating them over an assumed subgrid-scale distribution of cloud water amount in each grid cell (Morrison and Gettelman 2008).

An important advance over the past decade has been the development of Lagrangian particle-based microphysics schemes in which the multitude of cloud and precipitation hydrometeors are represented by a collection of “super-particles” that evolve as they are transported by the modeled flow (e.g., Shima et al. 2009; Andrejczuk et al. 2010; Unterstrasser and Söhl 2010). Unlike bin (and bulk) schemes that employ continuous-medium, Eulerian microphysical variables, Lagrangian-based schemes do not suffer from numerical diffusion errors.

Up to now, most of the work on microphysics parameterization has been focused on stratiform clouds. The treatment of microphysics in convection parameterizations has generally remained very simple and crude, despite the facts that cumulus clouds generate a large fraction of Earth’s precipitation, and detrainment from cumulus updrafts produces many radiatively important stratiform clouds. All of these important effects of cumulus clouds are influenced by microphysical
processes at work within the cumuli. Kerry Emanuel (1991) forcefully argued that more realistic microphysics is needed in cumulus parameterizations. There has been some recent progress in this area (Song and Zhang 2011; Elsaesser et al. 2017; Zhao et al. 2016). In addition, attempts have been made to unify cloud microphysics across cloud schemes with unified closures in climate models treating all clouds with the same microphysics (Bogenschutz et al. 2013). Overall, the ongoing convergence of models spanning scales from weather to climate requires detailed, yet efficient cloud microphysics schemes linked consistently to the parameterized turbulence, convection, and radiation.

d. Current issues in ocean modeling

1) OCEAN MODEL INTERCOMPARISONS

The integrity of climate models depends on the integrity of the physical parameterizations in the ocean component. Early coupled ocean–atmosphere models drifted away from an Earth-like climate because of inaccuracies in the representation and parameterization of physical processes in the models. An important milestone was reached by Boville and Gent (1998) and Gordon et al. (2000), whereby their use of improved ocean physical parameterizations was shown to enable a much more stable simulation without the use of “flux adjustments.” Gent (2013) offers an overview of climate modeling and the role of the ocean and ocean physical parameterizations. Such results lend support for the need to study the ocean and sea ice components of climate models separately from the fully coupled AOGCMs. For that purpose, the community has developed the Ocean Model Intercomparison Project (OMIP) started in the late 1990s. It took nearly 20 years to develop a suitable protocol and to improve model integrity sufficiently to support the OMIP exercise (Griffies et al. 2016). Such comparison projects have been the foundation for ongoing model improvements throughout much of the history of ocean modeling, and will remain so into the future.

2) MESOSCALE EDDIES AND PARAMETERIZATIONS OF ISOPYCNAL DIFFUSION

Realizing the importance of mesoscale eddies for ocean dynamics and the transport of heat, carbon, and other tracers, oceanographers became rather critical of numerical simulations that had no representation of these eddies. As with synoptic eddies in the atmosphere, ocean mesoscale eddies have scales largely determined by the first baroclinic Rossby radius due to their connection to baroclinic instability. Ocean mesoscale eddies range in size from 100 km in the tropics to less than 10 km near the poles and on continental shelves. One response to this situation was to focus on quasigeostrophic models, whose simpler dynamical equations and lack of thermodynamics allowed for the explicit representation of transient eddy features (Holland 1978). Another response was to tackle the problem of eddy parameterization while continuing to improve primitive equation models. Although much progress has been made since the 1970s, the eddy parameterization problem remains at the forefront of ocean theory and modeling to this day.

A conceptual framework for how mesoscale eddies act on the large-scale tracer field arose during the 1970s to 1990s. This framework arose largely from field measurements of transient radioactive ocean tracers as well as through atmospheric insights into transport from synoptic atmospheric eddies. The two key pieces to the framework are eddy-induced diffusion along neutral surfaces and eddy-induced stirring of density in a manner that reduces available potential energy. See the book by Griffies (2004) for a pedagogical treatment.

Neutral diffusion (more commonly known as isopycnal diffusion) was proposed by Solomon (1971) and Redi (1982). The neutral diffusion operator respected growing observational evidence (Veronis 1975) that tracers are stirred by mesoscale eddies along neutral directions rather than along constant geopotential surfaces (see McDougall 1987; McDougall et al. 2014, for discussion of neutral directions). Cox (1987) offered a numerical implementation of isopycnal diffusion, and Griffies et al. (1998) updated the Cox scheme to remove some pernicious numerical instabilities.

