Atmospheric modeling, data assimilation and predictability

Eugenia Kalnay
University of Maryland

CAMBRIDGE
UNIVERSITY PRESS

PUBLISHED BY THE PRESS SYNDICATE OF THE UNIVERSITY OF CAMBRIDGE
The Pitt Building, Trumpington Street, Cambridge, United Kingdom

CAMBRIDGE UNIVERSITY PRESS
The Edinburgh Building, Cambridge CB2 2RU, UK
40 West 20th Street, New York, NY 10011-4211, USA
477 Williamstown Road, Port Melbourne, VIC 3207, Australia
Ruiz de Alarcón 13, 28014 Madrid, Spain
Dock House, The Waterfront, Cape Town 8001, South Africa

http://www.cambridge.org

First published 2003

Printed in the United Kingdom at the University Press, Cambridge

Typefaces Times Roman $10\frac{1}{4}/13\frac{1}{2}$ pt and Joanna *System* LaTeX 2_ε [TB]

A catalogue record for this book is available from the British Library

Library of Congress Cataloguing in Publication data

Kalnay, Eugenia, 1942–
Atmospheric modeling, data assimilation and predictability / Eugenia Kalnay.
p. cm.
Includes bibliographical references and index.
ISBN 0-521-79179-0 – ISBN 0-521-79629-6 (pbk.)
1. Numerical weather forecasting. I. Title.
QC996 .K35 2002 551.63′4 – dc21 2001052687

ISBN 0 521 79179 0 (hardback)
ISBN 0 521 79629 6 (paperback)

Contents

Foreword xi

Acknowledgements xv

List of abbreviations xvii

List of variables xxi

1 **Historical overview of numerical weather prediction** 1

1.1 Introduction 1
1.2 Early developments 4
1.3 Primitive equations, global and regional models, and nonhydrostatic models 10
1.4 Data assimilation: determination of the initial conditions for the computer forecasts 12
1.5 Operational NWP and the evolution of forecast skill 17
1.6 Nonhydrostatic mesoscale models 24
1.7 Weather predictability, ensemble forecasting, and seasonal to interannual prediction 25
1.8 The future 30

2 **The continuous equations** 32

2.1 Governing equations 32
2.2 Atmospheric equations of motion on spherical coordinates 36
2.3 Basic wave oscillations in the atmosphere 37
2.4 Filtering approximations 47
2.5 Shallow water equations, quasi-geostrophic filtering, and filtering of inertia-gravity waves 53
2.6 Primitive equations and vertical coordinates 60

3 **Numerical discretization of the equations of motion** 68

3.1 Classification of partial differential equations (PDEs) 68
3.2 Initial value problems: numerical solution 72
3.3 Space discretization methods 91
3.4 Boundary value problems 114
3.5 Lateral boundary conditions for regional models 120

4 **Introduction to the parameterization of subgrid-scale physical processes** 127

4.1 Introduction 127
4.2 Subgrid-scale processes and Reynolds averaging 129
4.3 Overview of model parameterizations 132

5 **Data assimilation** 136

5.1 Introduction 136
5.2 Empirical analysis schemes 140
5.3 Introduction to least squares methods 142
5.4 Multivariate statistical data assimilation methods 149
5.5 3D-Var, the physical space analysis scheme (PSAS), and their relation to OI 168
5.6 Advanced data assimilation methods with evolving forecast error covariance 175
5.7 Dynamical and physical balance in the initial conditions 185
5.8 Quality control of observations 198

6 **Atmospheric predictability and ensemble forecasting** 205

6.1 Introduction to atmospheric predictability 205
6.2 Brief review of fundamental concepts about chaotic systems 208
6.3 Tangent linear model, adjoint model, singular vectors, and Lyapunov vectors 212
6.4 Ensemble forecasting: early studies 227
6.5 Operational ensemble forecasting methods 234
6.6 Growth rate errors and the limit of predictability in mid-latitudes and in the tropics 249
6.7 The role of the oceans and land in monthly, seasonal, and interannual predictability 254
6.8 Decadal variability and climate change 258

Appendix A The early history of NWP 261

Appendix B Coding and checking the tangent linear and the adjoint models 264

Appendix C Post-processing of numerical model output to obtain station weather forecasts 276

References 283

Index 328

1

Historical overview of numerical weather prediction

1.1 Introduction

In general, the public is not aware that our daily weather forecasts start out as initial-value problems on the major national weather services supercomputers. Numerical weather prediction provides the basic guidance for weather forecasting beyond the first few hours. For example, in the USA, computer weather forecasts issued by the National Center for Environmental Prediction (NCEP) in Washington, DC, guide forecasts from the US National Weather Service (NWS). NCEP forecasts are performed by running (integrating in time) computer models of the atmosphere that can simulate, given one day's weather observations, the evolution of the atmosphere in the next few days.[1] Because the time integration of an atmospheric model is an *initial-value problem*, the ability to make a skillful forecast requires both that *the computer model be a realistic representation of the atmosphere*, and that *the initial conditions be known accurately*.

NCEP (formerly the National Meteorological Center or NMC) has performed operational computer weather forecasts since the 1950s. From 1955 to 1973, the forecasts included only the Northern Hemisphere; they have been global since 1973. Over the years, the quality of the models and methods for using atmospheric observations has improved continuously, resulting in major forecast improvements.

1 In this book we will provide many examples mostly drawn from the US operational numerical center (NCEP), because of the availability of long records, and because the author's experience in this center facilitates obtaining such examples. However, these operational NCEP examples are only given for illustration purposes, and are simply representative of the evolution of operational weather forecasting in all major operational centers.

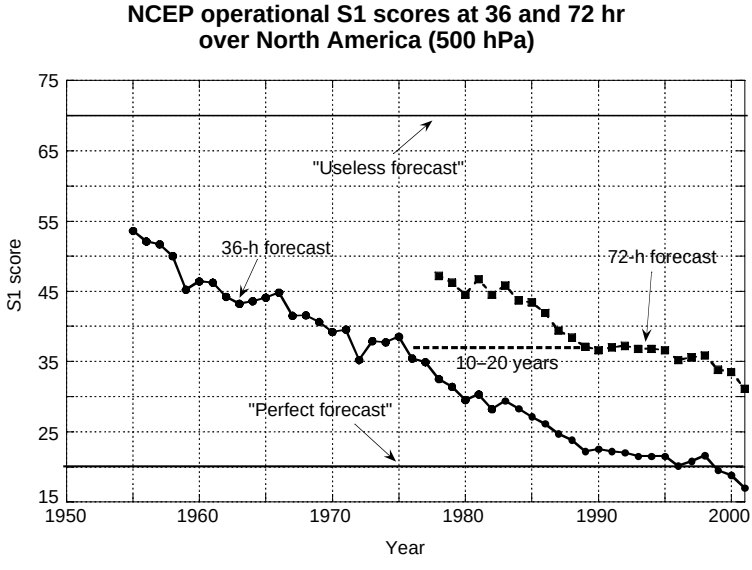
Figure 1.1.1(a) shows the longest available record of the skill of numerical weather prediction. The "*S*1" score (Teweles and Wobus, 1954) measures the relative error in the horizontal gradient of the height of the constant pressure surface of 500 hPa (in the middle of the atmosphere, since the surface pressure is about 1000 hPa) for 36-h forecasts over North America. Empirical experience at NMC indicated that a score of 70% or more corresponds to a useless forecast, and a score of 20% or less corresponds to an essentially perfect forecast. This was found from the fact that 20% was the average *S*1 score obtained when comparing analyses hand-made by several experienced forecasters fitting the same observations over the data-rich North American region.

