1 Half a century of numerical weather prediction

1.1 Introduction

Numerical weather prediction (NWP) is a very young discipline that developed essentially in the second half of the twentieth century with the continual benefit of advances in computing. The techniques implemented are used to solve equations describing the behaviour of the atmosphere, that is, to numerically compute future values of the atmosphere's characteristic parameters from initial values that are known from meteorological observations.

The equations used are the general equations of fluid mechanics that were already well established by the early twentieth century and to which certain simplifications are applied. Those simplifications are justified by the orders of magnitude of the various terms in the specific instance of the Earth's atmosphere and by the scales to be described. Computers are essential for solving these systems of nonlinear equations, which, in the general case, cannot be solved analytically.

A *numerical model* of the atmosphere is constructed in two separate stages: first, a system of equations is established to govern the continuous behaviour of the atmosphere; then, by the process of *discretization*, the equations relating to continuous variables are replaced by equations relating to discrete variables, the solutions to which are obtained by an appropriate algorithm. The results of a numerical prediction (that is, the solutions of discretized equations of dynamic meteorology) depend therefore on the discretization process employed.

Implementing this algorithm requires a sufficiently powerful computing tool. This is why advances in numerical weather prediction have followed in the wake of the fantastic development of electronic computers since they came into being at the end of the Second World War.

And last, it should be emphasized that weather forecasting achieved by forecasters using numerical models owes its success to the implementation of the global weather observing system that relies on both conventional and satellite measurements and provides an admittedly imperfect but nonetheless effective description of the atmosphere at a given initial time.

1.2 The early days

The history of numerical weather prediction features a number of stages that proved decisive in the development of the discipline.

Back in 1904, the Norwegian Vilhelm Bjerknes recognized that weather forecasting is fundamentally a deterministic *initial-value* problem in the mathematical sense (Bjerknes, 1904):

> *If it is true, as every scientist believes, that subsequent atmospheric states develop from the preceding ones according to physical law, then it is apparent that the necessary and sufficient conditions for the rational solution of forecasting problems are the following*:
>
> – *A sufficiently accurate knowledge of the state of the atmosphere at the initial time.*
> – *A sufficiently accurate knowledge of the laws according to which one state of the atmosphere develops from another.*

However, he realized that the difficulty lay in the need to solve a system of nonlinear partial differential equations for which there were no analytical solutions, in the general case.

Between 1916 and 1922, the Englishman Lewis Fry Richardson tried to solve weather forecast equations by numerical methods. He even made a 6-hour forecast by hand, although it proved quite unrealistic. Undaunted, though, he sought out the causes of his failure. His work was published in 1922 in his famous and truly visionary *Weather Prediction by Numerical Process* (Richardson, 1922). Noting that '$32 \times 2000 = 64{,}000$ *computers would be needed to race the weather for the whole globe*', Richardson let his imagination roam and dreamed of a weather-forecasting factory, with a myriad of people making synchronized computations under the control of a supervisor tasked with the orchestration of operations (Figure 1.1).

In 1928, the German mathematicians Courant, Friedrichs, and Lewy systematically studied how to solve partial derivative equations by using finite differences and specified the constraints to comply with when performing discretization (Courant et al., 1928).

In 1939, the Swede Carl-Gustav Rossby showed that the *absolute vorticity conservation equation* provided a correct interpretation of the observed displacement of atmospheric centres of action (Rossby, 1939).

In 1946, the first electronic computer, the ENIAC (Electronic Numerical Integrator and Computer) was installed at Pennsylvania University, in Philadelphia, while the Hungarian-born U.S. mathematician John von Neumann was also working on building improved machines at the Institute for Advanced Studies in Princeton.

In 1948, the American Jule Charney proposed a simplification of the general system of equations, known as the *quasi-geostrophic approximation*, and found, as a specific instance, the equation studied by Rossby (Charney, 1948).