The eddy-induced tracer transport was proposed by Gent and McWilliams (1990). As per the energetic impacts from mesoscale eddies, the eddy-induced velocity is designed to dissipate available potential energy (Gent et al. 1995; Griffies 1998). Complementary ideas from Greatbatch and Lamb (1990) noted the equivalence, for geostrophic flows, of vertical momentum form drag realized through vertical viscosity.

As shown by Danabasoglu et al. (1994), the combined use of neutral diffusion and eddy-induced stirring remedied a huge suite of model biases that had plagued the ocean GCMs of the time. Evolved versions of these two parameterizations are still in use today in all ocean GCMs that do not explicitly resolve transient mesoscale eddy features. Neutral diffusion is simpler to implement in isopycnal models than the rotated neutral diffusion of geopotential models, but precludes the representation of water mass transformation by thermobaricity.

Models that use quasi-Lagrangian methods [see section 7c(2)] to ensure that model coordinate surfaces are
isopycnal surfaces can directly represent neutral diffusion without the need for special numerical methods.

3) DIAPYCNAL MIXING WITHIN THE OCEAN INTERIOR AND BOUNDARY LAYERS

Much of the ocean interior is a quasi-ideal fluid in that there is very little irreversible mixing between isopycnals. In contrast, mixing is vigorous in the mixed layer of the upper ocean, as well as in “benthic” boundary layers next to the ocean bottom. Mixing between interior ocean isopycnals (i.e., diapycnal mixing) affects stratification, ventilation, and the time scales for dynamical processes such as waves. Hence, this mixing has a very large impact on ocean circulation. The sensitivity of ocean circulation models to the levels of diapycnal mixing were emphasized by the watermass study of Bryan and Lewis (1979). They prescribed an enhanced diffusivity at depth to account for increased mixing in deep ocean regions of low stratification. This Bryan–Lewis diffusivity profile became the norm for ocean circulation models for the next 20 years, because it greatly improved the realism of the simulations, particularly those where deep flows appear such as in the Southern Ocean. This sensitivity of ocean circulation to diapycnal mixing has also been emphasized by the work of Walter Munk (1966) and Frank Bryan (1987).

These two modeling studies pointed to the need for additional field measurements and process modeling to enable an understanding of the fundamental nature of interior ocean mixing. This work was recently reviewed by MacKinnon et al. (2013), who brought together ideas of interior mixing and summarized its connection to breaking internal gravity waves. Further, this study offers an example of how efforts at large-scale models, process-models, theory, and observations can be synergistically combined to render deeper understanding of how the ocean works.

The ocean is strongly forced at its surface through air-sea and ice-sea interactions, and at the bottom through interactions with the solid Earth. This forcing drives intense three-dimensional turbulent mixing with order unity vertical-to-horizontal aspect ratios (i.e., non-hydrostatic dynamics). It therefore must be parameterized in hydrostatic ocean models.

In ocean circulation models of the 1980s, the surface boundary layer was “parameterized” by using a top layer of order 50 m thick. However, as models refined their vertical grid spacing, the needs for more physically based schemes became apparent. In response to this need, Large et al. (1994) provided a review of the extant methods (e.g., bulk boundary layers and second-order turbulence closures). They proposed a new approach based on ideas that had been developed for atmospheric boundary layer parameterizations (Troen and Mahrt 1986; Holtslag et al. 1990; Holtslag and Boville 1993). Their K-profile parameterization (KPP) has been incorporated into many ocean climate models. Alternative methods based on energetic approaches have also provided the framework for boundary layer closure (e.g., Gaspar et al. 1990), particularly those used by the NEMO community. Such energetic approaches have also been traditionally used by isopycnal models (Hallberg 2003).

Much of the deep waters around Antarctica originate from overflows off the continental shelves. Similar processes occur in the Denmark Strait and Faroe Bank Channel regions of the North Atlantic. Faithfully incorporating such processes in ocean climate models is a combination of model frameworks (e.g., vertical coordinates) and parameterizations. The traditional geopotential vertical coordinate is ill suited to representing these processes because of high levels of spurious mixing, whereas isopycnal and terrain-following models are far better suited (Legg et al. 2006). Legg et al. (2009) summarized the results from a climate process team of global circulation modelers, theorists, process-physicists, and observationalists who focused on this overflow problem and offered recommendations for improving the climate-scale models.