Figure 1.1.1(a) shows that current 36-h 500-hPa forecasts over North America are close to what was considered essentially "perfect" 40 years ago: the computer forecasts are able to locate generally very well the position and intensity of the large-scale atmospheric waves, major centers of high and low pressure that determine the general evolution of the weather in the 36-h forecast. The sea level pressure forecasts contain smaller-scale atmospheric structures, such as fronts, mesoscale convective systems that dominate summer precipitation, etc., and are still difficult to forecast in detail (although their prediction has also improved very significantly over the years) so their *S*1 score is still well above 20% (Fig. 1.1.1(b)). Fig. 1.1.1(a) also shows that the 72-h forecasts of today are as accurate as the 36-h forecasts were 10–20 years ago. This doubling (or better) of skill in the forecasts is observed for other forecast variables, such as precipitation. Similarly, 5-day forecasts, which had no useful skill 15 years ago, are now moderately skillful, and during the winter of 1997–8, ensemble forecasts for the second week average showed useful skill (defined as anomaly correlation close to 60% or higher).

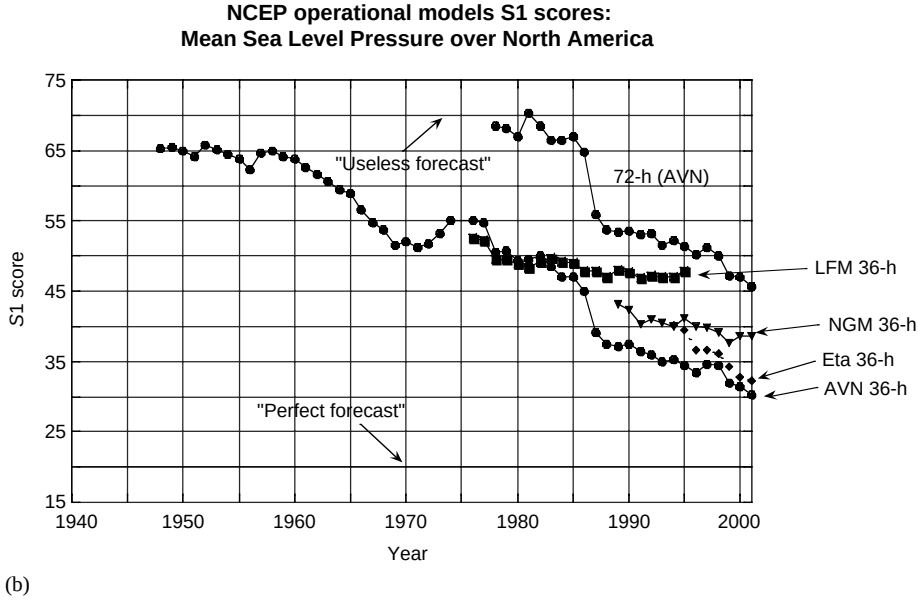
The improvement in skill of numerical weather prediction over the last 40 years apparent in Fig.1.1.1 is due to four factors:

- the increased power of supercomputers, allowing much finer numerical resolution and fewer approximations in the operational atmospheric models;
- the improved representation of small-scale physical processes (clouds, precipitation, turbulent transfers of heat, moisture, momentum, and radiation) within the models;
- the use of more accurate methods of data assimilation, which result in improved initial conditions for the models; and
- the increased availability of data, especially satellite and aircraft data over the oceans and the Southern Hemisphere.

In the USA, research on numerical weather prediction takes place in the national laboratories of the National Oceanic and Atmospheric Administration (NOAA), the National Aeronautics and Space Administration (NASA) and the National Center for Atmospheric Research (NCAR), and in universities and centers such as the



(a)



(b)

Figure 1.1.1: (a) Historic evolution of the operational forecast skill of the NCEP (formerly NMC) models over North America (500 hPa). The *S*1 score measures the relative error in the horizontal pressure gradient, averaged over the region of interest. The values $S1 = 70\%$ and $S1 = 20\%$ were empirically determined to correspond respectively to a "useless" and a "perfect" forecast when the score was designed. Note that the 72-h forecasts are currently as skillful as the 36-h were 10–20 years ago (data courtesy C.Vlcek, NCEP). (b) Same as (a) but showing *S*1 scores for sea level pressure forecasts over North America (data courtesy C.Vlcek, NCEP). It shows results from global (AVN) and regional (LFM, NGM and Eta) forecasts. The LFM model development was "frozen" in 1986 and the NGM was frozen in 1991.

Center for Prediction of Storms (CAPS). Internationally, major research takes place in large operational national and international centers (such as the European Center for Medium Range Weather Forecasts (ECMWF), NCEP, and the weather services of the UK, France, Germany, Scandinavian and other European countries, Canada, Japan, Australia, and others). In meteorology there has been a long tradition of sharing both data and research improvements, with the result that progress in the science of forecasting has taken place on many fronts, and all countries have benefited from this progress.

In this introductory chapter, we give an overview of the major components and milestones in numerical forecasting. They will be discussed in detail in the following chapters.

1.2 Early developments

Jule G. Charney (1917–1981) was one of the giants in the history of numerical weather prediction. In his 1951 paper "Dynamical forecasting by numerical process", he introduced the subject of this book as well as it could be introduced today. We reproduce here parts of the paper (with emphasis added):

> As meteorologists have long known, *the atmosphere exhibits no periodicities of the kind that enable one to predict the weather in the same way one predicts the tides*. No simple set of causal relationships can be found which relate the state of the atmosphere at one instant of time to its state at another. It was this realization that *led V. Bjerknes (1904) to define the problem of prognosis as nothing less than the integration of the equations of motion of the atmosphere*.[2] But it remained for *Richardson (1922)* to suggest the practical means for the solution of this problem. *He proposed to integrate the equations of motion numerically and showed exactly how this might be done. That the actual forecast used to test his method was unsuccessful was in no way a measure of the value of his work.* In retrospect it

2 The importance of the Bjerknes (1904) paper is clearly described by Thompson (1990), another pioneer of NWP, and the author of a very inspiring text on NWP (Thompson, 1961a). His paper "Charney and the revival of NWP" contains extremely interesting material on the history of NWP as well as on early computers:

It was not until 1904 that Vilhelm Bjerknes – in a remarkable manifesto and testament of deterministic faith – stated the central problem of NWP. This was the first explicit, coherent recognition that the future state of the atmosphere is, *in principle*, completely determined by its detailed initial state and known boundary conditions, together with Newton's equations of motion, the Boyle–Charles–Dalton equation of state, the equation of mass continuity, and the thermodynamic energy equation. Bjerknes went further: he outlined an ambitious, but logical program of observation, graphical analysis of meteorological data and graphical solution of the governing equations. He succeeded in persuading the Norwegians to support an expanded network of surface observation stations, founded the famous Bergen School of synoptic and dynamic meteorology, and ushered in the famous polar front theory of cyclone formation. Beyond providing a clear goal and a sound physical approach to dynamical weather prediction, V. Bjerknes instilled his ideas in the minds of his students and their students in Bergen and in Oslo, three of whom were later to write important chapters in the development of NWP in the US (Rossby, Eliassen and Fjörtoft).

