Cite this entry as:
Pu Z., Kalnay E. (2018) Numerical Weather Prediction Basics: Models, Numerical Methods, and Data Assimilation. In: Duan Q., Pappenberger F., Thielen J., Wood A., Cloke H., Schaake J. (eds) Handbook of Hydrometeorological Ensemble Forecasting. Springer, Berlin, Heidelberg

Numerical Weather Prediction Basics: Models, Numerical Methods, and Data Assimilation

Zhaoxia Pu and Eugenia Kalnay

Contents

Abstract

Numerical weather prediction has become the most important tool for weather forecasting around the world. This chapter provides an overview of the fundamental principles of numerical weather prediction, including the numerical framework of models, numerical methods, physical parameterization, and data assimilation. Historical revolution, the recent development, and future direction are introduced and discussed.

Z. Pu (✉)
Department of Atmospheric Sciences, University of Utah, Salt Lake City, UT, USA
e-mail: Zhaoxia.Pu@utah.edu

E. Kalnay
Department of Atmospheric and Oceanic Science, University of Maryland, College Park, MD, USA
e-mail: ekalnay@atmos.umd.edu

Q. Duan et al. (eds.), *Handbook of Hydrometeorological Ensemble Forecasting*,
https://doi.org/10.1007/978-3-642-40457-3_11-1

Keywords
Numerical weather prediction · Numerical methods · Physical parameterization · Data assimilation

1 Introduction: Basic Concept and Historical Overview

Along with advances in computer technology, numerical weather prediction (NWP) has become the central component of weather forecasting. For instance, in the United States, daily weather forecasting begins with a supercomputer at the National Oceanic and Atmospheric Administration (NOAA) in Washington, DC. In Europe, the European Centre for Medium-Range Weather Forecasts (ECMWF), the world's largest numerical weather prediction center, provides advanced weather guidance for all member countries of the European Union. Around the world, most countries use NWP as key guidance for their operational weather prediction.

The basic concept of NWP is to solve a set of partial differential equations (PDEs) that govern atmospheric motion and evolution (Kalnay 2003). As will be described in Sect. 2.1.2, this set of PDEs describes basic conservation laws, including the conservation of momentum, mass, energy, and water vapor. In order to predict the atmospheric state in the future, we must integrate this set of equations forward. Therefore, NWP is an initial value problem: given the current atmospheric conditions (initial conditions), we integrate the set of PDEs to obtain future atmospheric states.

This initial value problem was defined early in the 1900s (e.g., Bjerknes 1904). Bjerknes (1904) stated that the ultimate problem in meteorology is weather forecasting (predicting future atmospheric conditions) and outlined an approach for tackling it. According to his approach, two conditions must be satisfied to successfully predict future atmospheric states:

I. The present atmospheric conditions must be characterized as accurately as possible.
II. The intrinsic laws, according to which the subsequent states develop out of the preceding ones, must be known.

He outlined a program that was subdivided into three partial problems or components:

1. The observation component
2. The diagnostic or analysis component
3. The prognostic component

Components 1 and 2 are related to the characterization of the present state (condition I), while component 3 is related to condition II.

In today's terminology, Bjerknes's approach would be called deterministic because the forecast is assumed to be completely determined from the present state. In addition, component 1 here refers to the global observing system, although that was not established until the 1970s. Component 2 would now be named data assimilation, which combines observation information and short-range weather forecasts to form the best possible initial conditions. Component 3 involves solving the PDEs with numerical methods.

After Bjerknes, Richardson (1922) made the first attempt at NWP by hand. He used full primitive equations and a finite difference scheme. He divided the region of interest into cells, like the squares on a chessboard. He read the atmospheric conditions from a weather map using manual interpolation. Even though his methodology was impeccable, he obtained the forecast change in surface pressure at 145 mb in 6 h! The failure of Richardson's forecast set the NWP concept back into the theoretical world for many years. It was not until the 1950s that NWP was attempted again. Charney made the first successful numerical weather forecast with barotropic potential vorticity equations (Charney et al. 1950). Specifically, between the 1920s and 1950s, significant progress was made in the following areas:

- *Dynamic meteorology*. Atmospheric motion includes multiple temporal and spatial scales. Thus, the scale analysis method can be used to simplify the NWP equations based on the scale of motion in which one is interested in making a weather forecast. Based on scale analysis, a Rossby number is defined to validate the geostrophic flow in the midlatitude synoptic atmosphere. The quasi-geostrophic theory was derived to explain the circulations of synoptic flow in the midlatitudes. Conditions for baroclinic instability were also derived. The major benefit of dynamic meteorology to NWP is in solving the simplified equations according to the targeted scale of forecasts instead of having to deal with the full primitive equations (See details in Holton 2004 and Kalnay 2003).
- *Advances in numerical analysis*. Since analytical solutions for NWP equations do not exist, numerical methods must be used in order to archive the numerical integration in discrete grid spaces. The Courant-Friedrichs-Lewy (CFL) criterion sets a bottom-line of requirements between sizes of grid spaces and time steps in order to retain computational stability. Better understanding of nonlinear computational stability also helps in designing schemes that can be used to solve PDEs accurately and efficiently with numerical methods (See details in Kalnay 2003).
- *Atmospheric observations*. The invention of radiosonde made it possible to probe conditions in the upper atmosphere. During World War II, many countries started their radiosonde network. This has been a great help in improving the accuracy of the initial conditions required for NWP.
- *Invention of electronic computers*. The invention of electronic computers enhanced the efficiency of scientific calculation tremendously. Thus, computers help scientists implement NWP in operational practice, and operational NWP centers have been major users of supercomputers.

In 1950, when the first computer forecast was generated with the Electronic Numerical Integrator and Calculator (ENIAC), the first electronic computer in the world, the NWP became operationally practical.

Since the 1950s, continuous and rapid developments have been made in NWP:

- The maturity of global observing systems in the 1970s (http://www.wmo.int/pages/prog/www/OSY/GOS.html) and the use of satellite and radar observations
- Advances in data assimilation techniques (see Daley 1991; Kalnay 2003; Evensen 1994, also details in section "Data Assimilation")
- Significant advances in computer technology and the development of numerical methods, especially the development of global spectral models, semi-Lagrangian models, and high-resolution regional models (see Robert 1982; Williamson 2007; Lynch 2008)
- Rapid development in physical parameterizations (see Arakawa 1997, 2004; Stensrud 2007)

In addition, some new developments have taken place in recent years:

- Ensemble forecasting: Instead of a single deterministic forecast, ensemble forecasting has become the mainstream method of operational NWP today (see Kalnay 2003; Bauer et al. 2015).
- Advanced data assimilation methods for satellite and radar data (e.g., Lorenc 1986; Daley 1991; Kalnay 2003; Houtekamer and Zhang 2016).
- Coupled atmosphere-ocean-land models (e.g., Hodur 1997; Chen and Dudhia 2001; Ek et al. 2003; Tolman 2014).

Today, NWP has become a multidisciplinary science in both research and operational environments. Many countries have their own NWP systems for daily operational forecasting, including both global and regional model systems for NWP. In addition, NWP computer models have become useful tools for research and education. The development of community weather and climate models, led by the US institutes, e.g., National Center for Atmospheric Research (NCAR), Penn State University, etc. makes these weather and climate research and forecasting models available to many research institutes and universities around the world.

Along with the rapid development in computer power and computer science during the last 60 years, NWP skill has been steadily improved. Figure 1 shows the forecast skill improvements at ECMWF between 1981 and 2014 (note that other centers follow a similar trend of improvement). It is clear that NWP products are quite reliable within a 5-day range and useful in a 7-day range. The improvement attained since the late 1990s from the advanced use of satellite observations is remarkable, especially in the Southern Hemisphere.

2 NWP Models and Numerical Methods

2.1 Basic Equations

There is a complete set of seven equations with seven unknowns that governs the evolution of the atmosphere: Newton's second law or conservation of momentum (three equations for the three velocity components), the continuity equation or

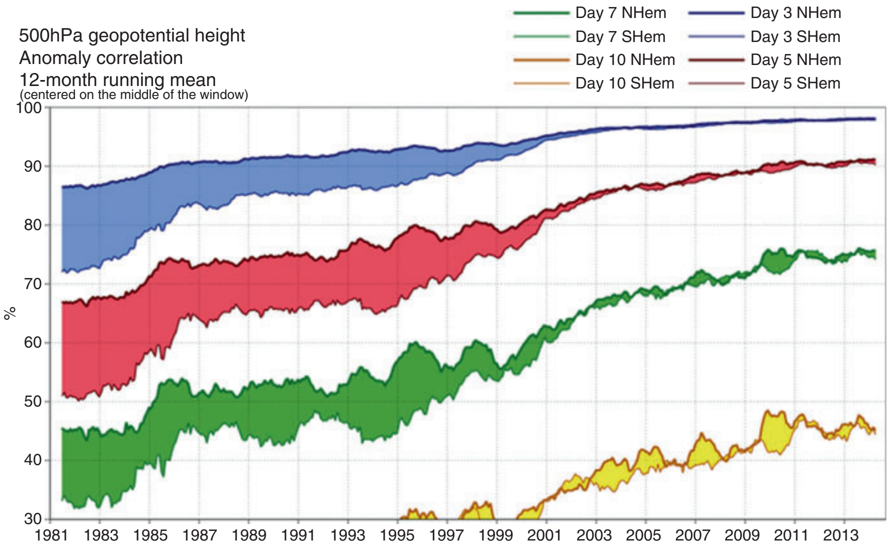

Fig. 1 Time series of the annual running mean of anomaly correlations of 500 hPa geopotential height forecasts evaluated against the operational analyses for the period of January 1981 till present. Values plotted at a particular month are averages over that month, the previous 5 months, and the following 6 months. Forecast lead times of 3, 5, 7, and 10 days are shown, for scores averaged over the northern (bold lines) and southern (thin lines) extratropics. The shading shows differences in scores between the two hemispheres at the forecast ranges indicated. (Adapted and extended from Simmons and Hollingsworth 2002. Verification follows updated WMO/CBS guidelines as specified in the Manual on the GDPFS, Volume 1, Part II, Attachment II.7, Table F, (2010 Edition – Updated in 2012); anomalies are computed with respect to ERA-Interim-based climate (Courtesy of ECMWF web site http://www.ecmwf.int; Also see Bauer et al. 2015)

conservation of mass, the equation of state for ideal gases, the first law of thermodynamics or conservation of energy, and a conservation equation for water mass.