e. Current issues in sea ice modeling

With satisfactory methods to solve sea ice dynamics, attention turned to improving the thermodynamics by implementing an ice-thickness distribution and brine-packet physics, first in the University of Victoria Climate Model (Bitz and Lipscomb 1999; Bitz et al. 2001), soon after in version 2 of the Community Earth System Model (CCSM2; Bitz et al. 2005; Holland et al. 2006) and in version 2 of the GFDL Climate Model (CM2; Winton 2000; Gnanadesikan et al. 2006), and now in the majority of models. Detailed melt pond parameterizations and radiative transfer that includes scattering (important for sea ice because of brine inclusions and air bubbles) are in many models now too (e.g., Briegleb and Light 2007; Flocco and Feltham 2007; Holland et al. 2012). A desire to simulate brine pocket dynamics more faithfully, with prognostic salinity and sea ice biogeochemistry, has motivated practical schemes to better approximate mushy-layer physics (e.g., Vancoppenolle et al. 2009; Turner et al. 2013).

Figure 12-11 is a schematic illustration of the grid cell of a state-of-the-art sea ice model with a thickness distribution. Each thickness category has a unique snow depth, melt pond depth and coverage, heat fluxes

11 Frank Bryan is not related to Kirk Bryan.
at the top and bottom, and vertical profile of temperature and salinity. A fraction of the grid cell may be open water. Models that do not parameterize the thickness distribution effectively have just one thickness category.

In weather forecast systems, the initial sea ice concentration is usually specified based on passive microwave satellite retrievals and held fixed throughout the forecast (e.g., Grumbine 2013). Other aspects of sea ice thermodynamics are usually rudimentary. Until recently, sea ice has been specified in medium-range and seasonal forecast models as well. However, with version 2 of the National Centers for Environmental Prediction Climate Forecast System (CFSv2; Saha et al. 2010), the GFDL sea ice component of CM2 was implemented. In 2017, ECMWF transitioned their subseasonal operational forecast model from fixed to active sea ice by adopting version 2 of the Louvain-la-Neuve Sea Ice Model (LIM; Fichefet and Morales-Maqueda 1997, F. Vitart 2017, personal communication, and see https://www.ECMWF.int/en/research/modelling-and-prediction/ marine).

f. Current issues in land surface modeling

Turbulent fluxes of latent and sensible heat and momentum can be estimated from high-frequency measurements of wind speed and scalars through a technique called eddy covariance, pioneered in the 1950s (Swinbank 1951). These fluxes are examples of the second moments considered in boundary layer parameterizations based on higher-order closure, as discussed in section 5d(1). The development of relatively inexpensive sonic anemometers and fast-response sensors led to rapid expansion in the use of eddy covariance in the 1990s. The application of the technique to measure the carbon balance of ecosystems has led to the creation of a worldwide network of many hundreds of semipermanent eddy covariance towers that monitor turbulent fluxes over land surfaces (Baldocchi et al. 2001; Baldocchi 2003). The availability of hourly estimates of turbulent fluxes of heat, moisture, and CO2 over all kinds of surfaces in all kinds of weather has been incredibly valuable for the development and maturation of land surface modeling (Friend et al. 2007; Stöckli et al. 2008).

The use of satellite imagery to prescribe vegetation and soil parameters made land models more realistic in the 1990s, but also limited their usefulness for prediction, since satellite imagery will never be available for the future. Two major developments in the 2000s intended to address this were prognostic phenology and dynamic global vegetation models (DGVMs). Rather than using satellite vegetation data as model input, land models seek to predict both seasonal and longer-term changes in vegetation properties. Observations of vegetation properties from satellites and other sources are then used to evaluate model output.

“Phenology” refers to the seasonal growth and shedding of leaves in response to changing environmental conditions. Models with prognostic phenology “grow their own leaves.” These models are based on empirical relationships between the timing of leaf activity and day length, temperature, and moisture (White et al. 1997; Lawrence and Slingo 2004; Arora and Boer 2005; Gienapp et al. 2005; Jolly et al. 2005; Stöckli et al. 2008a; Dickinson et al. 2008). The availability of global satellite coverage and hundreds of hourly flux tower records greatly accelerated the development of skillful parameterizations of phenology (Zhang et al. 2003; Reed et al. 1994; Gibelin et al. 2006; Bradley et al. 2007; Kathuroju et al. 2007; Stöckli et al. 2011).