> becomes obvious that the inadequacies of observation alone would have doomed any attempt, however well conceived, a circumstance of which Richardson was aware. The real value of his work lay in the fact that it crystallized once and for all the essential problems that would have to be faced by future workers in the field and it laid down a thorough groundwork for their solution.
>
> For a long time no one ventured to follow in Richardson's footsteps. The paucity of the observational network and the enormity of the computational task stood as apparently insurmountable barriers to the realization of his dream that one day it might be possible to advance the computation faster than the weather. But with the increase in the density and extent of the surface and upper-air observational network on the one hand, and the development of large-capacity high-speed computing machines on the other, interest has revived in Richardson's problem, and attempts have been made to attack it anew.
>
> These efforts have been characterized by a devotion to objectives more limited than Richardson's. Instead of attempting to deal with the atmosphere in all its complexity, one tries to be satisfied with *simplified models* approximating the actual motions to a greater or lesser degree. By *starting with models incorporating only what it is thought to be the most important of the atmospheric influences*, and by gradually bringing in others, one is able to proceed inductively and thereby to avoid the pitfalls inevitably encountered when a great many poorly understood factors are introduced all at once.
>
> A necessary condition for the success of this stepwise method is, of course, that the first approximations bear a recognizable resemblance to the actual motions. Fortunately, the science of meteorology has progressed to the point where one feels that at least the main factors governing the large-scale atmospheric motions are well known. *Thus integrations of even the linearized barotropic and thermally inactive baroclinic equations have yielded solutions bearing a marked resemblance to reality*. At any rate, it seems clear that the models embodying the collective experience and the positive skill of the forecast cannot fail utterly. This conviction has served as the guiding principle in the work of the meteorology project at The Institute for Advanced Study [at Princeton University] with which the writer has been connected.

As indicated by Charney, Richardson performed a remarkably comprehensive numerical integration of the full primitive equations of motion (Chapter 2). He used a horizontal grid of about 200 km, and four vertical layers of approximately 200 hPa, centered over Germany. Using the observations at 7 UTC (Universal Coordinate Time) on 20 May 1910, he computed the time derivative of the pressure in central Germany between 4 and 10 UTC. *The predicted 6-h change was 146 hPa, whereas in reality there was essentially no change observed in the surface pressure*. This huge error was discouraging, but it was due mostly to the fact that the initial conditions were *not balanced*, and therefore included fast-moving gravity waves which masked the *initial rate of change* of the meteorological signal in the forecast (Fig. 1.2.1). Moreover, if the integration had been continued, it would have suffered "computational blow-up" due to the violation of the Courant–Friedricks–Lewy (CFL) condition

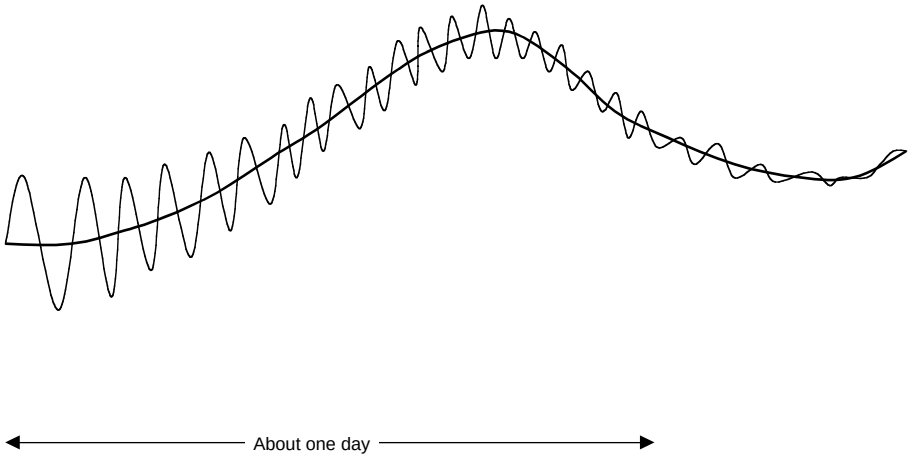


Figure 1.2.1: Schematic of a forecast with slowly varying weather-related variations and superimposed high-frequency gravity waves. Note that even though the forecast of the slow waves is essentially unaffected by the presence of gravity waves, the initial time derivative is much larger in magnitude, as obtained in the Richardson (1922) experiment.

(Chapter 3) which requires that the time step should be smaller than the grid size divided by the speed of the fastest traveling signal (in this case horizontally moving sound waves, traveling at about 300 m/s).

Charney (1948, 1949) and Eliassen (1949) solved both of these problems by deriving "filtered" equations of motion, based on quasi-geostrophic (slowly varying) balance, which filtered out (i.e., did not include) gravity and sound waves, and were based on pressure fields alone. Charney points out that this approach was justified by the fact that forecasters' experience was that they were able to predict tomorrow's weather from pressure charts alone:

> In the selection of a suitable first approximation, Richardson's discovery that the horizontal divergence was an unmeasurable quantity had to be taken into account. Here a consideration of forecasting practice gave rise to the belief that this difficulty could be surmounted: forecasts were made by means of geostrophic reasoning from the pressure field alone – forecasts in which the concept of horizontal divergence played no role.

In order to understand better Charney's comment, we quote an anecdote from Lorenz (1990) on his interactions with Jule Charney:

> On another[3] occasion when our conversations had turned closer to scientific matters, Jule was talking again about the early days of NWP. For a proper

3 The previous occasion was a story about an invitation Charney received to appear on the "Today" show, to talk about how computers were going to forecast the weather. Since the show was at 7 am, Charney, a late riser, had never watched it. "He told us that he felt that he ought to see the show at least once before agreeing to appear on it, and so, one morning, he managed to pull himself out of bed and turn on the TV set, and the first person he saw was a chimpanzee.

> perspective, we should recall that at the time when Charney was a student, pressure was king. The centers of weather activity were acknowledged to be the highs and lows. A good prognostic chart was one that had the isobars in the right locations. Naturally, then, the thing that was responsible for the weather changes was the thing that made the pressure change. This was readily shown to be the divergence of the wind field. The divergence could not be very accurately measured, and a corollary deduced by some meteorologists, including some of Charney's advisors, was that the dynamic equations could not be used to forecast the weather.
>
> Such reasoning simply did not make sense to Jule. The idea that the wind field might serve instead of the pressure field as a basis for dynamical forecasting, proposed by Rossby, gave Jule a route to follow.[4] He told us, however, that what really inspired him to develop the equations that later became the basis for NWP was a determination to prove, to those who had assured him that the task was impossible, that they were wrong.