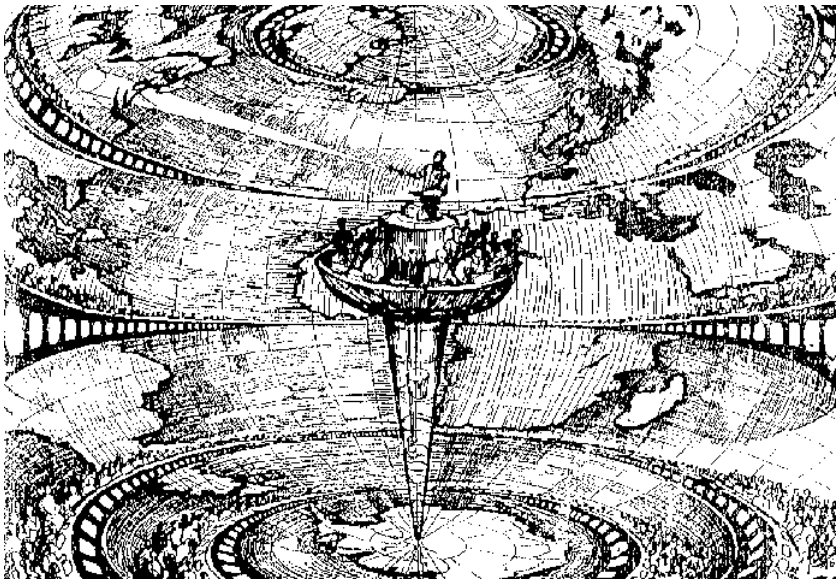

Figure 1.1 Richardson's 'dream.' (Artist's impression by Alf Lannerbaeck, published by the Swedish newspaper *Dagens Nyheter*, 22 September 1984)

Finally, in 1950 Jule Charney, the Norwegian Ragnar Fjörtoft, and John von Neumann made the first numerical weather prediction (Charney et al., 1950): they used the absolute vorticity conservation equation for this experiment and did the computing on the ENIAC at Aberdeen (Maryland). The results obtained for the forecast of geopotential height of the 500 hPa isobaric surface, characteristic of the middle atmosphere, were most encouraging, and the experiment marked the starting point of modern numerical prediction (Platzman, 1979). In answer to Charney, who had sent the paper describing the experiment to him, Richardson wrote in 1952: '*Allow me to congratulate you and your collaborators on the remarkable progress which has been made in Princeton; and on the prospects of further improvement which you indicate to establish a science of meteorology, with the aim of predicting future states of the atmosphere from the present state*' (Ashford, 1985).

1.3 Half a century of continual progress

The success of Charney, Fjörtoft, and von Neumann's experiment was to lead from the mid 1950s onwards to the development for operational purposes of a large number of increasingly complex prediction models of ever greater spatial resolution, allowing ever smaller scales to be covered.

1.3.1 The need to be fast and accurate

Richardson had fully understood that numerical weather prediction was a race between the computing process and the actual evolution of the atmosphere. The speed of computation depends on the various characteristics of the prediction model and on the speed of the computer used, in a form we shall examine in detail.

Suppose that the equations are discretized by dividing space into boxes defined by a horizontal grid and a number of vertical levels. Within each box, the atmosphere is assumed to be homogeneous and so it suffices to know the values of the various atmospheric quantities at some point within the box. The time required to make a prediction for a given time range can then be calculated by taking account of the various factors involved:

- The total number N_v of variables to be processed. The state of the atmosphere being described by a limited number of quantities (the two components of horizontal wind, temperature, specific humidity, and surface pressure), the number of variables is equal to the product of that number of parameters by the total number of points processed, which varies with the size of the geographical domain and the spatial resolution adopted horizontally and vertically.
- The number of calculations N_c to be made per variable for a time step Δt. This number of elementary arithmetical operations depends on the complexity of the model, with greater allowance for interactions among variables being reflected by an increased number of calculations.
- The number of time steps N_t needed to reach a given time range H, namely, $N_t = H/\Delta t$. This time step Δt depends on the spatial resolution characterized by the mesh size Δx of the grid, for it must satisfy the Courant, Friedrichs, and Lewy (CFL) condition that is expressed as:

$$U\Delta t/\Delta x<C,$$

  where U is the speed of propagation of the fastest waves described by the equations and C is a dimensionless number dependent upon the problem geometry and the chosen discretization. While some algorithms, to be discussed later, can take us beyond this limit, it is nonetheless true that the time step Δt must be reduced concomitantly with the mesh size Δx to process the space and time scales of mesoscale atmospheric phenomena with similar accuracy, as shown by examination of Table 1.1.
- The computer's calculating speed R. This is expressed as the number of elementary floating point operations per second, or *flops*, whether done by one computer or several computers in parallel.

The time T required to make a prediction for a given time range H is given by the ratio:

$$T = N_v N_c N_t / R.$$

Table 1.1 The different meteorological phenomena with their respective time and space scales. (After Orlanski (1975). *Bull. Am. Meteor. Soc.*, **56**, 528 © Amer. Met. Soc.)

CLASSIFICATION OF SCALES	Time	Climatological scale		Planetary and synoptic scales	Meso-scale	Micro-scale	
Length	T / L	1 month		1 day	1 hour	1 minute	1 sec.
Macro α scale	10 000 km	Standing waves	Ultra-long waves	Tidal waves			
Macro β scale	2 000 km			Barodinic waves			
Meso α scale	200 km			Fronts and Hurricanes			
Meso β scale	20 km				Nocturnal low level jet Squall lines Inertial waves Cloud clusters Mtn and lake disturbances		
Meso γ scale	2 km				Thunderstorms Internal gravity waves Clear air turbulence Urban effects		
Micro α scale	200 m					Tornadoes Deep convection Short gravity waves	
Micro β scale	20 m					Dust devils Thermals Wakes	
Micro γ scale							Plumes Roughness Turbulence

To take the example of the ARPEGE operational model used by Météo-France in 1998, the number of variables to be processed was $N_v \approx 23.10^6$ (600×300 horizontal points, 31 levels, 4 three-dimensional variables, and 1 two-dimensional variable), the number of calculations to be made for one variable was $N_c \approx 7.10^3$, and the number of time steps for a 24-hour forecast was $N_t = 96$ (15-minute time steps). The calculations were made on a FUJI VPP700 multiprocessor computer credited with a computational speed of up to 20 gigaflops (20 billion floating point operations per second) and the time required for a 24-hour forecast was just under a quarter of an hour.

As the time T is imposed by operational constraints, any increase in the computer's speed means the model's resolution may be augmented (horizontal grid spacing and

number of vertical levels) as may the number of calculations made for each of the variables. This evolution towards greater resolution and increased complexity has been the rule in recent decades; it has also been facilitated by the development of new algorithms allowing longer time steps.

1.3.2 The use of filtered equations

The earliest models used operationally relied on the quasi-geostrophic approximation that imposes a diagnostic (that is, time-independent) relation between the pressure field and the wind field, which reduces the number of degrees of freedom of the model. This approximation also has the effect of conserving only slow waves, known as *Rossby waves*, as solutions and eliminating rapidly propagating *inertia-gravity waves*; it thus allows us to use a comparatively large time step compatible with the CFL condition. Because of the filtering effect so obtained, the simplified equations are known as *filtered equations*. Such a three-level model (Charney, 1954) was put into service for operational forecasting in May 1955 by the U.S. Weather Bureau. However, it was not until it was improved by Cressman (1963) that forecasters could really use the tool (Shuman, 1989). In the 1960s and until the mid 1970s models with filtered equations were widely used by the leading meteorological services (Bushby, 1987; Pône, 1993; Cressman, 1996; Rochas and Javelle, 1993). Enhanced computer performances were then used to extend the domain and increase the horizontal resolution and the number of vertical levels (thereby increasing the number of variables N_v) so as to better describe the dynamics of the atmosphere.