Dynamic global vegetation models seek to predict not just the seasonal greening and browning of the land surface, but changes in long-term distribution of vegetation in response to climate. These models are important for century- and longer-scale climate simulation, allowing for feedback between the physical climate and the geographic patterns of surface properties (Cramer et al. 2001; Sitch et al. 2008). These models incorporate well-established physical and biological algorithms of earlier land surface parameterizations, but add algorithms for plant establishment, mortality, and competition for light and water (Cox 2001; Bonan et al. 2003; Woodward and Lomas 2004; Gerten et al. 2004; Krinner et al. 2005; Sitch et al. 2005; Lucht et al. 2006). Introduction of a new feedback process such as vegetation-climate interaction can result in large perturbations,
which may not be realistic. An early DGVM result was catastrophic dieback of the Amazon rain forest in one such coupled model (Cox et al. 2000, 2004), which released large amounts of CO₂ to the atmosphere and accelerated global warming. This result may reflect excessive drought stress in the hydrologic model component (Baker et al. 2008; Harper et al. 2014) than with a realistic assessment of carbon cycle instability (Cox et al. 2013).

Having linked land–atmosphere exchanges of energy and water with photosynthesis in the 1990s and then incorporated prognostic phenology and dynamic vegetation in the 2000s, the coupled models were then used to analyze sources and sinks of atmospheric CO₂. It has long been known that terrestrial ecosystems currently sequester about half of global fossil fuel emissions because of an excess of photosynthesis over decomposition (Tans et al. 1990; Le Quéré et al. 2009). Increased atmospheric CO₂ directly induces enhanced rates of photosynthesis (Norby et al. 2005; Luo et al. 2006), but nutrient limitation typically restricts growth (Oren et al. 2001; LeBauer and Treseder 2008; Thornton et al. 2007, 2009). Land carbon sinks also result from changes in land management and the age structure of forests (Shevlakova et al. 2009; Pan et al. 2011). Each of these carbon sink processes is likely to change in the future: some (e.g., CO₂ fertilization) are likely to get stronger while others (e.g., regrowing forests) are likely to get weaker.

A systematic intercomparison of coupled carbon–climate models was undertaken by Friedlingstein et al. (2006). They ran 11 Earth system models from 1850 to 2100, prescribing fossil fuel emissions, allowed ocean and land sinks to interact, and predicted both CO₂ and climate change. Their results showed a striking divergence in CO₂ and climate in the twenty-first century. Most simulations developed stronger and stronger land carbon sinks driven primarily by CO₂ fertilization, but the effect was highly variable across participating models. But several simulations showed sharply decreased land carbon uptake or even the release of hundreds of gigaton of land carbon as atmospheric CO₂ as death and decomposition overtook photosynthesis. This highly uncertain carbon–climate feedback (Dufresne et al. 2002; Friedlingstein et al. 2003) was shown to produce uncertainty of over 250 ppm in simulated CO₂ at the end of the runs (Friedlingstein et al. 2006), given identical fossil fuel emissions, with a resulting spread of about 1.5 K of global warming. A quantitative analysis of Earth system climate feedback showed that carbon–climate feedback is among the most uncertain, rivaling uncertainty in cloud feedbacks (Gregory et al. 2009).

9. The future

a. Increasing resolution

GCMs have always used the fastest computers available. Up to about the year 2000, the “clock speeds” of computer processors steadily increased. A faster clock means that a given arithmetic operation (e.g., addition or multiplication) can be performed in a shorter time. Faster clocks thus allowed longer simulations with the same model, or simulations of a given length with more “expensive” models, that is, models that use higher spatial resolution or more computationally demanding physical parameterizations.

The increase in clock speeds came to an end in large part because faster clocks demand increasing amounts of expensive electrical power; the cost has simply become unsupportable. Since about 2000, the supercomputers used to run ESMs have increased in performance largely through the use of increasing numbers of processors running in parallel. The most straightforward way for a modeling center to use more processors is to increase the horizontal resolution of the model. Unfortunately, however, having 4 times as many processors does not enable simulations of a given length with 4 times as many grid points, because the time step of the model will have to decrease at higher resolution. There are various practical difficulties of this type.

Increasing resolution brings a different issue. As a model’s grid cells become smaller, the character of the unresolved physical processes changes. For example, with grid cells 100 km across, an atmospheric model needs a parameterization that represents the gridcell-averaged heating and drying (and other processes) associated with deep cumulus convection, including vertical transports by strong but unresolved convective updrafts and downdrafts. This is because multiple deep cumulus clouds can fit in a grid cell 100 km wide. In stark contrast, a grid cell 1 km across can actually fit inside a deep cumulus cloud; a model with a horizontal grid spacing of 1 km can explicitly (if crudely) simulate the deep convective clouds, so that parameterizing them is inappropriate (and unnecessary), although of course parameterizations of microphysical processes, radiation, and small-scale turbulence are still needed.