Charney, R. Fjørtoft, and J. von Neuman (1950) computed a historic first one-day weather forecast using a barotropic (one-layer) filtered model. The work took place in 1948–9. They used one of the first electronic computers (the Electronic Numerical Integrator and Computer, ENIAC), housed at the Aberdeen Proving Grounds of the US Army in Maryland. It incorporated von Neuman's idea of "stored programming" (i.e., the ability to perform arithmetic operations over different operands (loops) without having to repeat the code). The results of the first forecasts were quite encouraging: Fig. 1.2.2, reproduced from Charney (1951) shows the 24-h forecast and verification for 30 January 1949. Unlike Richardson's results, the forecast remains meteorological, and there is a pattern correlation between the predicted and the observed pressure field 24-h changes.

It is remarkable that in his 1951 paper, just after the triumph of performing the first successful forecasts with filtered models, Charney already saw that much more progress would come from the use of the primitive (unfiltered) equations of motion as Richardson had originally attempted:

> The discussion so far has dealt exclusively with the quasi-geostrophic equations as the basis for numerical forecasting. Yet there has been no intention to exclude the possibility that the primitive Eulerian equations can also be used for this purpose. *The outlook for numerical forecasting would be indeed dismal if the quasi-geostrophic approximation represented the upper limit of attainable accuracy, for it is known that it applies only indifferently, if at all, to many of the small-scale but meteorologically significant motions.* We have merely indicated two obstacles that stand in the way of the applications of the primitive equations:

He decided he could never compete with a chimpanzee for the public's favor, and so he gracefully declined to appear, much to the dismay of the computer company that had engineered the invitation in the first place" (Lorenz, 1990).

4 The development of the "Rossby waves" phase speed equation $c = U - \beta L^2/\pi^2$ based on the linearized, non-divergent vorticity equation (Rossby *et al*., 1939, Rossby, 1940), and its success in predicting the motion of the large-scale atmospheric waves, was an essential stimulus to Charney's development of the filtered equations (Phillips, 1990b, 1998).

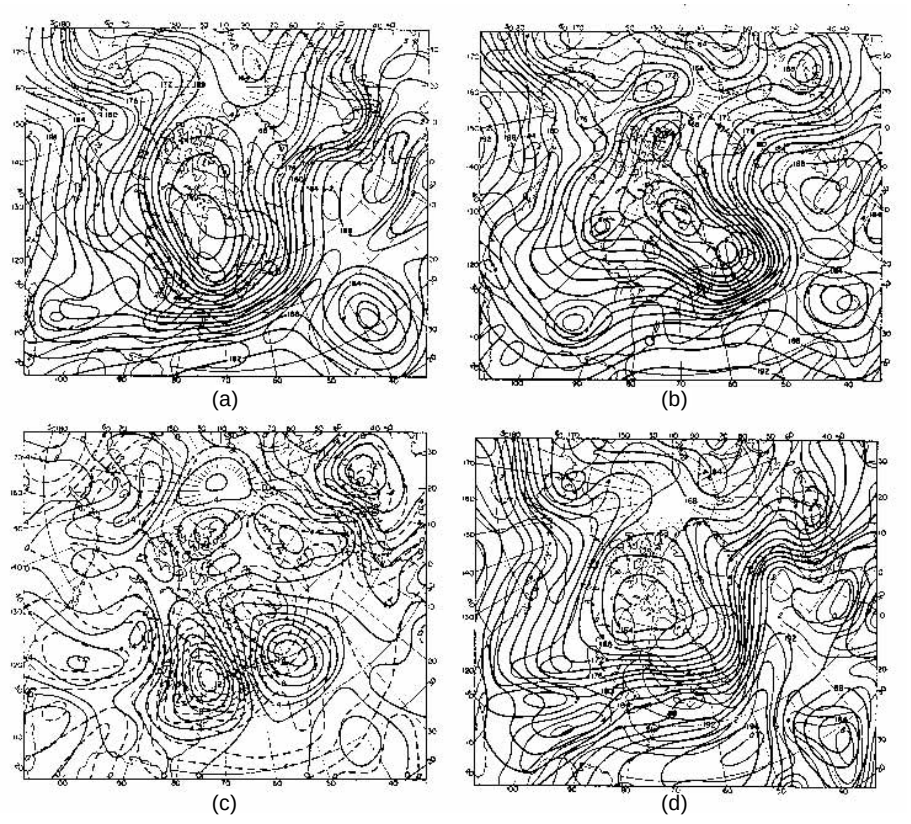


Figure 1.2.2: Forecast of 30 January 1949, 0300 GMT: (a) contours of observed z and $\zeta + f$ at $t = 0$; (b) observed z and $\zeta + f$ at $t = 24$ h; (c) observed (continuous lines) and computed (broken lines) 24-h height change; (d) computed z and $\zeta + f$ at $t = 24$ h. The height unit is 100 ft and the unit of vorticity is $1/3 \times 10^{-4}$ s^{-1}. (Reproduced from the *Compendium of Meteorology*, with permission of the American Meteorological Society.)

> First, there is the difficulty raised by Richardson that *the horizontal divergence cannot be measured with sufficient accuracy. Moreover, the horizontal divergence is only one of a class of meteorological unobservables which also includes the horizontal acceleration. And second, if the primitive Eulerian equations are employed, a stringent and seemingly artificial bound is imposed on the size of the time interval for the finite difference equations. The first obstacle is the most formidable, for the second only means that the integration must proceed in steps of the order of fifteen minutes rather than two hours*. Yet the first does not seem insurmountable, as the following considerations will indicate.

He proceeded to describe an unpublished study in which he and J.C. Freeman integrated barotropic primitive equations (i.e., shallow water equations, Chapter 2) which include not only the slowly varying quasi-geostrophic solution, but also fast gravity waves. They initialized the forecast assuming zero initial divergence, and compared the result with a barotropic forecast (with gravity waves filtered out). The results were similar to those shown schematically in Fig. 1.2.1: they observed

that over a day or so the gravity waves subsided (through a process that we call geostrophic adjustment) and did not otherwise affect the forecast of the slow waves. From this result Charney concluded that numerical forecasting could indeed use the full primitive equations (as eventually happened in operational practice). He listed in the paper the complete primitive equations in pressure coordinates, essentially as they are used in current operational weather prediction, but without heating (nonadiabatic) and frictional terms, which he expected to have minor effects in one- or two-day forecasts. Charney concluded this remarkable paper with the following discussion, which includes a list of the physical processes that take place at scales too small to be resolved, and are incorporated in present models through "parameterizations of the subgrid-scale physics" (condensation, radiation, and turbulent fluxes of heat, momentum and moisture, Chapter 4):

> *Nonadiabatic and frictional terms have been ignored in the body of the discussion because it was thought that one should first seek to determine how much of the motion could be explained without them. Ultimately they will have to be taken into account, particularly if the forecast period is to be extended to three or more days.*
>
> Condensational phenomena appear to be the simplest to introduce: one has only to add the equation of continuity for water vapor and to replace the dry by the moist adiabatic equation. Long-wave radiational effects can also be provided for, since our knowledge of the absorptive properties of water vapor and carbon dioxide has progressed to a point where quantitative estimates of radiational cooling can be made, although the presence of clouds will complicate the problem considerably.
>
> The most difficult phenomena to include have to do with the turbulent transfer of momentum and heat. A great deal of research remains to be done before enough is known about these effects to permit the assignment of even rough values to the eddy coefficients of viscosity and heat conduction. Owing to their statistically indeterminate nature, the turbulent properties of the atmosphere *place an upper limit to the accuracy obtainable by dynamical methods of forecasting*, beyond which we shall have to rely upon statistical methods. But it seems certain that much progress can be made before these limits can be reached.