In the Cartesian coordinate system, this set of equations can be written as follows:

$$\frac{\mathrm{d}\,\vec{\mathrm{V}}}{\mathrm{dt}} = -\alpha\,\vec{\nabla}\,p - \vec{\nabla}\,\Phi + \vec{F} - 2\Omega\times\vec{V} \tag{1}$$

$$\frac{\partial\rho}{\partial t} = -\,\vec{\nabla}\cdot\left(\rho\,\vec{V}\right) \tag{2}$$

$$p\alpha = \mathrm{RT} \tag{3}$$

$$Q = C_p\frac{\mathrm{dT}}{\mathrm{dt}} - \alpha\frac{\mathrm{dp}}{\mathrm{dt}} \tag{4}$$

$$\frac{\partial\rho q}{\partial t} = -\,\vec{\nabla}\cdot\left(\rho\,\vec{\mathrm{V}}\,\mathrm{q}\right) + \rho(E - C) \tag{5}$$

where $\vec{V} = (u,v,w)$ represents the velocity of air, t is arbitrary time, α is specific volume, ρ is density, p and T are pressure and temperature, Φ is geopotential height, q is the water vapor mixing ratio, Q is heating, E and C represent evaporation and condensation, respectively, R is the gas constant, and $\vec{F}$ is the friction force.

In spherical coordinates, assume that λ and ϕ are longitude and latitude and r is the radius of the Earth:

$$u = \text{zonal}(\text{positiveeastward}) = r\cos\varphi \frac{d\lambda}{\text{dt}}$$

$$v = \text{meridional}(\text{positivenorthward}) = r\frac{d\varphi}{\text{dt}} =$$

$$w = \text{vertical}(\text{positiveup}) = \frac{\text{dr}}{\text{dt}}$$

Since $\vec{V} = \text{u}\ \vec{\text{i}}\ +\text{v}\ \vec{\text{j}}\ +\text{w}\ \vec{\text{k}}$ and $r = a + z; \quad a \gg z; \quad r \approx z; \frac{\partial}{\partial r} = \frac{\partial}{\partial z}$.

The Eqs. (1), (2), (3), (4), and (5) can be written as:

$$\frac{\partial u}{\partial t} = -u\frac{\partial u}{\partial x} - v\frac{\partial u}{\partial y} - w\frac{\partial u}{\partial z} + \frac{\text{uv}\tan\phi}{a} - \frac{\text{uw}}{a} - \frac{1}{\rho}\frac{\partial p}{\partial x} - 2\Omega(w\cos\phi - v\sin\phi) + \text{Fr}_x \tag{6}$$

$$\frac{\partial v}{\partial t} = -u\frac{\partial v}{\partial x} - v\frac{\partial v}{\partial y} - w\frac{\partial v}{\partial z} - \frac{u^2\tan\phi}{a} - \frac{\text{uw}}{a} - \frac{1}{\rho}\frac{\partial p}{\partial y} - 2\Omega u\sin\phi + \text{Fr}_y \tag{7}$$

$$\frac{\partial w}{\partial t} = -u\frac{\partial w}{\partial x} - v\frac{\partial w}{\partial y} - w\frac{\partial v}{\partial z} - \frac{u^2 + v^2}{a} - \frac{1}{\rho}\frac{\partial p}{\partial z} + 2\Omega u\cos\phi - g + \text{Fr}_z \tag{8}$$

$$\frac{\partial T}{\partial t} = -u\frac{\partial T}{\partial x} - v\frac{\partial T}{\partial y} + (\gamma - \gamma_d)w + \frac{1}{c_p}\frac{\text{dH}}{\text{dt}} \tag{9}$$

$$\frac{\partial \rho}{\partial t} = -u\frac{\partial \rho}{\partial x} + -v\frac{\partial \rho}{\partial y} - w\frac{\partial \rho}{\partial z} - \rho\left(\frac{\partial u}{\partial x} + \frac{\partial v}{\partial y} + \frac{\partial w}{\partial z}\right) \tag{10}$$

$$\frac{\partial q_v}{\partial t} = -u\frac{\partial q_v}{\partial x} - v\frac{\partial q_v}{\partial y} - w\frac{\partial q_v}{\partial z} + Q_v \tag{11}$$

$$p\alpha = \text{RT} \tag{12}$$

where Υ and Υ_d are the lapse rate and dry adiabatic lapse rate, respectively. Note that the terms in the equations related to the diabatic effects (H), friction (F_r), gain, or losses of water through phase changes $(Q_{v)}$ must be defined within the model. Commonly, the set of equations above is called the primitive equations. These seven equations with seven unknowns represent a set of universal equations for NWP.

In reality, since the numerical models are built for various purposes that deal with different scales, it is expected that this set of equations will be simplified with some assumptions. For instance, the hydrostatic balance equation will replace the vertical motion equation if the model is designed for dealing with large scales only. The quasi-Boussinesq or anelastic approximation will be used to make the density a constant and also eliminate high-frequency waves in the solution in order to retain computational stability. In addition, the form of the equations can be changed as various coordinate systems (e.g., Cartesian vs. spherical or pressure coordinates), especially vertical coordinate systems (e.g., sigma vs. pressure vertical coordinates), are used.

To these equations we must add appropriate boundary conditions at the bottom and top of the atmosphere, then solve them using an integration process as suggested by Richardson (1922).

$$\begin{gathered}\frac{\partial \varphi}{\partial t} = F(\varphi,t) \\ \varphi|_{t+\Delta t} = \varphi|_t + F\left(\varphi|_t,t\right)\Delta t\end{gathered} \tag{13}$$

2.2 Numerical Frameworks of NWP Model

2.2.1 Finite Difference Equations (FDEs)

The analytical solution for the aforementioned set of equations (Eqs. 6, 7, 8, 9, 10, 11, and 12) is impossible. The equations must be solved using the discrete form with numerical methods. Therefore, finite difference equations (FDEs) can be used to find approximate solutions of the PDEs.

Using the advection equation as an example:

$$\frac{\partial u}{\partial t} + c\frac{\partial u}{\partial x} = 0 \tag{14}$$

we take discrete values for x and t: $x_j = j\Delta x$ and $t_n = n\Delta t$, where $\Delta x$is the grid space and Δt is the time step of integration. The solution of the FDE is defined at the discrete points $(x_j,\ t_n) = (j\Delta x,\ n\Delta t)$:

$$U_j^n = U(j\Delta x, n\Delta t) = U\left(x_j, t_n\right) \tag{15}$$

Here we use a small u to denote the solution of the PDE (continuous) and a capital U to denote the solution of the FDE (discrete).

The FDE that is used to approximate PDE (14) can be written as follows:

$$\frac{U_j^{n+1} - U_j^n}{\Delta t} + c\frac{U_j^n - U_{j-1}^n}{\Delta x} = 0 \tag{16}$$

This is called an upstream scheme if we assume $c > 0$. Note that both differences are noncentered with respect to the point $(x_j, t_n) = (j\Delta x, n\Delta t)$.

Since we employ an FDE to approximate a PDE, two fundamental conditions should be satisfied:

(i) The FDE should be consistent with the PDE.
(ii) For a given time $t > 0$, the solution of the FDE should converge to that of the PDE as $\Delta x \to 0$ and $\Delta t \to 0$.

In order to fulfill these two requirements, the numerical schemes used in the FDE must be as accurate as possible. We say that the FDE is consistent with the PDE if, in the limit $\Delta x \to 0$, $\Delta t \to 0$, the FDE coincides with the PDE. This requires that the solutions of the FDE be consistent approximations of the solutions of the PDE. The difference between the PDE and FDE is the discretization error or local (in space and time) truncation error.

In addition, it is very important to keep computational stability during the integration (prediction) process. Commonly, the Courant-Friedrichs-Lewy, or CFL, condition must be satisfied when specifying the time step Δt for a given grid size Δx:

$$0 \le c\frac{\Delta t}{\Delta x} \le 1 \tag{17}$$

where $c\frac{\Delta t}{\Delta x}$ is the so-called CFL number and c is a translation velocity. It can commonly be specified by the typical maximum wind speed at the scale of the synoptic/weather event one is dealing with in the model.

This CFL condition, however, is only a necessary condition that ensures that an FDE is computationally stable so that the solution of the FDE at a fixed time $\mathrm{t} = \mathrm{n}\,\Delta t$ remains bounded as $\Delta t \to 0$. Due to the nonlinearity of the NWP equations, computational instability can occur even when CFL conditions are satisfied. Commonly, the time step should be set smaller than what satisfies the CFL conditions. Nevertheless, some implicit numerical schemes allow using large time steps (e.g., the finite volume implicit schemes). In order to achieve accuracy and stable numerical schemes for time and spatial discretization, many advances have been made in computational mathematics, such as semi-Lagrangian schemes, finite volume schemes, and high-order Runge-Kutta schemes (see Robert 1982; Durran 1999; Lin and Rood 1996; Lin 1997).

2.2.2 Spectral Models

In addition to grid space discretization, spectral models describe the present and future states of the atmosphere using the Galerkin approach to perform space discretization with a sum of basis functions $\boldsymbol{\varphi}(x)$:

$$U(x,t) = \sum_{k=1}^{K} A_k(t)\varphi_k(x) \tag{18}$$

The space derivatives are computed directly from the known $\frac{d\varphi(x)}{\mathrm{dx}}$. This procedure leads to a set of ordinary differential equations (ODEs) for the coefficients $A_k(t)$. For instance, considering the one-way advection equation:

$$\frac{\partial u}{\partial t} + c\frac{\partial u}{\partial x} = 0 \tag{19}$$

we use the Fourier transform pairs:

$$\begin{aligned} \xi(u) = U(k,t) &= \int_{-\infty}^{\infty} u(x,t)\exp[-\mathrm{ikx}]\mathrm{dx} \\ \xi^{-1}(U) = u(x,t) &= \frac{1}{2\pi}\int_{-\infty}^{\infty} U(k,t)\exp[\mathrm{ikx}]\mathrm{dk} \end{aligned} \tag{20}$$

where ξ stands for the forward Fourier transform operator and ξ^{-1} denotes the inverse Fourier transform operator. From the definitions above, we can show that:

$$\begin{aligned} \xi\left(\frac{\partial u}{\partial t}\right) &= \frac{\mathrm{dU}(k,t)}{\mathrm{dt}} \\ \xi\left(\frac{\partial u}{\partial x}\right) &= \mathrm{ikU}(k,t) \end{aligned} \tag{21}$$

so that:

- The Fourier transform of a time derivative of a function is equal to the time derivative of the Fourier transform of the function.
- The Fourier transform of a space derivative of a function is equal to the Fourier transform of the function itself multiplied by ik.