There is an intermediate range of grid spacings, on the order of 10 km, that is too coarse to allow explicit simulation of deep cumulus clouds, but too fine to permit an accurate statistical representation of such clouds. This troublesome range of grid spacings, used today in some global models, is often called the “gray zone.” An analogous gray zone can be defined with respect to the turbulent eddies of the boundary layer, which are an order of magnitude smaller than deep cumulus clouds.
In fact, because the atmosphere and ocean contain eddies on all scales larger than a millimeter or so, a gray zone can be defined for any practical choice of horizontal resolution.

The gray zone for deep cumulus convection is thought to be particularly important, however. Many of today’s models have grid spacings that are in or approaching the gray zone for deep convection. Ongoing research aims to create resolution-independent parameterizations that can work for a wide range of horizontal grid spacings, including those that fall within the gray zone (e.g., Arakawa and Wu 2013). This will allow a single code, based on a single set of equations, to be used with a wide range of horizontal grid spacings—a very practical and convenient modeling system.

b. The future of atmospheric dynamical cores

The ongoing increase in horizontal resolution, mentioned in the preceding section, has motivated the development of “nonhydrostatic” dynamical cores for global models, which do not use the quasi-static approximation. Some current research is aimed at evaluating the relative merits of using the “fully compressible” system of equations, which allows vertically propagating sound waves, versus alternative systems that filter such waves (e.g., Arakawa and Konor 2009).

Because its cost grows faster than the number of degrees of freedom, and because of issues such as “spectral ringing” in the presence of sharp gradients, the imminent demise of the spectral method has been predicted for several decades! The communication burden of the spectral transforms on massively parallel machines may be the final nail in the coffin.

Semi-Lagrangian advection schemes are complex both algorithmically and in terms of their communication patterns. At the same time, their advantage in being able to take large time steps is less important on quasi-uniform grids. We may see a move away from semi-Lagrangian schemes in the future.

Finally, semi-implicit integration schemes require the solution of global elliptic problems, which are perceived to be difficult to solve efficiently on massively parallel machines. Consequently, new nonhydrostatic model developments aimed at massively parallel machines have tended to time splitting or vertically implicit integration schemes (Satoh et al. 2008; Skamarock et al. 2012; Zängl et al. 2015), though some attempts have been made to demonstrate the feasibility and competitiveness of parallel elliptic solvers (Heikes et al. 2013; Sandbach et al. 2015). There is now a vast number and variety of numerical methods for atmospheric modeling under consideration by the research community. A range of quasi-uniform grids is being explored, the most popular being cubed spheres, triangular and hexagonal icosahedra, and the overset yin–yang grid (Fig. 12-12). Spatial discretizations include finite-difference methods, finite-volume methods, and a variety of finite-element methods, which are analogous to spectral methods but use local (rather than global) basis functions. These are coupled with a range of explicit, implicit, subcycling, and Riemann-solver-based time integration schemes.

Current work is exploring some approaches that have the potential for a major impact on the field, if they can be made to work well enough. Grids with geographically variable (but temporally fixed) resolution are being tested (e.g., Rauscher and Ringler 2014; Zarzycki and Jablonowski 2015). An idea that has great potential to improve the computational efficiency of weather and climate simulations is to use a grid that dynamically adapts to the solution, placing the highest resolution where it is most needed. Alternative approaches are moving the grid while retaining the grid topology, or

FIG. 12-12. Three examples of quasi-uniform spherical grids. (left) Cubed sphere; (middle) hexagonal–icosahedral grid; (right) yin–yang grid. In practice the resolutions used would be much finer than shown here.
inserting and removing grid points where needed. Experiments with both approaches appeared in the 1990s (e.g., Dietachmayer and Droegemeier 1992; Skamarock and Klemp 1993). Adaptive vertical grids are also being investigated (Marchand and Ackerman 2011; Yamaguchi et al. 2017). However, there are significant challenges both in defining suitable criteria for where to refine the grid and in maintaining conservation and balance and avoiding noise as the grid adapts. The only operational adaptive grid forecast model to date appears to be that of Bacon et al. (2000). Some of the challenges mentioned above are being addressed (e.g., St-Cyr et al. 008; Dubos and Kevlahan 2013; Bauer et al. 2014; see also the 28 November 2009 issue of Philos. Trans. Roy. Soc.). Also, with the evolution of supercomputers toward ever greater numbers of processors, a significant challenge is to devise algorithms with sufficient parallelism to take advantage. This has led to some exploration of parallel-in-time algorithms (e.g., Haut and Wingate 2014).