1.3.3 Back to the primitive equations and initialization

The growing calculating speed of computers meant it was then possible to return to the equations for the evolution of a fluid in hydrostatic equilibrium used earlier by Richardson, which were from then on termed the *primitive equations*. They admit rapidly propagating inertia-gravity waves as solutions, and for compliance with the CFL condition require the choice of a time step some six times smaller than with filtered equations, thereby increasing the number of time steps N_t. Work on the primitive equations begun by Eliassen (1956) led to successful tests in the United States (Smagorinsky, 1958) and in Germany (Hinkelmann, 1959). In the United States, the primitive equation model with six vertical levels developed by Shuman and using a 381 km mesh on an octagonal domain covering most of the northern hemisphere (Shuman and Hovermale, 1968) began its operational career on 6 June 1966, thus opening up the path to generalized use of this type of model for many meteorological services.

Primitive equation models are relatively easy to implement but require that the *initialization* problem be solved. Pressure and wind fields coupled through evolution equations must respect a certain balance at the initial time; otherwise they will give rise to substantial oscillations owing to the propagation of gravity waves of unrealistic amplitudes (Hinkelmann, 1951). The difficulty in obtaining a balanced initial state from pressure and wind observations was what brought about Richardson's unrealistic result in the first attempt at numerical prediction (Lynch, 1994).

Static initialization methods whereby the wind field is deduced from the pressure field using a linear or nonlinear equation proved comparatively ineffective; moreover, wind observations were not really used then for defining the initial state. It was in the late 1970s that an elegant solution to the problem of initialization of global fields was found, independently by Baer and Tribbia (1977) and by Machenhauer (1977). The idea was to decompose the initial state of the atmosphere into normal modes (that is, into solutions of a linearized version of the model) and then to correct the inertia-gravity modes in the initial state so as to make them stationary when the model evolved. This technique of *nonlinear normal mode initialization* meant primitive equation models could be used effectively to take full advantage of initial pressure and wind data.

1.3.4 Global processing and the spectral method

Elementary reasoning based on the speed at which perturbations move shows that the working area has to be extended and so the number of points N_v increased when one wishes to make predictions over longer time ranges (Figure 1.2).

The models for limited geographical areas were replaced by hemispheric models, and then finally by global models allowing interactions between the two hemispheres to be handled properly. This meant grids had to be defined on a sphere and the problem of instability resulting from smaller mesh size close to the poles had to be solved.

Alongside grid point models using the finite difference method for computing partial derivatives, the use of spectral models has developed. In these, fields defined on the sphere are represented by series expansion in terms of basis functions: the surface