We can use this and take the Fourier transform of the advection equation as follows:

$$\xi\left(\frac{\partial u}{\partial t}\right) + c\,\xi\left(\frac{\partial u}{\partial x}\right) = 0 \tag{22}$$

which is the same as:

$$\frac{\mathrm{dU}}{\mathrm{dt}} + \mathrm{ickU} = 0 \tag{23}$$

By using Fourier transforms, we have turned a PDE into an ODE. We can then integrate the ODE forward in time to find the future value of $U(k,\, t)$ and then take the inverse transform to find $u(x,\, t)$. This is the essence of spectral methods. We convert the PDEs in real space into ODEs in wave space and then solve them.

The space discretization based on a spectral representation is extremely accurate (the space truncation errors are of "infinite" order), because the space derivatives are computed analytically, not numerically. Given this advantage, spectral models better lend themselves to longer-range forecasts than grid-point models with the same resolution. Thus, many operational global models today are spectral models (e.g., NCEP Global Forecast System). However, local forcing processes (e.g., latent heat release, differential surface heat fluxes) are sometimes discontinuous and can be represented only in physical space. In addition, when a linear combination of waves (e.g., spectral harmonics) is used to represent a large gradient or discontinuity, spurious waves can result (the Gibbs phenomenon). For higher resolutions, spectral models are computationally more demanding than grid-point models. Furthermore, spectral models do not conserve mass or energy with precision. For these reasons, only a few regional, limited-area spectral models have been developed and employed for research and operational prediction. One of the most widely used is the NCEP Regional Spectral Model (Juang and Kanamitsu 1994).

2.2.3 Grid Staggering Methods

Once the continuous PDEs are discrete in the grid mesh, all model variables are defined in the grids. Even in spectral models, since the transformations of spectral space to grids and from grids to spectral space are necessary and commonly used, model variables are defined in the grid space to some extent. The arrangement of model variables on different grid points becomes one of the considerations when designing numerical schemes for an NWP model. Instead of arranging all variables at the same grid point, many numerical models adopt a staggered grid approach.

The staggered grid combines several types of nodal points located in different geometrical positions and looks rather complex. However, the staggered grid allows for a natural and more accurate formulation of several crucial PDEs with finite differences; thus it is widely used in numerical models. Figure 2 shows an example of a staggered grid in the horizontal direction. In the vertical direction, most models have adopted a staggered grid, for instance, with the vertical velocity defined at the boundary of the layers and the prognostic variables in the center of the layer (Fig. 3). A nonstaggered vertical grid, allowing the simple implementation of higher-order differences in the vertical, would also be possible, but it would also have more computational modes present in the solution.

2.2.4 Boundary Conditions

Since numerical models commonly deal with part of the universe; boundary conditions are necessary. For instance, top and bottom boundary conditions should be given in a global atmospheric model. A regional atmospheric model requires lateral boundary conditions in addition to top and boundary conditions.

Upper boundary conditions: The altitude of a model top is usually above features of meteorological interest. Commonly, it is in the stratosphere or above. Since longer timescale processes dominate the stratosphere, climate models use a high upper boundary. Above the upper boundary, the only input of interest is from incoming solar radiation, which typically is parameterized. There are many ways to represent the upper boundary. For instance, a rigid lid can cap the model at some specified

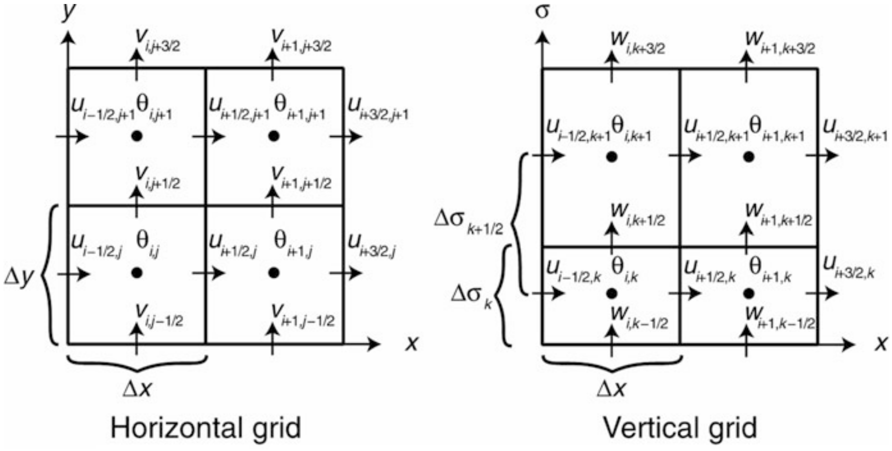

Fig. 2 The Arakawa C staggered grid method (Adapted from Skamarock et al. 2008)

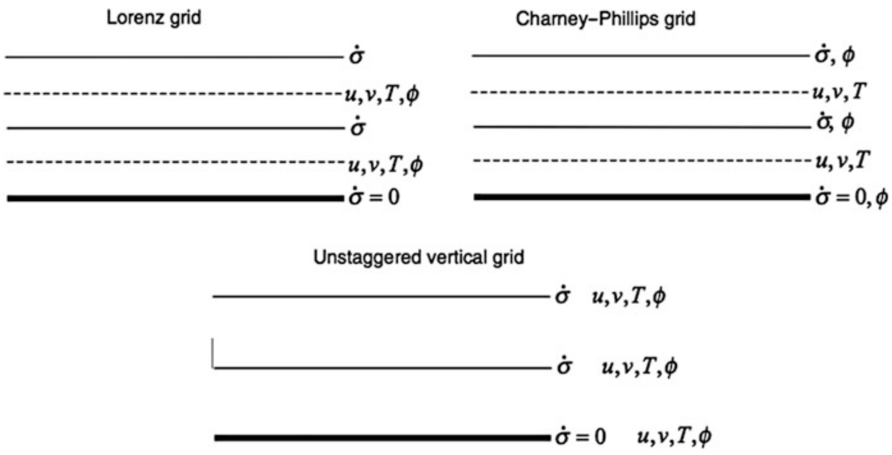

Fig. 3 Staggering in vertical grids (After Arakawa and Konor 1996)

altitude so that energy reaching this lid is reflected downward. A free-surface method treats the model atmosphere and higher altitude as two distinct, nonmixing fluids and also reflects energy downward. Since the key issue of representing upper boundary conditions is how to handle the transfer of energy by gravity waves upward and out of the domain, an absorption/damping layer is incorporated with both the rigid-lid and free- surface methods. This damping layer is usually placed right below the top of the model and applies a diffusion/damping operator to selected vertical levels in order to dampen upward-propagating energy. However, it must be relatively thick to mitigate the development of large vertical gradients and wave reflection issues and can dampen to a predefined reference state or to one defined by the model atmosphere. A radiative boundary condition is also used in some models to mimic the effects of wave energy propagating upward and out of the domain at the top of the model. One should usually choose an upper boundary that is sufficiently high to mitigate the issues from upper boundary conditions. See details in Durran (1999), Kalnay (2003), and Warner (2011).

Bottom boundary conditions: The bottom boundary conditions of NWP models are very complicated, as surface characteristics vary significantly. Therefore, the bottom conditions of NWP models are commonly parameterized or represented by a thermal diffusion surface model, land surface and ocean model, or surface drag schemes (as described in the next section).

Lateral boundary conditions (LBCs): The use of regional models for weather prediction has arisen from the desire to reduce model errors through an increase in horizontal resolution that cannot be afforded in a global model. Operational regional models have been embedded or "nested" into coarser-resolution hemispheric or global models since the 1970s. For instance, the current NCEP North American Mesoscale Forecast System (NAM) model is nested inside the NCEP Global Forecast System (GFS) model. The nesting of regional models requires the use of updated lateral boundary conditions obtained from the global model. Commonly, a lateral boundary condition should be satisfied if (a) it transmits incoming waves from the "host" model and provides boundary information without appreciable change in phase or amplitude, and (b) at the outflow boundaries, reflected waves do not reenter the domain of interest with appreciable amplitude.

In practice, boundary conditions are chosen pragmatically and tested numerically to check their appropriateness. Popular choices for lateral boundary conditions include both one-way and two-way nested schemes. In the one-way lateral boundary conditions, the host model, with coarser resolution, provides information about the boundary values to the nested regional model, but it is not affected by the regional model solution. In a two-way interaction in the boundary conditions, i.e., the (presumably more accurate) regional solution, in turn, also affects the global solution.

In addition, to nest a regional model inside a global model, many regional models use the nested domain technique to achieve high-resolution simulations and forecasts, with the high-resolution domain nested inside the coarser regional model domain. In this case, the lateral boundary conditions should also be addressed in either a one-way or two-way interaction. Furthermore, variable resolution models have been developed in recent years. With the use of continuously stretched horizontal coordinates, only the region of interest is solved with high resolution in a variable resolution model. It is evident that with this approach, the equations in regional high-resolution areas do not require special boundary conditions and they do influence the solutions in the regions of coarser resolution so that they can be considered as two-way interactive nesting (see Kalnay 2003; Warner 2011). Their disadvantage is that the smallest grid size requires the use of short-time steps for the whole domain.

2.3 Global and Regional Models

Both global and regional models are used for NWP. Global models are generally used for guidance in medium-range forecasts (more than 3 days) and for climate simulations. At NCEP, for example, global models are run through 16 days every day. Because the horizontal domain of these global models is the whole Earth, they usually cannot be run at high resolution. However, with advances in computer power, the resolution of global models has increased significantly. For instance, the NCEP Global Forecast System (GFS) and ECMWF medium-range forecast model now run at nearly 16 km (T1297) horizontal resolution, about ten times the horizontal resolution of 20 years ago!

For more detailed forecasts, it is necessary to increase the resolution, and this can be done over only limited regions of interest. Regional models are used for shorter-range forecasts (typically 1–3 days) and are run with a resolution two or more times higher than that of global models. For example, in 1997 the NCEP global model was run with 28 vertical levels, with a horizontal resolution of 100 km for the first week and 200 km for the second week. The regional (Eta) model was run with a horizontal resolution of 29 km and 50 levels. Today, the NCEP regional North American Mesoscale Forecast System (NAM) model runs at a grid spacing of several kilometers (<10 km) with about 100 vertical levels. Because of their higher resolution, regional models have the ability to reproduce smaller-scale phenomena such as fronts, squall lines, and much better orographic forcing than global models. On the other hand, regional models are not “self-contained” because they require lateral boundary conditions at the borders of the horizontal domain. These boundary conditions must be as accurate as possible, because otherwise the interior solution of the regional models quickly deteriorates. Therefore, it is customary to “nest” the regional models within another model with coarser resolution, whose forecast provides the evolving boundary conditions. For this reason, regional models are used only for short-range forecasts. After a certain period, which is proportional to the size of the model, the information contained in the high-resolution initial conditions is “swept away” by the influence of the boundary conditions, and the regional model becomes merely a “magnifying glass” for the coarser model forecast in the regional domain. This can still be useful, for example, in climate simulations performed for long periods (seasons to multiple years), which therefore tend to be run at coarser resolution. A “regional climate model” can provide a more detailed version of the coarse climate simulation in a region of interest. Several other major NWP centers in Europe, including the United Kingdom, France, and Germany and in Japan, Australia, and Canada also have similar global and regional models, whose details can be obtained at their web sites.