The future evolution of computer architecture (which itself is uncertain) is likely to continue to influence the development of numerical methods. Efforts are currently under way to test the feasibility of running global atmospheric models on machines that include graphics processing units (GPUs) to achieve greater speed (e.g., Leutwyler et al. 2016; Abdi et al. 2017).

c. The future of radiation parameterizations

Radiation is unique as a parameterization problem for atmospheric modeling because fundamental understanding of the problem is so complete. For this reason, the parameterization of radiative processes focuses on how to use incomplete information from a model to compute fluxes of sufficient accuracy with acceptable computational cost. Future research will likely be focused on strategies for mitigating computational cost and increasing accuracy and accounting for the horizontal transport of radiation.

As described in section 6c, the high computational cost of spectrally integrated calculations means that radiative fluxes are typically computed more sparsely in time than any other subgrid-scale diabatic processes, potentially degrading simulations by blurring the coupling between fast-changing clouds and radiative fluxes. One promising approach is to devote specific computational resources to computing radiative fluxes (e.g., Balaji et al. 2016), allowing more frequent radiation computations and speeding time to solution at the cost of using more resources overall. Because radiation calculations integrate over a spectral dimension the problem is well suited to exploit heterogenous computing environments. Highly parallel processors such as GPUs in particular offer tantalizing hints of very high efficiency (e.g., Price et al. 2013; Clement et al. 2018).

New frontiers for accuracy include better coupling of radiation among the atmospheric, oceanic, cryospheric, and terrestrial components of Earth system models and steps to relax the strong one-dimensional plane-parallel assumption. In all ESMs of which we are aware, radiative fluxes are computed independently in each domain, that is, in the atmosphere, ocean, and sea ice, using independent models that are nonetheless based on the same underlying equations (e.g., Yuan et al. 2017). Results from each domain serve as boundary conditions for the other domains; the cost of coupling components often requires that spectral resolution is degraded at the potential cost of accuracy. A more natural and potentially more accurate approach would be to compute radiative fluxes in the atmosphere and ocean simultaneously (e.g., Lee and Liou 2007); extending this approach to sea ice, whose albedo can vary dramatically, might improve prediction in the Arctic. Problems arising in the computation of radiation in heterogenous vegetation canopies (e.g., Yuan et al. 2017) have much in common with similar efforts in clouds, suggesting that progress might come from the two communities working more closely together (Hogan et al. 2018).

Despite the manifest three-dimensionality of the atmosphere, essentially all parameterizations of radiative transfer used in global models adopt plane-parallel geometry and make use of the assumption that all radiation travels straight up and down. Emerging new techniques (Schäfer et al. 2016; Hogan et al. 2016) relax the one-dimensional assumption, accounting parametrically for effects such as the casting of cloud shadows, the illumination of cloud sides, and the increased cooling from cloud edges (Hogan et al. 2016) within each column. These effects are small but systematic: finite clouds uniformly increase surface and top-of-atmosphere fluxes relative to their plane-parallel counterparts while impacts of solar illumination vary with solar zenith angle, and hence latitudinally and seasonally. As parametric treatments are evaluated more rigorously efforts to include these effects in coarse-resolution models may become more common.

d. The future of cloud and microphysics parameterizations

Parameterizing microphysics remains highly challenging because of the complexity of the underlying physics and a lack of fundamental knowledge on these processes, especially for ice microphysics. This is a critical challenge for weather and climate modeling because simulations are often quite sensitive to microphysical parameter settings, and the increasing complexity of
schemes has not changed this picture. Overall, continued advancement of parameterizations will require greater knowledge of the underlying physical processes in order to reduce parameter uncertainty, including from laboratory studies, cloud observations, and detailed process modeling. Representing subgrid-scale cloud processes consistently across all model scales continues to be another major challenge despite increasing model resolution. Efforts have been made to develop subgrid representations of clouds and dynamics to consistently drive cloud microphysics across a range of scales and cloud types (e.g., Thayer-Calder et al. (2015)). These "unified" cloud parameterization efforts will likely be an important part of weather and climate model development in the coming years.