This paper, which although written in 1951 has not become dated, predicted with almost supernatural vision the path that numerical weather forecasting was to follow over the next five decades. It described the need for objective analysis of meteorological data in order to replace the laborious hand analyses. We now refer to this process as data assimilation (Chapter 5), which uses both observations and short forecasts to estimate initial conditions. Note that at a time at which only one-day forecasts had ever been attempted, Charney already had the intuition that there was an *upper limit* to weather predictability, which Lorenz (1965) later estimated to be about two weeks. However, Charney attributed the expected limit to model deficiencies (such as the parameterization of turbulent processes), rather than to the chaotic nature of the atmosphere, which imposes a limit of predictability even if the model is perfect

(Lorenz, 1963b; Chapter 6). Charney was right in assuming that in practice model deficiencies, as well as errors in the initial conditions, would limit predictability. At the present time, however, the state of the art in numerical forecasting has advanced enough that, when the atmosphere is highly predictable, the theoretically estimated limit for weather forecasting (about two weeks) is occasionally reached and even exceeded through techniques such as ensemble forecasting (Chapter 6).

Following the success of Charney *et al*. (1950), Rossby moved back to Sweden, and was able to direct a group that reproduced similar experiments on a powerful Swedish computer known as BESK. As a result, the first operational (real time) numerical weather forecasts started in Sweden in September 1954, six months before the start-up of the US operational forecasts[5] (Döös and Eaton, 1957, Wiin-Nielsen, 1991, Bolin, 1999).

1.3 Primitive equations, global and regional models, and nonhydrostatic models

As envisioned by Charney (1951, 1962) the filtered (quasi-geostrophic) equations, although very useful for understanding of the large-scale extratropical dynamics of the atmosphere, were not accurate enough to allow continued progress in NWP, and were eventually replaced by primitive equation models (Chapter 2). The primitive equations are conservation laws applied to individual parcels of air: conservation of the three-dimensional momentum (equations of motion), conservation of energy (first law of thermodynamics), conservation of dry air mass (continuity equation), and equations for the conservation of moisture in all its phases, as well as the equation of state for perfect gases. They include in their solution fast gravity and sound waves, and therefore in their space and time discretization they require the use of smaller time steps, or alternative techniques that slow them down (Chapter 3). For models with a horizontal grid size larger than 10 km, it is customary to replace the vertical component of the equation of motion with its hydrostatic approximation, in which the vertical acceleration is considered negligible compared with the gravitational acceleration (buoyancy). With this approximation, it is convenient to use atmospheric pressure, instead of height, as a vertical coordinate.

The continuous equations of motions are solved by discretization in space and in time using, for example, finite differences (Chapter 3). It has been found that the accuracy of a model is very strongly influenced by the spatial resolution: in general, the higher the resolution, the more accurate the model. Increasing resolution, however, is extremely costly. For example, doubling the resolution in the three space dimensions also requires halving the time step in order to satisfy conditions for computational

5 Anders Persson (1999 personal communication) kindly provided the notes on the historical development of NWP in the USA and Sweden reproduced in Appendix A.

stability. Therefore, the computational cost of doubling the resolution is a factor of 2^4 (three space and one time dimensions). Modern methods of discretization attempt to make the increase in accuracy less onerous by the use of semi-implicit and semi-Lagrangian time schemes. These schemes (pioneered by Canadian scientists under the leadership of Andre Robert) have less stringent stability conditions on the time step, and more accurate space discretization. Nevertheless, there is a constant need for higher resolution in order to improve forecasts, and as a result running atmospheric models has always been a major application of the fastest supercomputers available.

When the "conservation" equations are discretized over a given grid size (typically from a few to several hundred kilometers) it is necessary to add "sources and sinks" terms due to small-scale physical processes that occur at scales that cannot be explicitly resolved by the models. As an example, the equation for water vapor conservation on pressure coordinates is typically written as

$$\frac{\partial \overline{q}}{\partial t} + \overline{u}\frac{\partial \overline{q}}{\partial x} + \overline{v}\frac{\partial \overline{q}}{\partial y} + \overline{\omega}\frac{\partial \overline{q}}{\partial p} = \overline{E} - \overline{C} + \frac{\partial \overline{\omega' q'}}{\partial p} \tag{1.3.1}$$

where q is the ratio between water vapor and dry air mass, x and y are horizontal coordinates with appropriate map projections, p is pressure, t is time, u and v are the horizontal air velocity (wind) components, $\omega = dp/dt$ is the vertical velocity in pressure coordinates, and the product of primed variables represents turbulent transports of moisture on scales unresolved by the grid used in the discretization, with the overbar indicating a spatial average over the grid of the model. It is customary to call the left-hand side of the equation, the "dynamics" of the model, which is computed explicitly (Chapter 3).

The right-hand side represents the so-called "physics" of the model. For the moisture equation, it includes the effects of physical processes such as evaporation and condensation $\overline{E} - \overline{C}$, and turbulent transfers of moisture which take place at small scales that cannot be explicitly resolved by the "dynamics". These *subgrid-scale physical processes*, which are sources and sinks for the equations, are then "parameterized" in terms of the variables explicitly represented in the atmospheric dynamics (Chapter 4).

Two types of models are in use for NWP: global and regional models (Chapter 5). Global models are generally used for guidance in medium-range forecasts (more than 2 d), and for climate simulations. At NCEP, for example, the global models are run through 16 d every day. Because the horizontal domain of global models is the whole earth, they usually cannot be run at high resolution. For more detailed forecasts it is necessary to increase the resolution, and this can only be done over limited regions of interest.

Regional models are used for shorter-range forecasts (typically 1–3 d), and are run with a resolution two or more times higher than global models. For example, the NCEP global model in 1997 was run with 28 vertical levels, and 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 48 km and 38 levels, and later in the day with 29 km and 50 levels. Because of their higher resolution, regional models have the advantage of higher accuracy and 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 have the disadvantage that, unlike global models, they 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 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 multiyears), and 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 (United Kingdom (http://www.met-office.gov.uk/), France (http://www.meteo.fr/), Germany (http://www.dwd.de/)), Japan (http://www.kishou.go.jp/), Australia (http://www.bom.gov.au/nmoc/ab_nmc_op.shtml), and Canada (http://www.ec.gc.ca/) also have similar global and regional models, whose details can be obtained at their web sites.