2.4 Physical Parameterizations

2.4.1 Basic Principles

The basic equations of an NWP model (e.g., Eqs. 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, and 12) include terms for friction, heating source, and evaporation and condensation processes. Specifically, the momentum equation has the effect of eddy fluxes of momentum; the thermodynamics equation includes radiative heating and cooling, sensible heat fluxes, and condensation and evaporation; and the water vapor equation includes condensation and evaporation, as well as moisture flux. These physical processes in numerical models represent their contribution. Thus, the model should include surface and planetary boundary layer processes, radiative transfer, and cloud microphysics in order to represent their contributions.

Atmospheric motion includes a broad spectrum of temporal and spatial scales. The timescale spans from 1 to 10^6 s and beyond, including the life cycle of a small turbulent air blob to a local storm, synoptic motions, and seasonal to interannual variations. The spatial scale ranges from 1 cm to 10,000 km, including the turbulent microscale, convective scale, mesoscale, and large scale.

Due to the use of numerical discretization methods to solve PDEs, the grid resolution of the atmospheric model is always limited. Therefore, any processes that occur on a scale smaller than the grid space cannot be explicitly represented in the numerical model, even though their contribution cannot be ignored. To give an example, here we apply Reynolds' average to the u component of the motion equation (Eq. 6).

Assume that any variable (e.g., u, v, w, T, p) can be separated into resolvable and unresolvable components, i.e., one can split all dependent variables into mean and turbulent parts, respectively. The mean is defined as an average over a grid cell, as described by Pielke (2002). For example:

$$\begin{aligned} u &= \bar{u} + u', \\ T &= \bar{T} + T', \text{ and} \\ p &= \bar{p} + p'. \end{aligned}$$

These expressions are substituted into Eqs. (6), (7), (8), (9), (10), (11), and (12); produce the expansions such the following one for Eq. (6):

$$u\frac{\partial u}{\partial x} = (\bar{u} + u')\frac{\partial}{\partial x}(\bar{u} + u') = \bar{u}\frac{\partial \bar{u}}{\partial x} + \bar{u}\frac{\partial u'}{\partial x} + u'\frac{\partial \bar{u}}{\partial x} + u'\frac{\partial u'}{\partial x}.$$

$$\overline{u\frac{\partial u}{\partial x}} = \overline{\bar{u}\frac{\partial \bar{u}}{\partial x}} + \overline{\bar{u}\frac{\partial u'}{\partial x}} + \overline{u'\frac{\partial \bar{u}}{\partial x}} + \overline{u'\frac{\partial u'}{\partial x}}.$$

Since

$$\begin{aligned} &\overline{a'} = 0, \\ &\bar{\bar{a}} = \bar{a} \text{ and } \overline{\bar{\mathrm{a}}\mathrm{b}} = \bar{\bar{\mathrm{a}}}\bar{\bar{\mathrm{b}}} = \bar{\mathrm{a}}\bar{\mathrm{b}}, \text{ and} \\ &\overline{\bar{\mathrm{a}}\mathrm{b}'} = \bar{\bar{\mathrm{a}}}\overline{\mathrm{b}'} = \bar{\mathrm{a}}\overline{\mathrm{b}'} = 0. \end{aligned}$$

Therefore,

$$\overline{u\frac{\partial u}{\partial x}} = \bar{u}\frac{\partial \bar{u}}{\partial x} + \cancelto{0}{\bar{u}\frac{\partial \overline{u'}}{\partial x}} + \cancelto{0}{\overline{u'}\frac{\partial \bar{u}}{\partial x}} + \overline{u'\frac{\partial u'}{\partial x}} = \bar{u}\frac{\partial \bar{u}}{\partial x} + \overline{u'\frac{\partial u'}{\partial x}}.$$

Thus,

$$\frac{\partial u}{\partial t} = -u\frac{\partial u}{\partial x} - v\frac{\partial u}{\partial y} - w\frac{\partial u}{\partial z} - \frac{1}{\rho}\frac{\partial p}{\partial x} + \mathrm{fv} + \frac{1}{\rho}\left(\frac{\partial \tau_{\mathrm{xx}}}{\partial x} + \frac{\partial \tau_{\mathrm{yx}}}{\partial y} + \frac{\partial \tau_{\mathrm{zx}}}{\partial z}\right).$$

where

$$\tau_{\mathrm{zx}} = \mu\frac{\partial u}{\partial z},$$

$$\frac{\partial \bar{u}}{\partial t} = -\bar{u}\frac{\partial \bar{u}}{\partial x} - \bar{v}\frac{\partial \bar{u}}{\partial y} - \bar{w}\frac{\partial \bar{u}}{\partial z} - \frac{1}{\bar{\bar{\rho}}}\frac{\partial \bar{p}}{\partial x} + \mathrm{f}\bar{\mathrm{v}} - \overline{u'\frac{\partial u'}{\partial x}} - \overline{v'\frac{\partial u'}{\partial y}} - \overline{w'\frac{\partial u'}{\partial z}} + \frac{1}{\bar{\bar{\rho}}}\left(\frac{\partial \bar{\tau}_{\mathrm{xx}}}{\partial x} + \frac{\partial \bar{\tau}_{\mathrm{yx}}}{\partial y} + \frac{\partial \bar{\tau}_{\mathrm{zx}}}{\partial z}\right).$$

$$\frac{\partial u'}{\partial x} + \frac{\partial v'}{\partial y} + \frac{\partial w'}{\partial z} = 0.$$

$$\frac{\partial \bar{u}}{\partial t} = -\bar{u}\frac{\partial \bar{u}}{\partial x} - \bar{v}\frac{\partial \bar{u}}{\partial y} - \bar{w}\frac{\partial \bar{u}}{\partial z} - \frac{1}{\bar{\bar{\rho}}}\frac{\partial \bar{p}}{\partial x} + \mathrm{f}\bar{\mathrm{v}} - \overline{\frac{\partial \mathrm{u'u'}}{\partial x}} - \overline{\frac{\partial \mathrm{u'v'}}{\partial y}} - \overline{\frac{\partial \mathrm{u'w'}}{\partial z}} + \frac{1}{\bar{\bar{\rho}}}\left(\frac{\partial \bar{\tau}_{\mathrm{xx}}}{\partial x} + \frac{\partial \bar{\tau}_{\mathrm{yx}}}{\partial y} + \frac{\partial \bar{\tau}_{\mathrm{zx}}}{\partial z}\right).$$

$$T_{\mathrm{xx}} = -\bar{\rho}\mathrm{u'u'},$$
$$T_{\mathrm{yx}} = -\bar{\rho}\mathrm{u'v'},$$
$$T_{\mathrm{zx}} = -\bar{\rho}\,\overline{\mathrm{u'w'}}.$$

$$\begin{aligned}\frac{\partial \bar{u}}{\partial t} &= -\bar{u}\frac{\partial \bar{u}}{\partial x} - \bar{v}\frac{\partial \bar{u}}{\partial y} - \bar{w}\frac{\partial \bar{u}}{\partial z} - \frac{1}{\bar{\bar{\rho}}}\frac{\partial \bar{p}}{\partial x} + \mathrm{f}\bar{\mathrm{v}} \\ &+ \frac{1}{\bar{\bar{\rho}}}\left(\frac{\partial}{\partial x}(\tau_{\mathrm{xx}} + T_{\mathrm{xx}}) + \frac{\partial}{\partial y}(\tau_{\mathrm{yx}} + T_{\mathrm{yx}}) + \frac{\partial}{\partial z}(\tau_{\mathrm{zx}} + T_{\mathrm{zx}})\right).\end{aligned} \tag{24}$$

In these equations, the first component of the five terms can be explicitly represented by model grid values. The second component of the three terms inside the parentheses cannot be explicitly resolved at model grid points, but their contributions cannot be ignored, because these subgrid-scale processes depend on and in turn affect the large-scale fields and processes that are explicitly resolved by numerical models. Therefore, parameterization schemes are then necessary in order to properly describe the impact of these subgrid-scale mechanisms on the large-scale flow of the atmosphere. In other words, the ensemble effect of the subgrid-scale processes has to be formulated in terms of the resolved grid-scale variables. Furthermore, forecast weather parameters, such as 2-m temperature, precipitation, and cloud cover, are computed by the physical parameterization of the model.

Overall, an NWP model consists of two major parts, as shown in Fig. 4: The "dynamics of the model" indicates schematically the resolved processes and the "model physics," the processes that must be parameterized. Since all physical processes interact not only with the dynamics of the model but also with each other, an NWP model is numerically complicated and computationally expensive.

2.4.2 Overview of Physical Parameterizations

Physical parameterization schemes in a numerical model should be designed to (1) represent the physical processes that interact with the dynamics; and (2) explicitly

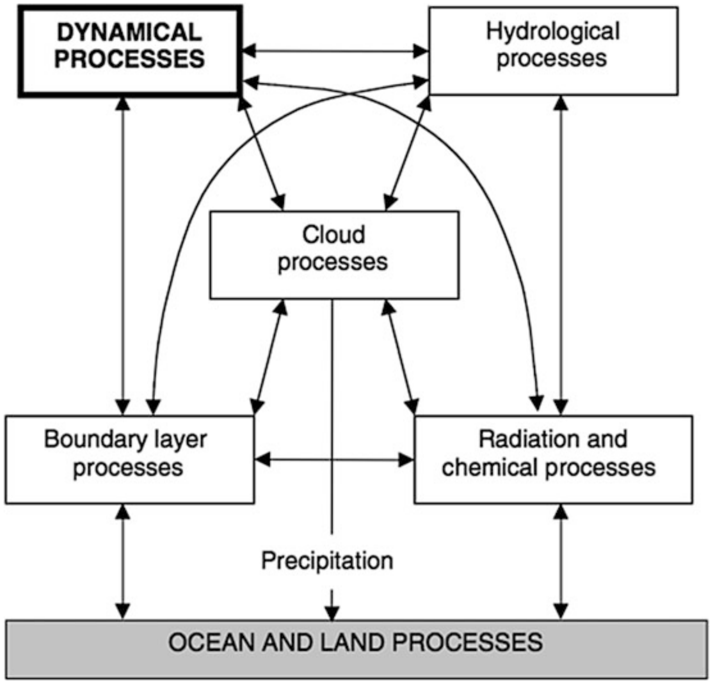

Fig. 4 Physical processes in the atmosphere and their interactions. The dynamical processes for resolvable scales, in bold, are explicitly computed by the model "dynamics." The other subgrid-scale processes are parameterized in terms of the resolved-scale fields (Adapted from Arakawa 1997, 2004)

calculate the contributions from the subgrid-scale processes parameterized as a function of the large-scale, resolved scales. The common parameterization schemes included in a numerical model, taking as an example those included in the current ECMWF global model, can be briefly be described as follows.