New approaches to superparameterization are also under development. For example, (Parishani et al. 2017) report encouraging results with an "ultra-parameterization" in which the horizontal grid spacing of the embedded cloud-resolving models is reduced to 250 m, and the vertical resolution is also increased, so that the eddies associated with shallow clouds can be explicitly simulated. Jung and Arakawa (2014) have developed a "quasi-three-dimensional" (Q3D) superparameterization, in which the CRMs take the form of narrow channels that form closed loops on the global model’s grid, for example, around meridians or latitude circles. The channels cross but do not intersect; they communicate only through the host GCM. With the Q3D approach, it is possible to include realistic topography (Jung 2016) including orographically enhanced precipitation, as well as vertical momentum transport by both convection and gravity waves, as explicitly simulated on the CRM grids.

Meanwhile, efforts are under way to use machine learning to create accurate and computationally efficient parameterizations (Chevallier et al. 1998; Brenowitz and Bretherton 2018; Gentine et al. 2018; Schneider et al. 2017). It seems likely that this approach can lead to improved simulations with tolerable computational cost, at least for the current climate. Can it also be used to simulate different climate states? Can it also be used to learn more about the actual physical mechanisms through which the cloud systems interact with larger-scale motions? Work is needed to address these questions.

e. The future of ocean models

Since the 1970s, much of the focus of global ocean circulation modeling has been at understanding, representing, and parameterizing the impacts from mesoscale eddies. This focus remains a large part of today’s efforts. For example, prototype centennial-scale climate simulations have been run with a vigorous eddy field. In particular, Griffies et al. (2015) emphasize the role of mesoscale eddies in the vertical transport of heat in the ocean, thus directly impacting on the rate of transient climate change. Small et al. (2014) emphasize the role of small-scale ocean features in forcing the atmosphere circulation through the surface fluxes. However, new avenues of research are focused on the submesoscales, which are intermediate between the balanced motions at the mesoscale and unbalanced motions at the gravity wave scale (Fox-Kemper et al. 2008; Thomas et al. 2008; McWilliams 2016). Submesoscale processes impact the vertical transfer of properties in the upper ocean, and mediate the downscale cascade of energy and tracer variance to the small scales. In parallel, modelers are increasingly pushing the frontiers of coastal and shelf processes within the global climate models by grid refinement or nesting approaches. It is here that impacts from the changing climate will have their largest footprint on civilization because of changes in ecosystems and sea level.

We expect that numerical models of the ocean will continue to improve through advances in numerical methods and physical parameterizations, including many of the approaches outlined here (e.g., ALE for the vertical and unstructured meshes for the horizontal). Improvements to observational datasets will also be necessary to evaluate the simulations. The history of ocean modeling has not been linear, with examples of advances in one subfield spawning new understanding and development in unexpected areas. Nonetheless, ongoing advances in ocean models and modeling practices, along with new theoretical insights, will ensure that numerical models remain a fundamental component of oceanography and climate science into the future.

f. The future of sea ice models

The next developments for sea ice are likely to be more realistic sea ice dynamics that replicate the effects of anisotropy on lead formation (e.g., Sulsky et al. 2007; Tsamados et al. 2013) and joint thickness-floe distributions models (Horvat and Tziperman 2015; Roach et al. 2018)—the latter permitting better representation of the region near the sea ice edge where ocean surface waves interact with floes and floe size influences ice–albedo feedback.

g. The future of land surface models

As more processes are added to Earth system models, there is more room for unexpected interactions. Just as the coupling of ocean and atmosphere GCMs produced nonphysical climate drift that required flux corrections (e.g., Cubasch et al. 1992), fully coupled land–atmosphere models produced highly uncertain
carbon–climate feedback (Friedlingstein et al. 2006). In response to the large spread in Earth system model outcomes, the land modeling community has embarked on a series of systematic model intercomparisons, evaluations, and benchmarking exercises using a wide range of global datasets (Luo et al. 2012; Huntzinger et al. 2012).

Land–atmosphere coupling in the CMIP5 ensemble of Earth system models produced an even wider spread of outcomes (Arora et al. 2013) than had been documented a decade earlier as more model complexity was added. An important approach to improving predictability of land–atmosphere climate futures is the application of emergent constraints on carbon–climate feedbacks (Wenzel et al. 2014). A subset of CMIP5 models forced with identical emissions and allowed to predict the behavior of land and ocean sinks and atmospheric CO$_2$ found a spread of almost 350 ppm in CO$_2$ concentration in 2100 (Hoffman et al. 2014).