More recently the resolution of some regional models has been increased to just a few kilometers in order to resolve better storm-scale phenomena. Storm-resolving models such as the Advanced Regional Prediction System (ARPS) cannot use the hydrostatic approximation which ceases to be accurate for horizontal scales of the order of 10 km or smaller. Several major nonhydrostatic models have been developed and are routinely used for mesoscale forecasting. In the USA the most widely used are the ARPS, the MM5 (Penn State/NCAR Mesoscale Model, Version 5), the RSM (NCEP Regional Spectral Model) and the COAMPS (US Navy's Coupled Ocean/Atmosphere Mesoscale Prediction System). There is a tendency towards the use of nonhydrostatic models that can be used globally as well.

1.4 Data assimilation: determination of the initial conditions for the computer forecasts

As indicated previously, NWP is an initial-value problem: given an estimate of the present state of the atmosphere, the model simulates (forecasts) its evolution. The problem of determination of the initial conditions for a forecast model is very

important and complex, and has become a science in itself (Daley, 1991). In this section we introduce methods that have been used for this purpose (successive corrections method or SCM, optimal interpolation or OI, variational methods in three and four dimensions, 3D-Var and 4D-Var, and Kalman filtering or KF). We discuss this subject in more detail in Chapter 5, and refer the reader to Daley (1991) as a much more comprehensive text on atmospheric data analysis.

In the early experiments, Richardson (1922) and Charney *et al*. (1950) performed hand interpolations of the available observations to grid points, and these fields of initial conditions were manually digitized, which was a very time consuming procedure. The need for an automatic "objective analysis" quickly became apparent (Charney, 1951), and interpolation methods fitting data to grids were developed (e.g., Panofsky, 1949, Gilchrist and Cressman, 1954, Barnes, 1964, 1978). However, there is an even more important problem than spatial interpolation of observations to gridded fields: the data available are not enough to initialize current models. Modern primitive equations models have a number of degrees of freedom of the order of 10^7. For example, a latitude–longitude model with a typical resolution of 1° and 20 vertical levels would have $360 \times 180 \times 20 = 1.3 \times 10^6$ grid points. At each grid point we have to carry the values of at least four prognostic variables (two horizontal wind components, temperature, moisture), and the surface pressure for each column, giving over 5 million variables that need to be given an initial value. For any given time window of ± 3 hours, there are typically 10–100 thousand observations of the atmosphere, two orders of magnitude less than the number of degrees of freedom of the model. Moreover, their distribution in space and time is very nonuniform (Fig. 1.4.1), with regions like North America and Eurasia which are relatively data-rich, while others much more poorly observed.

For this reason, it became obvious rather early that it was necessary to use additional information (denoted *background*, *first guess* or *prior information*) to prepare initial conditions for the forecasts (Bergthorsson and Döös, 1955). Initially climatology was used as a first guess (e.g., Gandin, 1963), but as the forecasts became better, a short-range forecast was chosen as the first guess in the operational data assimilation systems or "analysis cycles". The intermittent data assimilation cycle shown schematically in Fig. 1.4.2 is continued in present-day operational systems, which typically use a 6-h cycle performed four times a day.

In the 6-h data assimilation cycle for a global model, the background field is a model 6-h forecast x^b (a three-dimensional array). To obtain the background or first guess "observations", the model forecast is interpolated to the observation location, and if they are different, converted from model variables to observed variables y^o (such as satellite radiances or radar reflectivities). The first guess of the observations is therefore $H(x^b)$, where H is the observation operator that performs the necessary interpolation and transformation from model variables to observation space. The difference between the observations and the model first guess $y^o - H(x^b)$ is denoted "observational increments" or "innovations". The analysis x^a is obtained by

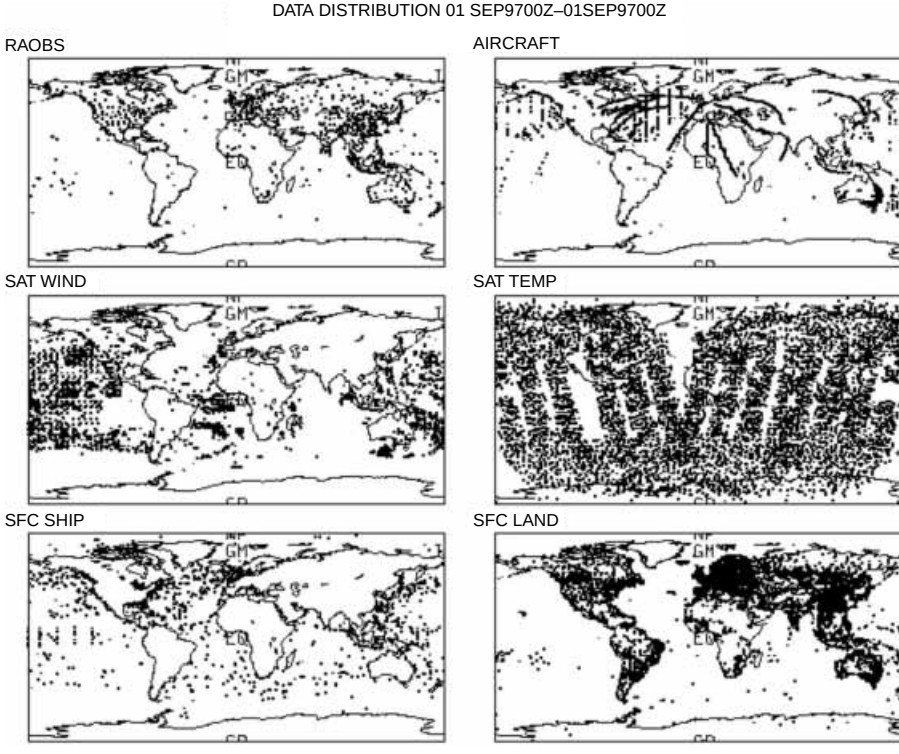


Figure 1.4.1: Typical distribution of observations in a ±3-h window.

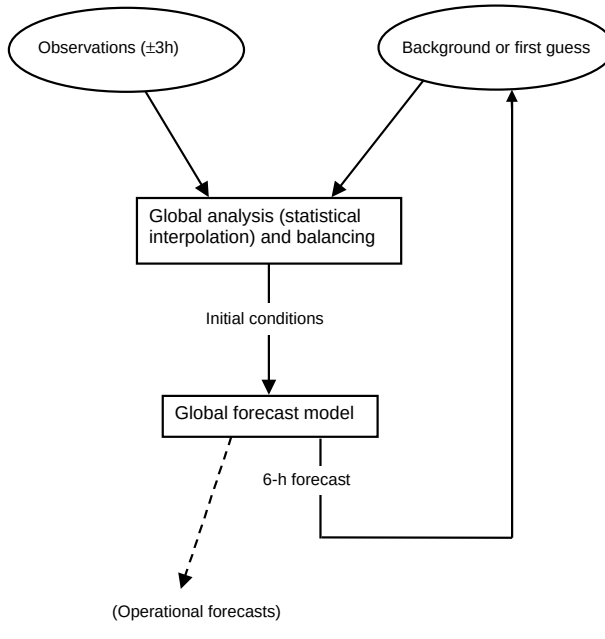


Figure 1.4.2: Flow diagram of a typical intermittent (6-h) data assimilation cycle.