Radiation and chemical processes: The radiation scheme performs computations of shortwave and long-wave radiative fluxes using the predicted values of temperature, humidity, cloud, and monthly mean climatologies for aerosols and the main trace gases (CO_2, O_3, CH_4, N_2O, etc.). The radiation parameterization describes the radiative transfer processes. Cloud-radiation interactions are usually taken into account. Since finding the solution of radiative transfer equations to obtain the fluxes is computationally very expensive, depending on the model configuration, full radiation calculations are commonly performed on a reduced (coarser) radiation grid and/or at a reduced time frequency. The results are then interpolated back to the original grid. (See Liou 1980; Stephens 1984 for details about the radiative processes.)

Convection: The moist convection scheme represents deep (including congestus), shallow, and midlevel (elevated moist layers) convection. The distinction between deep and shallow convection is made in the convection scheme. Moist convection also resolves the entrainment process and diurnal variation of the convection. The effects of updrafts and downdrafts are also simulated (See Arakawa 2004).

Cloud microphysics and precipitation: Cloud microphysics encompasses all cloud processes that occur on the scales of the cloud droplets and the hydrometeors, including cloud droplets, raindrops, ice crystals, snow flake, rimed ice particles, graupel particles, and hail stones, rather than on the scale of the cloud itself. Microphysical parameterizations aim to represent, as thoroughly as possible, the processes described in the microphysical processes, including condensation, accretion, evaporation, ice and snow aggregation, accretion by frozen particles, vapor deposition, melting, and freezing.

In a large-scale model, clouds and large-scale precipitation are parameterized with a number of prognostic equations for cloud liquid, cloud ice, rain and snow water content, and a subgrid fractional cloud cover. The cloud scheme represents the sources and sinks of clouds and precipitation due to the major generation and destruction processes, including cloud formation by detrainment from cumulus convection, condensation, deposition, evaporation, and collection, melting, and freezing (see Houze 1993; Straka 2009).

In high-resolution models, especially regional models at the cloud-permitting scale, cloud microphysical processes are explicitly represented by the microphysics of the liquid, ice, and vapor with detailed configurations and phase changes. In models at the cloud-permitting scale, since the clouds are explicitly resolved, cumulus convection schemes can be eliminated.

Soil/surface: The surface parameterization scheme represents the surface fluxes of energy and water and the corresponding subsurface quantities. The scheme should describe different subgrid surface types for vegetation, bare soil, snow, and open water. The surface energy balance equation is also included. Soil layers are represented as well as snow mass and density. The evaporative fluxes consider the fractional contributions from snow cover, wet and dry vegetation, and bare soil. An interception layer collects water from precipitation and dewfall, and infiltration and runoff should be represented, depending on soil texture and subgrid orography. A carbon cycle may be included, and land-atmosphere exchanges of carbon dioxide may be parameterized to respond to diurnal and synoptic variations in the water and energy cycles. The soil/surface parameterization can be classified as a simple bulk scheme or as a full land-surface model.

Turbulent diffusion and planetary boundary layer scheme: The turbulent diffusion scheme represents the vertical exchange of heat, momentum, and moisture through subgrid-scale turbulence. Vertical turbulent transport is treated differently in the surface layer than it is above. For instance, in the surface layer of the ECMWF global model as in 2013, the turbulence fluxes are computed using a first-order K-diffusion closure based on the Monin-Obukhov (MO) similarity theory. Above the surface layer, a K-diffusion turbulence closure is used everywhere, except for unstable boundary layers where an eddy-diffusivity mass-flux (EDMF) framework is applied, to represent the nonlocal boundary-layer eddy fluxes. The scheme is written in moist conserved variables (liquid static energy and total water) and predicts total water variance. A total water distribution function is used to convert from the moist conserved variables to the prognostic cloud variables (liquid/ice water content and cloud fraction) but only for the treatment of stratocumulus. Convective clouds are treated separately by the shallow convection scheme. (See Stull 1988 for details about the boundary layer and turbulence.)

Orographic drag: Limited by their resolution, NWP models cannot fully resolve the orographic features of the terrain. The effects of unresolved orography on the atmospheric flow are parameterized as a sink of momentum (drag). The turbulent diffusion scheme includes a parameterization in the lower atmosphere to represent the turbulent orographic drag induced by small-scale (<5 km) orography. In addition, in stably stratified flow, the orographic drag parameterization represents the effects of low-level blocking due to unresolved orography (blocked flow drag) and the absorption and/or reflection of vertically propagating gravity waves (gravity wave drag) on the momentum budget.

Non-orographic gravity wave drag: The non-orographic gravity wave drag parameterization accounts for the effects of unresolved non-orographic gravity waves. These waves are generated in nature by processes like deep convection, frontal disturbances, and shear zones. Propagating upward from the troposphere, the waves break in the middle atmosphere, comprising the stratosphere and mesosphere, where they deposit momentum and exert a strong drag on the flow (see Teixeira 2014).

2.5 Land-Surface and Ocean Models and Coupled Numerical Models in NWP

2.5.1 Land-Surface Models

The representation of soil, vegetation, snow, mountains, and water bodies is an integral part of the NWP system. Land can affect the weather, the magnitude of the weather effects, and the evolution of human activities. The effects of land-surface state anomalies can persist for several days, thus increasing the importance of correct initial conditions and model evolution. A refined representation of land-surface processes and their accurate initialization hold potential for further improvement of weather prediction up to the monthly range, as indicated in predictability studies.

Land-surface models (LSMs) are used to represent and parameterize land-surface processes. LSMs are important because they provide the necessary lower boundary conditions for NWP and climate models (Ek et al. 2003). They also calculate radiation flux, heat flux, and moisture flux to NWP or climate models. Such fluxes are frequently dominant driving mechanisms for mesoscale circulations. Furthermore, land-surface processes can influence near-surface forecasting such as 2-m temperature, 10-m wind speed, boundary layer structures, and precipitation forecasting.

The energy and water budgets at the land-surface control the temperature and moisture content of the substrate and vegetation, which interact with the atmosphere. The energy conservation equation can be written for a unit mass or unit area of the surface that is experiencing energy gain or loss:

$$R_n + G + \mathrm{LE} + H = 0 \tag{25}$$

where R_n is the net radiation at the surface, namely, the sum of downward longwave radiation, downward shortwave radiation, upward shortwave radiation (reflected by the surface, controlled by the albedo of the surface), and upward longwave radiation (surface emission). G is ground heat flux. It can also be interpreted as minus the rate of heat storage beneath the surface. E is the rate of evaporation, LE represents latent heat flux, and H is sensible heat flux. Thus, the LSM calculates sensible and latent fluxes using parameters in surface and canopy layers.

The soil temperature transfer equation is part of the LSM. For instance, a Fourier law of diffusion can be used to govern the soil heat and moisture transfer:

$$(\rho C)_s \frac{\partial T_{\ s}}{\partial t} = \frac{\partial}{\partial z}\left[\lambda_T \frac{\partial T}{\partial z}\right] \tag{26}$$

where $(\rho C)_s$ is the volumetric soil heat capacity (J m^{-3} K^{-1}). It is a function of soil texture and soil moisture. T is soil temperature, z is the vertical coordinate (distance from the surface), t is time (s), and λ_T is the thermal conductivity.

The soil heat capacity can be estimated as a weighted sum of the heat capacity of its phases. Then

$$(\rho C)_s = (1 - \theta_s)(\rho C)_m + \theta_m(\rho C)_w$$

The subscripts m and w refer to the soil matrix and water, respectively.

The soil water movement obeys “Richard’s” equation:

$$\frac{\partial \theta}{\partial t} = \frac{\partial}{\partial z}\left[\lambda \frac{\partial \theta}{\partial}\right] - \frac{\partial \gamma}{\partial z} - S_\theta \tag{27}$$

λ and γ are functions of soil texture and soil moisture. S_θ represents sources (rainfall) and sinks (evaporation).

The surface water budget equation expresses water conservation as:

$$\mathrm{DS} = P\text{-}R\text{-}E \tag{28}$$

where DS is the change in soil water content, P is precipitation, R is runoff, and E is evaporation. Specifically, evaporation is a function of soil moisture and vegetation type, rooting depth/density, green vegetation cover, etc. Specifically:

$$E = E_{\mathrm{dir}} + E_t + E_c + E_{\mathrm{snow}} \tag{29}$$

E is the total surface evaporation from combined soil/vegetation; E_{dir} is direct evaporation from soil; E_t is transpiration through the plant canopy; E_c is evaporation from canopy-intercepted rainfall; and E_{snow} is sublimation from the snowpack.

With comprehensive representation of the surface energy and water balance and heating and water transport, the inclusion of LSMs in an NWP model has been shown to be beneficial for numerical prediction of near-surface atmospheric conditions as well as quantitative precipitation forecasting (e.g., Chen and Dudhia 2001).

There are many LSMs that have been developed by different centers for various purposes. For instance, the current mesoscale community Weather Research and Forecasting (WRF) model (Skamarock et al. 2008) has six options of land-surface models. The one developed by multiple US government agencies (include NCEP and the Air Force) and universities (e.g., Oregon State University) is the so-called NOAH land-surface model (Ek et al. 2003), which is also used in the NCEP NAM regional model.

2.5.2 Ocean Models

The marine component of the Earth has an important influence on the atmosphere on a range of timescales. A fully coupled model of the marine system may include surface waves, ocean, and sea ice.

It has long been known that waves affect the marine boundary layer of the atmosphere by modifying the surface roughness. Most climate models are coupled to an ocean model. Some NWP models (depending on their applications) are also coupled with ocean models. Ensemble and seasonal forecast systems use a coupled atmosphere-ocean model, which includes a simulation of the general circulation of the ocean and any associated coupled feedback processes.