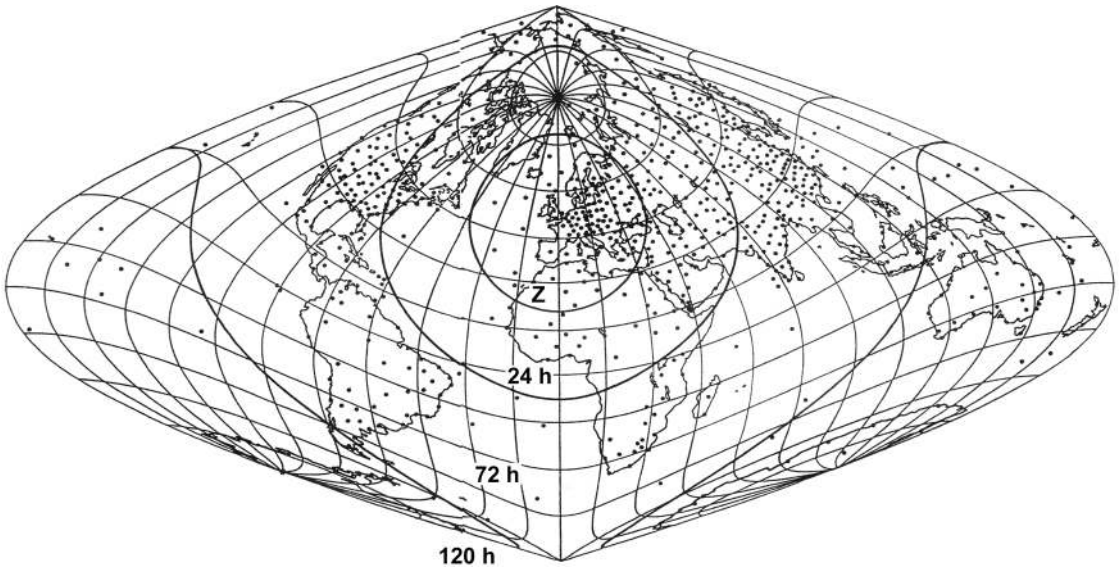

Figure 1.2 Worldwide distribution of radiosounding stations and indication of regions for which observations are required for making 1-, 3-, and 5-day forecasts over the central area Z. (ECMWF image)

spherical harmonics. This method allows better evaluation of wave speeds than by the finite difference method. It had long been reserved for models with just a few degrees of freedom, because of the high cost of direct computation of the expansion coefficients for nonlinear terms. With the advent of the fast Fourier transform (Cooley and Tukey, 1965), it proved far more advantageous to use the *transform method*, consisting in calculating the nonlinear terms at the nodes of an intermediate grid (Orszag, 1970; Eliassen et al., 1970). This technique made the spectral method highly competitive (Bourke, 1972), and in the 1980s it superseded the grid point method almost everywhere for producing global models.

1.3.5 Limited area models

In addition to extending the area, which was necessary to extend the time range of forecasts, it turned out to be advantageous for short-range forecasting (1–2 days) to continue working on a restricted area using a grid with a fine enough mesh to simulate small-scale motion correctly and reproduce features caused by topography. Thus, limited area models (LAM) were developed enabling short-range, small-scale predictions to be made (Rousseau et al., 1995). Mathematical analysis shows that field values have to be specified on the area boundary for each time step. The field values may be obtained by interpolating fields from a larger scale model. However, a dissipation term should be introduced into the limited area model to damp perturbations that are engendered by forcing fields at the boundary and that propagate towards the interior of the domain (Davies, 1976). This leads to the *nested models* that are the basis of operational prediction systems in most meteorological services.

As the spectral method and normal mode nonlinear initialization had proved effective for global models, it was tempting to apply the same techniques to models for limited areas. Among the various approaches proposed was that of Machenhauer and Haugen (1987), consisting in extending the fields over a larger domain so as to make them doubly periodic; this artefact allows the spectral method to be used on a limited area by performing the series expansion in terms of trigonometric functions.

As for the normal mode nonlinear initialization method, it is possible, under certain assumptions about the definition of the linearized part of the model, to stationarize inertia-gravity modes in physical space (Brière, 1982; Juvanon du Vachat, 1986). This process was successfully applied for initializing primitive equation limited area models. Subsequently a method of digital filtering of high frequencies corresponding to inertia-gravity waves was proposed by Lynch and Huang (1992). It provides satisfactory solutions to the initialization problem for limited area models and for models whose geometry makes it impossible to determine any normal modes.

1.3.6 Algorithms for an increased time step

The use of explicit time integration schemes with primitive equations requires the use of time steps six times smaller than those of the filtered models just to satisfy the

CFL condition. Robert (1969) came up with a fresh approach, proposing to process the terms responsible for gravity wave propagation implicitly. This *semi-implicit time integration* algorithm yields a new, much less restrictive CFL condition as it involves only the maximum speed of the synoptic wind and no longer the speeds of the fastest waves. This possibility of increasing the time step has its downside, as a system of linear equations then needs to be solved. Despite this, the semi-implicit algorithm maintains a clear lead, allowing the run time for grid point models to be divided fourfold and by even more for spectral models. This explains why it has been so popular and become so widespread since the 1970s.