adding the innovations to the model forecast (first guess) with weights W that are determined based on the estimated statistical error covariances of the forecast and the observations:

$$\mathbf{x}^a = \mathbf{x}^b + \mathbf{W}[\mathbf{y}^o - \mathbf{H}(\mathbf{x}^b)] \tag{1.4.1}$$

Different analysis schemes (SCM, OI, 3D-Var, and KF) are based on (1.4.1) but differ by the approach taken to combine the background and the observations to produce the analysis. Earlier methods such as the SCM (Bergthorsson and Döös, 1955, Cressman, 1959, Barnes, 1964) were of a form similar to (1.4.1), 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 OI (Gandin, 1963) the matrix of weights W is determined from the minimization of the analysis errors at each grid point. 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 (Sasaki, 1970). The cost function is minimized directly to obtain the analysis. Lorenc (1986) showed that OI and the 3D-Var approach are equivalent if the cost function is defined as:

$$J = \frac{1}{2}\{[\mathbf{y}^o - H(\mathbf{x})]^T R^{-1}[\mathbf{y}^o - H(\mathbf{x})] + (\mathbf{x} - \mathbf{x}^b)^T B^{-1}(\mathbf{x} - \mathbf{x}^b)\} \tag{1.4.2}$$

The cost function J in (1.4.2) 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 x^b (the second term in the cost function). The distances are scaled by the observation error covariance R and by the background error covariance B respectively. The minimum of the cost function is obtained for $x = x^a$, which is defined as the "analysis". The analysis obtained in (1.4.1) and (1.4.2) is the same if the weight matrix in (1.4.1) is given by

$$\mathbf{W} = \mathbf{B}\mathbf{H}^T(\mathbf{H}\mathbf{B}\mathbf{H}^T + \mathbf{R}^{-1})^{-1} \tag{1.4.3}$$

The difference between OI (1.4.1) and the 3D-Var approach (1.3) is in the method of solution: in OI, the weights W are obtained for each grid point or grid volume, using suitable simplifications. In 3D-Var, the minimization of (1.4.2) is performed directly, allowing for additional flexibility and a simultaneous global use of the data (Chapter 5).

More recently, the variational approach has been extended to four dimensions, by including within the cost function the distance to observations over a time interval (assimilation window). A first version of this considerably more expensive method was implemented at ECMWF at the end of 1997 (Bouttier and Rabier, 1997). Research on the even more advanced and computationally expensive KF (e.g., Ghil *et al.*, 1981), and ensemble KF (Evensen, 1994, Houtekamer and Mitchell, 1998) is discussed in Chapter 5. That chapter also includes a discussion about the problem of enforcing a balance in the analysis so that the presence of gravity waves does not

mask the meteorological signal, as happened to Richardson (1922) (Fig. 1.2.1). The method used for many years to solve this "initialization" problem was "nonlinear normal mode initialization" (Machenhauer, 1977, Baer and Tribbia, 1977). The balance in the initial conditions is usually obtained by either adding a constraint to the cost function (1.4.2) (Parrish and Derber, 1992), or through the use of a digital filter (Lynch and Huang, 1992, Chapter 5).

In the analysis cycle, no matter which analysis scheme is employed, the use of the model forecast is essential in achieving "four-dimensional data assimilation" (4DDA). This means that the data assimilation cycle is like a long model integration, in which the model is "nudged" by the observational increments in such a way that it remains close to the real atmosphere. The importance of the model cannot be overemphasized: it transports information from data-rich to data-poor regions, and it provides a complete estimation of the four-dimensional state of the atmosphere. Figure 1.4.3 presents the rms difference between the 6-h forecast (used as a first guess) and the rawinsonde observations from 1978 to the present (in other words, the rms of the observational increments for 500-hPa heights). It should be noted that the rms differences are not necessarily forecast errors, since the observations also contain errors. In the Northern Hemisphere the rms differences have been halved from about 30 m in the late 1970s, to about 13 m in 2000, equivalent to a mean temperature error of about 0.65 K, similar to rawinsonde observational errors. In the Southern Hemisphere the improvements are even larger, with the differences decreasing from about 47 m to about 12 m. The improvements in these short-range forecasts are a

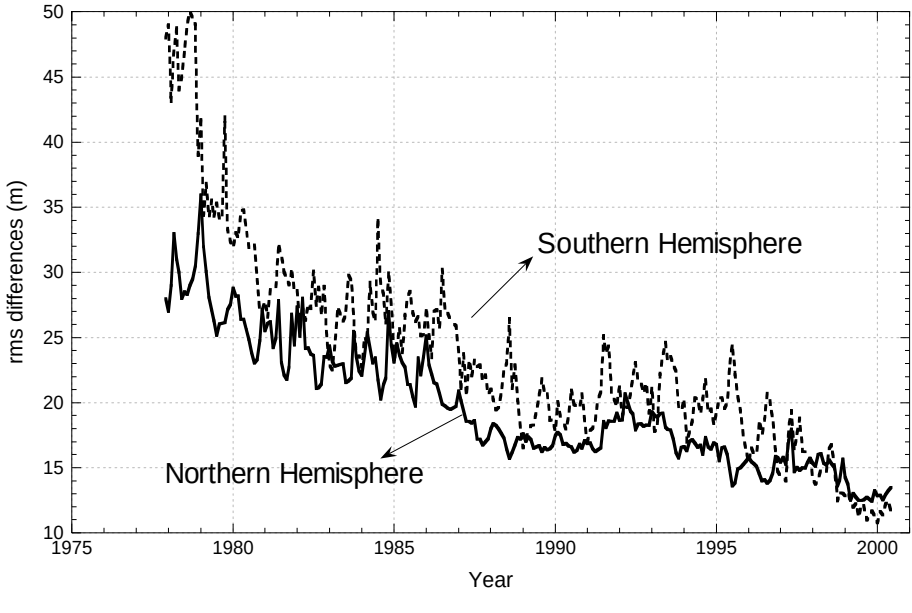


Figure 1.4.3: Rms observational increments (differences between 6-h forecast and rawinsonde observations) for 500-hPa heights (data courtesy of Steve Lilly, NCEP).

reflection of improvements in the model, the analysis scheme used to assimilate the data, and the quality and quality control of the data (Chapter 5).

1.5 Operational NWP and the evolution of forecast skill

Major milestones of operational numerical weather forecasting include the paper by Charney *et al*. (1950) with the first successful forecast based on the primitive equations, and the first operational forecasts performed in Sweden in September 1954, followed 6 months later by the first operational (real time) forecasts in the USA. We describe in what follows the evolution of NWP at NCEP, but as mentioned before, similar developments took place at several major operational NWP centers: in the UK, France, Germany, Japan, Australia and Canada.