Ocean wave modeling: Ocean wave modeling is used to predict the genesis and evolution of ocean waves and their associated energy. Many operational centers have developed ocean wave models in either stand-alone or coupled modes. For instance, ECMWF has developed the wave model (WAM), which is coupled to the atmospheric model or runs as a stand-alone model in the limited-area wave (LAW) configuration. Since 1998 ECMWF has been running a coupled forecasting system in which the atmospheric component of the Integrated Forecasting System (IFS) communicates with the WAM through exchange of the Charnock parameter, which determines the roughness of the sea surface (Janssen 2004). At NOAA NCEP, a wave model, WAVEWATCH III® (e.g., Tolman 2014), is developed. WAVEWATCH III solves the random phase spectral action density balance equation for wavenumber-direction spectra. The implicit assumption of this equation is that properties of medium (water depth and current) as well as the wave field itself vary on time and space scales that are much larger than the variation scales of a single wave (see details in Tolman 2014).

Ocean modeling: In addition to the ocean wave model, ocean models have also been developed to compute the time evolution of sea surface elevation, currents, salinity, and temperature. Similar to atmospheric models, ocean models are designed to be either hydrostatic or non-hydrostatic. A baroclinic, primitive-equation ocean model contains conservation equations for mass (continuity) and momentum. Thermodynamics are also used to describe the salinity and temperature. Many numerical schemes and parameterization methods for ocean models are similar to those used in atmospheric models, although there have been many new developments in recent years. Initialization is necessary for ocean models. 3D variational data assimilation methods are used for many existing ocean models. Ocean models have been used in the operational ensemble prediction system and the seasonal forecast system. For instance, ECMWF currently uses an ocean model with 1° of resolution, initialized with the 3D variational assimilation system. Since 2013, ensemble forecasts have been coupled with the atmosphere-wave-ocean model from the start of the forecast. This is important, as it allows capture of the two-way feedback between the atmosphere and sea surface temperatures; for example, when a tropical cyclone moves slowly, it can cool the sea surface. NCEP is implementing a new Local Ensemble Transform Kalman Filter (LETKF) and a gain hybrid that combines the 3D-VAR GODAS and the LETKF, following Penny (2014).

Sea ice modeling: Sea ice is an important component of the Earth's system; it is highly reflective, altering the amount of solar radiation that is absorbed; it changes the salinity of the ocean where it forms and melts; and it acts as a barrier to the exchange of heat and momentum fluxes between the atmosphere and ocean. Current operational weather forecast systems do not commonly predict sea ice dynamically, however. In coupled forecast systems, sea ice modeling is coupled with ocean modeling to represent the dynamic and thermodynamic evolution of sea ice, mainly for seasonal and interannual prediction.

2.5.3 Coupled Numerical Models in NWP

With the rapid development of land and ocean models, many NWP models have become coupled numerical models. For instance, it has been proven that the coupling of land-surface models with atmospheric models can significantly improve quantitative precipitation forecasting (QPF). In addition, in order to accurately predict tropical cyclone formation, intensification, and dissipation, coupled ocean and atmospheric models are found to be necessary. In regional models that emphasize hurricane intensity forecasts (such as hurricane WRF or HWRF), sea spray has also been taken into account in the parameterizations. Coupled atmosphere, land, and ocean models are necessary for medium-range weather forecasting and extended-range forecasting (beyond 10 days).

Land-atmosphere coupling can be achieved by land-surface models, with input from near-surface and atmospheric conditions and characteristics of the Earth's surface to the land-surface model and output to the atmospheric model by providing the water and energy fluxes.

In coupled ocean-atmosphere modeling systems, there are up to three models in use: an atmospheric model, an ocean model, and a wave model. However, owing to computational costs, timescales of interest, and the intended application, many systems today couple two of these models, either an ocean-atmosphere-coupled modeling system or a wave-atmosphere-coupled modeling system. The coupling takes place at the air-sea interface. For example, when all three models are used, the atmospheric model provides the surface stress to the wave model, which uses this information to derive the two-dimensional wave energy spectrum. The wave model provides the wave-induced roughness length to the atmospheric model for use in calculating surface fluxes, which also requires the SST provided by the ocean model. The wave-induced stress from the wave model along with the surface fluxes and radiation from the atmospheric model is used by the ocean model to derive the SST.

While synoptic observational data are generally sufficient to start an atmospheric model, observations are sparse below the surface of the ocean. This leads to the ocean model being run for months or years prior to the start time of any simulation or forecast in order to develop a representative three-dimensional ocean state. During the assimilation period, the ocean model is forced by surface wind stresses provided by global analyses or an atmospheric model, observed sea surface height anomalies derived from satellite data, and the observed SSTs. Wave models generally do not need to be initialized prior to the start of the coupled model simulation unless wave characteristics during the first day are important.

3 Data Assimilation

As mentioned in the beginning of the chapter, to make a forecast, we need to know the current state of the atmosphere and the Earth's surface (land and oceans). Modern numerical weather prediction makes extensive use of terrestrial and satellite observations, along with conventional observations (e.g., surface observations, radiosondes from weather stations, ships, buoys, and other components). These observations provide atmospheric, ocean, and land-surface information. Satellites now provide most data, although more observations are still important.

The weather forecasts produced by operational centers use *data assimilation* to estimate initial conditions for the forecast model from observations. The quality of forecasts depends on how well we use information received in real time from the global observing system, which consists of numerous satellite instruments, weather stations, ships, buoys, and other components. The purpose of data assimilation is to determine a best possible atmospheric state using observations and short-range forecasts. Data assimilation is typically a sequential time-stepping procedure, in which a previous model forecast is compared with newly received observations, the model state is then updated to reflect the observations, a new forecast is initiated, and so on. The update step in this process is usually referred to as the *analysis*; the short model forecast used to produce the analysis is called the *background*.

3.1 Least Squares Theory

The best estimate of the state of the atmosphere (analysis) is obtained, as indicated by Talagrand (1997), from combining prior information about the atmosphere (background or first guess) with observations, but in order to combine them optimally, we also need *statistical information* about the errors in these "pieces of information." A classic example of determining the best estimate of the true value of a scalar (e.g., the true temperature T_t) given two independent observations (or pieces of information), T_1 and T_2, serves as an introduction to statistical estimation:

$$T_1 = T_t + \varepsilon_1; \qquad T_2 = T_t + \varepsilon_2 \tag{30}$$

The observations have errors ε_i that we don't know. Let $E(\ )$ represent the *expected value*, i.e., the average that one would obtain if making many similar measurements. We assume that the instruments that measure T_1 and T_2 are unbiased: $E\,(T_1 - T_t) = E\,(T_2 - T_t) = 0$ or, equivalently,

$$E(\varepsilon_1) = E(\varepsilon_2) = 0 \tag{31}$$

and that we know the variances of the observational errors:

$$E\,\varepsilon_1^2 = \sigma_1^2 \text{ and } E\,\varepsilon_2^2 = \sigma_2^2 \tag{32}$$

We also assume that the errors of the two measurements are uncorrelated:

$$E(\varepsilon_1\varepsilon_2) = 0 \tag{33}$$

Equations (31), (32), and (33) represent the statistical information we need about the actual observations. We try to estimate T_t from a linear combination of the two observations, since they represent all the information that we have about the true value of T:

$$T_a = a_1 T_1 + a_2 T_2 \tag{34}$$

The "analysis" T_a should be unbiased:

$$E(T_a) = E(T_t) \tag{35}$$

which implies

$$a_1 + a_2 = 1 \tag{36}$$

T_a will be the *best estimate* of T_t if the coefficients are chosen to minimize the mean squared error of T_a:

$$\sigma_a^2 = E\left[(T_a - T_t)^2\right] = E\left[\left(a_1(T_1 - T_t)^2 + a_2(T_2 - T_t)^2\right)\right] \tag{37}$$

subject to the constraint (37). Substituting $a_2 = 1 - a_1$, the minimization of σ_a^2 with respect to a_1 gives:

$$a_1 = \frac{\frac{1}{\sigma_1^2}}{\frac{1}{\sigma_1^2} + \frac{1}{\sigma_2^2}}, \qquad a_2 = \frac{\frac{1}{\sigma_2^2}}{\frac{1}{\sigma_1^2} + \frac{1}{\sigma_2^2}} \tag{38}$$

or

$$a_1 = \frac{\sigma_2^2}{\sigma_1^2 + \sigma_2^2}, \qquad a_2 = \frac{\sigma_1^2}{\sigma_1^2 + \sigma_2^2} \tag{39}$$

i.e., the weights of the observations are proportional to the "precision" or accuracy of the measurements (defined as the inverse of the variances of the observational errors). Moreover, substituting the coefficients (39) in (37), we obtain a relationship between the analysis variance and the observational variances:

$$\frac{1}{\sigma_a^2} = \frac{1}{\sigma_1^2} + \frac{1}{\sigma_2^2} \tag{40}$$

i.e., *if the coefficients are optimal, and the statistics of the errors are exact, then the "precision" of the analysis (defined as the inverse of the variance) is the sum of the precisions of the measurements*. More importantly, Eq. (40) also indicates that the error variance of the "analysis" $\left(\sigma_a^2\right)$ is smaller than the error variance of either the "background" or "observations."

According to the least squares theory, one could achieve an analysis out of two uncertain pieces of information (background and observations) and make the analysis more accurate than either one of them alone could. Many data assimilation systems follow this theory.

3.2 Assimilation Methods

In practice, the analysis x^a is obtained by adding the innovations to the background (model forecast or first guess) with weights W that are determined based on the estimated statistical error covariances of the forecast and the observations:

$$\mathbf{x}^a = \mathbf{x}^b + \mathbf{W}\left[\mathbf{y}^o - \mathbf{H}\left(\mathbf{x}^b\right)\right] \tag{41}$$

The innovations are here defined as the difference between the observations and the model "guess" observations, namely, $\mathbf{y}^o - \mathbf{H}(\mathbf{x}^b)$. Specifically, the background (model forecast) is interpolated to the observation location, and if they are of different type, they are converted from model variables to observed variables $\mathbf{y}^o$ (such as satellite radiances or radar reflectivities). The first guess of the observations is therefore $\mathbf{H}(\mathbf{x}^b)$, where $\mathbf{H}$ is the observation operator that performs the necessary interpolation and transformation from model variables to observation space.

Different analysis schemes are based on (41) but differ in the approach taken to combine the background and observations to produce the analysis. Earlier methods such as the successive correction method were of a form similar to (41), with weights determined empirically. The weights are a function of the distance between the observation and the grid point, and the analysis is iterated several times. In optimal interpolation, the matrix of weights $\mathbf{W}$ is determined from the minimization of the analysis errors at each grid point. Even modern advanced data assimilation methods can be interpreted in a similar way, as introduced below. See details in Kalnay (2003).