Lagrangian processing of advection was initially used by Fjörtoft (1952) to solve a simple model graphically. The method then inspired Lepas (1963) in constructing a numerical prediction model. Krishnamurti (1962) and Sawyer (1963) also proposed using the technique to improve the accuracy of numerical advection schemes. However, the credit goes again to Robert (1981) for showing that the method used in conjunction with semi-implicit processing could free us from the CFL condition. Time discretization is performed on the total derivative (or *Lagrangian derivative*) and forces us to interpolate the model variables at the starting point of particles arriving at the grid points during their movement in one time step. We thus obtain the *semi-Lagrangian semi-implicit scheme* algorithm allowing us to further increase the time step Δt (and so reduce the number of time steps N_t) within the limits compatible with the required accuracy for representing the relevant time scales.

It should also be emphasized that this highly effective algorithm has also made the use of variable grid models (that is, with increasing resolution over a chosen area) into a competitive solution for the nested model system once the time step is no longer dependent on the smallest grid dimension in the working domain (Courtier and Geleyn, 1988; Côté et al., 1993).

1.3.7 The move to nonhydrostatic equations

The semi-Lagrangian semi-implicit scheme opened up new horizons for nonhydrostatic models that are essential for correctly handling spatial scales of the order of 1 km. Their operational implementation had until then come up against the problem of the very small time step arising from the need to comply with the CFL condition relative to the propagation of sound waves (also known as acoustic waves). However, Lagrangian processing of advection combined with implicit processing of the terms responsible for the propagation of gravity waves and sound waves leads to an unconditionally stable algorithm so that it is now possible to envisage using nonhydrostatic models (Tanguay et al., 1990; Laprise, 1992; Bubnova et al., 1995) to simulate atmospheric motion from the planetary scale to mesoscale (Table 1.1).

1.3.8 Physical processes

It soon proved necessary to evaluate the source and sink terms of momentum, heat, and water vapour resulting from more or less complex physical processes that have

to be introduced into equations to reproduce the evolution of the atmosphere realistically. Given that the scales to be taken into account to accurately simulate the relevant physical processes are generally smaller than the scales described by the model variables (these are sometimes referred to as *subgrid* scales), these processes have to be parameterized: their average effect on the model variables alone is sought. These additional computations form the model *physics* and are grafted onto the numerical processing of equations, which is the model *dynamics*.

After allowing in a simple way for the effects of friction to avoid depressions deepening excessively, a real improvement was made in describing the atmospheric water cycle and its associated energy exchanges. The addition of an extra equation describing the transport of water vapour is required to have the means for handling the effects of changes in water phases and for calculating precipitation (Smagorinsky, 1962).

Proper description of the turbulent transfer mechanisms between the soil and the atmosphere, not just for momentum but also for sensible heat and water vapour, implies calculating turbulent fluxes near the surface (Businger et al., 1971; Deardorff, 1972; Louis, 1979). This calculation involves additional variables, apart from dynamic model variables calculated for the lowermost level, such as surface temperature and moisture as well as data characterizing the soil such as roughness length or plant cover (Deardorff, 1977).

It is obvious that the evolution of surface variables is directly related to energy inputs from radiation flux, which in turn is highly dependent upon the time of day and cloud cover. This is why it is essential to calculate the effects of interaction between radiation and the various constituents of the atmosphere, especially the water present in its various phases. The effects of absorption, scattering, and re-emission of radiation, which differ particularly depending on whether the atmosphere is clear or cloudy, must be computed (Rodgers and Walshaw, 1966; Katayama, 1974).

Because of the hydrostatic hypothesis, the primitive equations cannot deal explicitly with convective motion resulting from local vertical instability of the atmosphere. The *convective adjustment* methods (Manabe and Strickler, 1964), which were designed to correct the vertical profiles leading to unstable solutions for the model, have been superseded by more elaborate methods that account for the effects of interaction between convective clouds and their environment (Kuo, 1965, 1974; Arakawa and Schubert, 1974; Bougeault, 1985).

The comparatively recent inclusion of energy dissipation of the vertically propagating mountain waves has also improved the prediction of the intensity of jet streams above mountainous regions (Palmer et al., 1986).

1.3.9 Objective analysis and data assimilation

Alongside the improvement made to the forecasting models, very important theoretical and practical work has also been done in precisely determining a given state of the atmosphere allowing for the various observations available (Daley, 1980). This