Uncertain carbon–climate feedback resulted in a spread in radiative forcing of more than 2 W m$^{-2}$, comparable to emission scenario uncertainty. Hoffman et al. (2014) compared the models predicted CO$_2$ concentrations in 2010 to observations and found that their biases in the present day were good linear predictors of the spread in 2100. Using integral constraints on anthropogenic carbon inventories in the ocean and atmosphere, they adjusted carbon sinks to match. This reduced the spread of CO$_2$ in 2100 by a factor of 7 relative to the control (CMIP) simulations, showing the potential for leveraging emergent constraints to solve the carbon–climate feedback problem.

The International Land Model Benchmarking Project (ILAMB; Hoffman et al. 2016) provides a comprehensive suite of observational datasets from flux towers, field experiments, satellite imagery, and atmospheric sampling in a transparent framework for model evaluation and intercomparison. Dozens of land modeling groups from around the world have participated in the development of the benchmarks, and in model intercomparison and evaluation studies. As of late 2017, model evaluation and improvement is among the highest priorities for predictive modeling of land–atmosphere futures in the Earth system (Huntzinger et al. 2017).

10. The road goes ever on

Developments in atmospheric dynamics and physics, instrumentation and observing practice, and digital computing have made the utopian visions of Abbe, Bjerknes, and Richardson an everyday reality. Global numerical weather prediction models are now at the center of operational forecasting and enable us to predict the weather for several days in advance with a high degree of confidence. Progress has been rapid; the useful range of deterministic prediction has been increasing by about one day per decade (Bauer et al. 2015). In addition, Earth system models are now being used to simulate future climate changes that will have enormous societal consequences. Using Earth system models, we are gaining great insight into the factors causing changes in our climate, and the likely timing and severity of those changes.

As a result of these spectacular achievements, meteorology and oceanography are now firmly established as quantitative sciences, and their value and validity are demonstrated daily by the acid test of any science: its ability to predict the results of measurements. The advances in Earth system modeling over the past century have been truly revolutionary. The development of comprehensive Earth system models is a major and insufficiently appreciated scientific achievement of the twentieth century. Today’s most advanced models simulate not only the physics of the atmosphere, oceans and land surface, but also a wide range of chemical and biological processes and the associated couplings and feedbacks. The conceptual breadth of the models has rapidly increased over the last few decades, and is now rather breathtaking.

It is also essential, however, to maintain a focus on the conceptual depth of the models. The ever-expanding range of parameterized processes must be tied back to fundamental physics, as securely as possible. Although it is exciting and important to add new processes to a model, it is at least equally important (and, for some of us, equally exciting) to strengthen the conceptual foundations of a model’s “legacy” components, including such things as parameterizations of clouds and turbulence, and the numerical methods used to solve the equations that govern fluid motion, over a wide range of scales, on a great big rotating sphere.

A comprehensive ESM can simulate many of the emergent phenomena that we see in nature, but the output of such a simulation is just a pile of numbers; it is not an explanation of the natural world. To claim that we understand the results of a highly detailed and successful simulation, and by extension that we understand the real world, we must work to create much simpler models that can semiquantitatively reproduce the key results of the comprehensive models. Meeting this inspiring challenge is the highest goal of our science.

Acknowledgments. We gratefully acknowledge valuable input provided by our friend Albert J. Semtner, Jr., who passed away in December 2018. Bert was a pioneer of ocean and sea ice modeling.
Paul Edwards of Stanford University and Wayne Schubert of Colorado State University helped us to locate some of the early references. Additional input came from Richard Somerville of the Scripps Institution for Oceanography, and Milton Halem of the University of Maryland.

We are grateful to the three reviewers for very helpful comments on the manuscript. Bjorn Stevens, in particular, liberally annotated the 186-page-long first version of the manuscript.

David Randall acknowledges support from NSF Grant AGS-1538532. Cecilia M. Bitz is grateful for support from NSF PLR-1643431. Stephen Griffies thanks his longstanding support from GFDL for vigorous ocean circulation models.

The National Center for Atmospheric Research is a major facility sponsored by the National Science Foundation under Cooperative Agreement 1852977.

REFERENCES


——, and S. G. Benjamin, 1993: Regional weather prediction with a model combining terrain-following and isentropic coordinates. Part


Page 12.55


Lee, W.-L., and K. Liou, 2007: A coupled atmosphere–ocean radiative transfer system using the analytic four-stream radiative transfer system using the analytic four-stream


Stier, P., and Coauthors, 2005: The aerosol-climate model