3.2.1 A Three-Dimensional Variational (3D-VAR) Data Assimilation Method

In the 3D-VAR approach, one defines a cost function proportional to the square of the distance between the analysis and both the background and the observations. The cost function is minimized directly to obtain the analysis. Lorenc (1986) showed that optimal interpolation (OI) and the 3D-VAR approaches are equivalent if the cost function is defined as:

$$J = 1/2\left\{[\mathbf{y}^o - H(\mathbf{x})]^T\mathbf{R}^{-1}[\mathbf{y}^o - H(\mathbf{x})] + 1/2\left(\mathbf{x} - \mathbf{x}^b\right)^T\mathbf{B}^{-1}\left(\mathbf{x} - \mathbf{x}^b\right)\right\} \tag{42}$$

The cost function J in (42) measures the distance of a field x to the observations (the first term in the cost function) and the distance to the first guess or background $\mathbf{x}^b$ (the second term in the cost function). The distances are scaled by the observation error covariance $\mathbf{R}$ and by the background error covariance $\mathbf{B}$, respectively. The minimum of the cost function is obtained for $x = x^a$, which is defined as the "analysis." The analysis obtained in (41) and (42) is the same if the weight matrix in (41) is given by:

$$\mathbf{W} = \mathbf{BH}^T\left(\mathbf{HBH}^T + \mathbf{R}^{-1}\right)^{-1} \tag{43}$$

In 3D-VAR, the minimization of (42) is performed directly, allowing for additional flexibility and a simultaneous global use of the data.

3.2.2 A Four-Dimensional Variational (4D-VAR) Data Assimilation Method

4D-VAR is an important extension of 3D-VAR that allows for observations distributed within a time interval (t_0, t_n) (e.g., Courtier and Talagrand 1987; Derber 1989; Bouttier and Rabier 1997). The cost function includes a term measuring the distance to the background *at the beginning of the interval* and a summation over time of the cost function for each observational increment computed with respect to the model integrated to the time of the observation:

$$\begin{aligned}\mathbf{J}[\mathbf{x}(t_0)] &= \frac{1}{2}\left[\mathbf{x}(t_0) - \mathbf{x}^b(t_0)\right]^T\mathbf{B}_0^{-1}\left[\mathbf{x}(\mathbf{t}_0) - \mathbf{x}^b(t_0)\right] \\ &+ \frac{1}{2}\sum_{i=0}^{N}\left[\mathbf{H}(\mathbf{x}_i) - \mathbf{x}_i^0\right]^T\mathbf{R}_i^{-1}\left[\mathbf{H}(\mathbf{x}_i) - \mathbf{y}_i^o\right]\end{aligned} \tag{44}$$

The control variable (the variable with respect to which the cost function is minimized) is the *initial* state of the model with the time interval $\mathbf{x}(t_0)$, whereas the analysis at the end of the interval is given by the *model integration* from the solution $\mathbf{x}(t_n) = M_0\ [\mathbf{x}(t_0)]$. Thus, the model is used as *a strong constraint*, i.e., the analysis solution has to satisfy the model equations. In other words, 4D-VAR seeks an initial condition such that the forecast best fits the observations within the assimilation interval. The fact that the 4D-VAR method assumes a perfect model is a disadvantage since, for example, it will give the same credence to older observations at the beginning of the interval as to newer observations at the end of the interval. Derber (1989) suggested a method of correcting for a constant model error (a constant shape within the assimilation interval).

In order to minimize the cost function, the gradient of the cost function with respect to the background and observation components can be given by:

$$\frac{\partial J_b}{\partial \mathbf{x}(t_0)} = \mathbf{B}_0^{-1}\left[\mathbf{x}(t_0) - \mathbf{x}^b(t_0)\right] \tag{45}$$

$$\left[\frac{\partial J_o}{\partial \mathbf{x}(t_0)}\right] = \sum_{i=0}^{N} \mathbf{L}(t_i, t_0)^T \mathbf{H}_i^T \mathbf{R}_i^{-1}\left[H(\mathbf{x}_i) - \mathbf{y}_i^o\right] \tag{46}$$

Equation (46) shows that the 4D-VAR minimization requires the computation of the gradient, i.e., computing the increments $[H\,(\mathbf{x}_i) - \mathbf{y}_i{}^o]$ at the observation times t_i during a forward integration, multiplying them by $\mathbf{H_i}^T\,\mathbf{R_i}^{-1}$, and integrating these weighted increments back to the initial time using the adjoint model. Since parts of the backward adjoint integration are common to several time intervals, the summation in (46) can be arranged more conveniently. Assume, for example, that the interval of assimilation is from 00 to 12 Z and that there are observations every 3 h. We compute during the forward integration the weighted negative observation increments $\overline{\mathbf{d}_i} = \mathbf{H_i^T}\mathbf{R_i}^{-1}[H(\mathbf{x}_i) - \mathbf{y}^o] = \mathbf{H_i^T}\mathbf{R_i}^{-1}\mathbf{d}_i$. The adjoint model $\mathbf{L}^T(t_i, t_{i-1}) = \mathbf{L}^T$ applied on a vector "advances" it from t_i to t_{i-1}. Then we can write (46) as:

$$\frac{\partial J_o}{\partial \mathbf{x}_o} = \bar{\mathbf{d}}_o + \mathbf{L}_0^T\left\{\bar{\mathbf{d}}_1 + \mathbf{L}_1^T\left[\bar{\mathbf{d}}_2 + \mathbf{L}_2^T\left(\bar{\mathbf{d}}_3 + \mathbf{L}_3^T\bar{\mathbf{d}}_4\right)\right]\right\} \tag{47}$$

From (45) plus (46) or (47), we obtain the gradient of the cost function, and the minimization algorithm modifies appropriately the control variable $\mathbf{x}(t_0)$. After this change, a new forward integration and new observational increments are computed, and the process is repeated.

The most important advantage of 4D-VAR is that if we assume that (a) the model is perfect and (b) the a priori error covariance at the initial time $\mathbf{B}_0$ is correct, *it can be shown that the 4D-VAR analysis at the final time is identical to that of the extended Kalman filter* (Lorenc 1986; Daley 1991). This means that *implicitly 4D-VAR is able to evolve the forecast error covariance from* $\mathbf{B}_0$ *to the final time*. Unfortunately, this implicit covariance is not available at the end of the cycle and neither is the new analysis error covariance. In other words, 4D-VAR is able to find the best linear unbiased estimation but not its error covariance, except in an approximation form.

Meanwhile, in 3D-VAR, the background term is defined statistically and is static with time. Therefore, neither 3D-VAR nor 4D-VAR can represent the follow-dependent background (model forecast) error covariance in the data assimilation.

3.2.3 Ensemble Kalman Filter

In a stochastic ensemble Kalman filter, an ensemble of K data assimilation cycles is carried out simultaneously (Evensen 1994). All the cycles assimilate the same real observations, but different sets of *random perturbations have to be added to the observations* assimilated in each member of the ensemble data assimilations. Another type of ensemble Kalman filter is known as deterministic or square-root filter (Tippett et al. 2003; Ott et al. 2004). Here there is a single data assimilation, and the new analysis perturbations are obtained from the background (forecast) ensemble perturbations using a square-root algorithm.

After completing the ensemble of analyses at time t_{i-1}, and the K forecasts $\mathbf{x}^f\,(t_i) = M^k\,[\mathbf{x}^a(t_{i-1})]$, one can obtain an estimate of the forecast error covariance from the K forecasts $\mathbf{x}_k^f\,(t_i)$, as:

$$\mathbf{P}^f \approx \frac{1}{K-1}\sum_{k=1}^{K}\left(\mathbf{x}_k^f - \overline{\mathbf{x}}^f\right)\left(\mathbf{x}_k^f - \overline{\mathbf{x}}^f\right)^T \tag{48}$$

where the overbar represents the ensemble average. This tends to underestimate the variance of the forecast errors due to nonlinearities. Thus, an inflation of covariance is usually implemented with the ensemble Kalman filters.

3.2.4 Hybrid 3D-VAR/4D-VAR and Ensemble Kalman Filter

Hamill and Snyder (2000) also suggested a hybrid between 3D-VAR and ensemble Kalman filtering, where the forecast error covariance is obtained from a linear combination of the (constant) 3D-VAR covariance $\mathbf{B}_{3D\text{-Var}}$:

$$P_l^{f\,(\text{hybrid})} = (1-\alpha)P_l^f + \alpha\mathbf{B}_{3D\text{-Var}} \tag{49}$$

where α is a tunable parameter that varies from 0 for pure 3D-VAR and pure ensemble Kalman filtering from (49) to (1). Since the ensemble Kalman filtering covariance is estimated from only a limited sample of ensemble members, its rank is $K - 1$, much smaller than the number of degrees of freedom of the model, so that it is rank deficient. The combination with 3D-VAR, computed from many estimated forecast errors (using, e.g., the method of Parrish and Derber 1992), may ameliorate this sampling problem and "fill out" the error covariance. In the experiments of Hamill and Snyder (2000), the best results were obtained for low values of α, between 0.1 and 0.4, indicating a good impact of the use of the ensemble-evolved forecast error covariance. They found that 25–50 ensemble members were enough to provide the benefit of ensemble Kalman filtering (but this may be different when using a more complex model than the quasi-geostrophic model used here). Recently, both hybrid 3D-VAR and 4D-VAR and ensemble Kalman filter scheme (refer to

3dEnVar and 4dEnVar, respectively) have been implemented and run at NCEP (Wang et al. 2013; Kleist and Ide 2015) with major improvements compared with the previous 3D-VAR analysis scheme, namely the Gridpoint Statistical Interpolation (GSI) system. A different hybrid (Penny 2014) combines the Kalman gain rather than the covariances and has been tested with very good results by ECMWF (Hamrud et al. 2015).

The ensemble Kalman filtering approach has several advantages: (a) K is of the order of 10–100, so that the computational cost (compared with OI or 3D-VAR) is increased by a factor of 10–100. Although this increased cost may seem large, it is small compared to extended Kalman filtering, which requires a cost increase on the order of the number of degrees of freedom of the model. (b) Ensemble Kalman filtering does not require the development of a linear and adjoint model. (c) It does not require the linearization of the evolution of the forecast error covariance. (d) It can provide excellent initial perturbations for ensemble forecasting. Ensemble Kalman filtering appears at the present time to be one of the most promising approaches for the future (Houtekamer and Zhang 2016).

4 Recent Developments and Challenges

Along with an increase in computer power, advances in science, and demands from various applications, NWP has not only become a major forecasting tool but is also active in research and application (Bauer et al. 2015). In recent years, notable developments have been made in several areas with challenges at the same time.

First, as of today (2016), convection-permitting and cloud-resolving scale modeling have become practically feasible, along with the successful usage of large eddy simulation in developing subgrid-scale parameterizations for these models. Many national hydrometeorology centers are now running regional models in the 2–5 km grid size range and will be increasing resolution at a steady rate such that several centers may be around 1 km in next a few years. Physical parameterizations face a challenge to deal with so-called gray zones, in which the explicit model dynamics is almost capable of resolving features that were parameterized at the coarser scale. Nevertheless, one might anticipate that with increasing resolution, the need of parameterization would be gradually reduced. For radiation and cloud processes and land-surface models, this is a matter of moving current schemes toward fully explicit models. For convection, the situation is more complicated because large tropical convective clouds or organized convection occurs even at currently resolved scale (~15 km), while embedded small-scale convective plumes may not be resolved even at 1 km and will still require parameterization (Hong and Dudhia 2012; Bauer et al. 2015). In addition, more physical and chemical processes will be added into the NWP models. Additional physical processes will be needed to represent the coupling of atmosphere with ocean, land-surface, and sea ice models. Thus, studies on developing physical parameterization schemes, stochastically physical parameterizations (e.g., Palmer and Williams 2008), and super-parameterizations (e.g., Khairoutdinov et al. 2005) will remain active areas.

Second, using more of the existing and new observations, and advances in data assimilation, poses more science challenges for NWP. How to better utilize the available conventional surface observations and satellite and radar observations still remain some challenges. Developing more feasible data assimilation in the "big data" era is still needed (Miyoshi et al. 2016). In addition, NWP is also limited by insufficient observational data. Beyond the maintenance of the backbone satellite and ground-based observing systems, measurements of vertical profiles of temperature, moisture, clouds, and near-surface weather, fundamental observations are missing. Moreover, coupled data assimilation will become critical for the initialization of the future coupled models (Brunet et al. 2010). The assimilation will need to include atmospheric composition (aerosols, trace gases) as well as ocean, land surface, and sea ice.

In light of the challenges in both physical parameterization and data assimilation, it is anticipated that the ensemble forecasting will remain as the mainstream of the future developments in NWP.

References

A. Arakawa, C. S. Konor, Vertical differencing of the primitive equations based on the Charney–Phillips grid in hybrid $\sigma - p$ vertical co-ordinates. Mon. Wea. Rev. **124**, 511–528 (1996)

A. Arakawa, Adjustment mechanisms in atmospheric motions. J. Meteorol. Soc. Jpn. **75**, 155–179 (1997)

A. Arakawa, The cumulus parameterization problem: past, present, and future. J. Clim. **17**, 2493–2525 (2004)

P. Bauer, A. Thorpe, G. Brunet, The quiet revolution of numerical weather prediction. Nature **525**, 47–55 (2015)

V. Bjerknes, Das Problem der Wettervorhersage betrachtet vomStandpunkt der Mechanik und Physik. Meteorol. Z. **21**, 1–7 (1904)

F. Bouttier, F. Rabier, The operational implementation of 4D-Var. ECMWF Newsl. **78**, 2–5 (1997)

G. Brunet et al., Collaboration of the weather and climate communities to advance subseasonal-to-seasonal prediction. Bull. Am. Meteorol. Soc. **91**, 1397–1406 (2010)

J.G. Charney, R. Fjoertoft, J.V. Neumann, Numerical integration of the barotropic vorticity equation. Tellus **2**, 237–254 (1950)

F. Chen, J. Dudhia, Coupling an advanced land-surface/hydrology model with the Penn State/NCAR MM5 modeling system. Part I: model description and implementation. Mon. Weather Rev. **129**, 569–585 (2001)

P. Courtier, O. Talagrand, Variational assimilation of meteorological observations with the adjoint vorticity equations, Part II, numerical results. Quart. J. Roy. Meteor. Soc. **113**, 1329–1347 (1987)

J. Derber, A variational continuous assimilation technique. Mon. Weather Rev. **117**, 2437–2446 (1989)

R. Daley, *Atmospheric Data Analysis* (Cambridge University Press, Cambridge, 1991)

D.R. Durran, *Numerical Methods for Wave Equations in Geophysical Fluid Dynamics* (Springer, New York, 1999)

M.B. Ek, K.E. Mitchell, Y. Lin, E. Rogers, P. Grunmann, V. Koren, G. Gayno, J.D. Tarpley, Implementation of Noah land surface model advances in the National Centers for Environmental Prediction operational mesoscale Eta model. J. Geophys. Res. **22**, 8851 (2003)

G. Evensen, Sequential data assimilation with a nonlinear quasigeostrophic model using Monte Carlo methods to forecast error statistics. J. Geophys. Res. **99**, 10143–10162 (1994)

T.M. Hamill, C. Snyder, A hybrid ensemble Kalman filter-3D variational analysis scheme. Mon. Weather Rev. **128**, 2905–2919 (2000)

M. Hamrud, M. Bonavita, L. Isaksen, EnKF and hybrid gain ensemble data assimilation. Part I: EnKF implementation. Mon. Weather Rev. **143**, 4847–4864 (2015)

S.Y. Hong, Dudhia, Next-generation numerical weather prediction: bridging parameterization, explicit clouds, and large eddies. Bull. Am. Meteorol. Soc. **93**, ES6–ES9 (2012)

R.M. Hodur, The naval research laboratory's coupled ocean/atmosphere mesoscale prediction system (COAMPS). Mon. Weather Rev. **125**, 1414–1430 (1997)

J. Holton, An introduction to dynamic meteorology. Fourth edition. (Elsevier Academic Press, 2004)

P.L. Houtekamer, F. Zhang, Review of the ensemble Kalman filter for atmospheric data assimilation. Mon. Weather Rev. **144**, 4489–4452 (2016)

R.A. Houze Jr., *Cloud Dynamics* (Academic, London, 1993)

P.A.E.M. Janssen, *The Interaction of Ocean Waves and Wind* (Cambridge University Press, Cambridge, UK, 2004)

H.M. Juang, M. Kanamitsu, The NMC nested regional spectral model. Mon. Weather Rev. **122**, 3–26 (1994)

E. Kalnay, *Atmospheric Modeling, Data Assimilation, and Predictability* (Cambridge University Press, 2003)

M.F. Khairoutdinov, D.A. Randall, C. DeMott, Simulations of the atmospheric general circulation using a cloud-resolving model as a super- parameterization of physical processes. J. Atmos. Sci. **62**, 2136–2154 (2005)

D.T. Kleist, K. Ide, An OSSE-based evaluation of hybrid variational-ensemble data assimilation for the NCEP GFS. Part II: 4DEnVar and hybrid variants. Mon. Weather Rev. **143**, 452–470 (2015)

S.-J. Lin, A finite-volume integration method for computing pressure gradient forces in general vertical coordinates. Q. J. R. Meteorol. Soc. **13**, 1749–1762 (1997)

S.J. Lin, R.B. Rood, Multidimensional flux-form semi-Lagrangian transport scheme. Mon. Weather Rev. **124**, 2046–2070 (1996)

K.-N. Liou, *An Introduction to Atmospheric Radiation* (Academic, London, 1980)

A.C. Lorenc, Analysis methods for numerical weather prediction. Q. J. R. Meteorol. Soc. **112**, 1177–1194 (1986)

P. Lynch, The origins of computer weather prediction and climate modeling. J. Comput. Phys. **227**, 3431–3444 (2008)

T. Miyoshi et al., "Big Data Assimilation" revolutionizing severe weather prediction. Bull. Am. Meteorol. Soc. **97**, 1347–1354 (2016)

E. Ott et al., A local ensemble Kalman filter for atmospheric data assimilation. Tellus **56A**, 415–428 (2004)

D.F. Parrish, J.C. Derber, The National Meteorological Center's spectral statistical interpolation analysis system. Mon. Weather Rev. **120**, 1747–1763 (1992)

S.G. Penny, The hybrid local ensemble transform Kalman filter. Mon. Weather Rev. **142**, 2139–2149 (2014)

T.N. Palmer, P.D. Williams, Introduction: stochastic physics and climate modelling. Phil. Trans. R. Soc. A **366**, 2421–2427 (2008)

R.A. Pielke Sr., *Mesoscale meteorological modelling*. Second edition (Academic Press, 2002)

L.F. Richardson, *Weather Prediction by Numerical Process* (Cambridge University Press, Cambridge, UK, 1922)

A.J. Robert, A semi-Lagrangian and semi-implicit numerical integration scheme for the primitive meteorological equations. J. Meteorol. Soc. Jpn. **60**, 319–324 (1982)

A.J. Simmons, A. Hollingsworth, Some aspects of the improvement in skill of numerical weather prediction. Q. J. R. Meteorol. Soc. **128**, 647–677 (2002)

W.C. Skamarock, J.B. Klemp, J. Dudhia, D.O. Gill, M. Barker, K.G. Duda, X.Y. Huang, W. Wang, J.G. Powers, A description of the advanced research WRF version 3. NCAR Tech. Note, NCAR/TN-475+STR, 113 pp. (2008)

D.J. Stensrud, *Parameterization Schemes: Keys to Understanding Numerical Weather Prediction Models* (Cambridge University Press, Cambridge, UK, 2007)

G.L. Stephens, The parameterization of radiation for numerical weather prediction and climate models. Mon. Weather Rev. **112**, 826–867 (1984)

J.M. Straka, *Cloud and Precipitation Microphysics: Principles and Parameterization* (Cambridge University Press, Cambridge, UK, 2009)

R.B. Stull, *An Introduction to Boundary Layer Meteorology* (Kluwer Academic Publishers, Dordrecht, 1988)

O. Talagrand, Assimilation of observations, an introduction. J. Met. Soc. Jpn. Spec. Issue **75**(1B), 191–209 (1997)

M. Teixeira, The physics of orographic gravity wave drag. Front. Phys. **2**, 43 (2014)

M. Tiedtke, The general problem of parameterization. ECMWF Lecture Note (1984), http://www.ecmwf.int/en/learning/education-material/introductory-lectures-nwp

M.K. Tippett, J.L. Anderson, C.H. Bishop, T.M. Hamill, J.S. Whitaker, Ensemble square-root filters. Mon. Weather Rev. **131**, 1485–1490 (2003)

H.L. Tolman, User manual and system documentation of WAVEWATCH III version 4.18. NOAA/NWS/NCEP/MMAB Technical Note 316, 194 pp. (2014)

X. Wang, D. Parrish, D. Kleist, J. Whitaker, GSI 3DVarbased ensemble-variational hybrid data assimilation for NCEP global forecast system: single-resolution experiments. Mon. Weather Rev. **141**, 4098–4117 (2013)

T. Warner, *Numerical Weather and Climate Prediction* (Cambridge Press, Cambridge, UK, 2011)

D.L. Williamson, The evolution of dynamical cores for global atmospheric models. J. Meteorol. Soc. Jpn. B **85**, 241–269 (2007)