The history of operational NWP at the NMC (now NCEP) has been reviewed by Shuman (1989) and Kalnay *et al*. (1998). It started with the organization of the Joint Numerical Weather Prediction Unit (JNWPU) on 1 July 1954, staffed by members of the US Weather Bureau (later the National Weather Service, NWS), the Air Weather Service of the US Air Force, and the Naval Weather Service.[6] Shuman pointed out that in the first few years, numerical predictions could *not* compete with those produced manually. They had several serious flaws, among them overprediction of cyclone development. Far too many cyclones were predicted to deepen into storms. With time, and with the joint work of modelers and practising synopticians, major sources of model errors were identified, and operational NWP became the central guidance for operational weather forecasts.

Shuman (1989) included a chart with the evolution of the $S1$ score (Teweles and Wobus, 1954), the first measure of error in a forecast weather chart which, according to Shuman (1989), was designed, tested, and modified to correlate well with expert forecasters' opinions on the quality of a forecast. The $S1$ score measures the average relative error in the pressure gradient (compared to a verifying analysis chart). Experiments comparing two independent subjective analyses of the same data-rich North American region made by two experienced analysts suggested that a "perfect" forecast would have an $S1$ score of about 20%. It was also found empirically that forecasts with an $S1$ score of 70% or more were useless as synoptic guidance.

Shuman pointed out some of the major system improvements that enabled NWP forecasts to overtake and surpass subjective forecasts. The first major improvement took place in 1958 with the implementation of a barotropic (one-level) model, which was actually a reduction from the three-level model first tried, but which included better finite differences and initial conditions derived from an objective analysis scheme (Bergthorsson and Döös, 1955, Cressman, 1959). It also extended the domain of the

6 In 1960 the JNWPU reverted to three separate organizations: the National Meteorological Center (National Weather Service), the Global Weather Central (US Air Force) and the Fleet Numerical Oceanography Center (US Navy).

model to an octagonal grid covering the Northern Hemisphere down to 9–15° N. These changes resulted in numerical forecasts that for the first time were competitive with subjective forecasts, but in order to implement them JNWPU had to wait for the acquisition of a more powerful supercomputer, an IBM 704, to replace the previous IBM 701. This pattern of forecast improvements which depend on a combination of the better use of the data and better models, and would require more powerful supercomputers in order to be executed in a timely manner has been repeated throughout the history of operational NWP. Table 1.5.1 (adapted from Shuman (1989)) summarizes the major improvements in the first 30 years of operational numerical forecasts at the NWS. The first primitive equations model (Shuman and Hovermale, 1968) was implemented in 1966. The first regional system (Limited Fine Mesh or LFM model, Howcroft, 1971) was implemented in 1971. It was remarkable because it remained in use for over 20 years, and it was the basis for Model Output Statistics (MOS). Its development was frozen in 1986. A more advanced model and data assimilation system, the Regional Analysis and Forecasting System (RAFS) was implemented as the main guidance for North America in 1982. The RAFS was based on the multiple Nested Grid Model (NGM, Phillips, 1979) and on a regional OI scheme (DiMego, 1988). The global spectral model (Sela, 1980) was implemented in 1980.

Table 1.5.2 (from Kalnay *et al.*, 1998 and P. Caplan, personal communication, 2000) summarizes the major improvements implemented in the global system starting

Table 1.5.1. *Major operational implementations and computer acquisitions at NMC between 1955 and 1985 (adapted from Shuman, 1989)*

Year	Operational model	Computer
1955	Princeton three-level quasi-geostrophic model (Charney, 1954). Not used by the forecasters	IBM 701
1958	Barotropic model with improved numerics, objective analysis initial conditions, and octagonal domain.	IBM 704
1962	Three-level quasi-geostrophic model with improved numerics	IBM 7090 (1960) IBM 7094 (1963)
1966	Six-layer primitive equations model (Shuman and Hovermale, 1968)	CDC 6600
1971	LFM model (Howcroft, 1971) (first regional model at NMC)	
1974	Hough functions analysis (Flattery, 1971)	IBM 360/195
1978	Seven-layer primitive equation model (hemispheric)	
1978	OI (Bergman,1979)	Cyber 205
Aug 1980	Global spectral model, R30/12 layers (Sela, 1980)	
March 1985	Regional Analysis and Forecast System based on the NGM (Phillips, 1979) and OI (DiMego, 1988)	

Table 1.5.2. *Major changes in the NMC/NCEP global model and data assimilation system since 1985 (adapted from Kalnay et al. 1998 and P. Caplan, pers. comm., 2000)*

Year	Operational model	Computer
April 1985	GFDL physics implemented on the global spectral model with silhouette orography, R40/18 layers	
Dec 1986	New OI code with new statistics	
1987		2nd Cyber 205
Aug 1987	Increased resolution to T80/18 layers, Penman–Montieth evapotranspiration and other improved physics (Caplan and White, 1989, Pan, 1990)	
Dec 1988	Implementation of hydrostatic complex quality control (CQC) (Gandin, 1988)	
1990		Cray YMP/8cpu/ 32 megawords
Mar 1991	Increased resolution to T126 L18 and improved physics, mean orography. (Kanamitsu *et al.*, 1991)	
June 1991	New 3D-Var (Parrish and Derber, 1992, Derber *et al.*, 1991)	
Nov 1991	Addition of increments, horizontal and vertical OI checks to the CQC (Collins and Gandin, 1990)	
7 Dec 1992	First ensemble system: one pair of bred forecasts at 00Z to 10 days, extension of AVN to 10 days (Toth and Kalnay, 1993, Tracton and Kalnay, 1993)	
Aug 1993	Simplified Arakawa–Schubert cumulus convection (Pan and Wu, 1995). Resolution T126/28 layers	
Jan 1994		Cray C90/16cpu/ 128 megawords
March 1994	Second ensemble system: five pairs of bred forecasts at 00Z, two pairs at 12Z, extension of AVN, a total of 17 global forecasts every day to 16 days	
10 Jan 1995	New soil hydrology (Pan and Mahrt, 1987), radiation, clouds, improved data assimilation. Reanalysis model	
25 Oct 1995	Direct assimilation of TOVS cloud-cleared radiances (Derber and Wu, 1998). New planetary boundary layer (PBL) based on nonlocal diffusion (Hong and Pan, 1996). Improved CQC	Cray C90/16cpu/ 256 megawords

Table 1.5.2. (*cont.*)

Year	Operational model	Computer
5 Nov 1997	New observational error statistics. Changes to assimilation of TOVS radiances and addition of other data sources	
13 Jan 1998	Assimilation of noncloud-cleared radiances (Derber *et al.*, pers.comm.). Improved physics.	
June 1998	Resolution increased to T170/40 layers (to 3.5 days). Improved physics. 3D ozone data assimilation and forecast. Nonlinear increments in 3D-Var. Resolution reduced to T62/28levels on Oct. 1998 and upgraded back in Jan. 2000	IBM SV2 256 processors
June 2000	Ensemble resolution increased to T126 for the first 60 h	
July 2000	Tropical cyclones relocated to observed position every 6 h	

in 1985 with the implementation of the first comprehensive package of physical parameterizations from GFDL (Geophysical Fluid Dynamics Laboratory). Other major improvements in the physical parameterizations were made in 1991, 1993, and 1995. The most important changes in the data assimilation were an improved OI formulation in 1986, the first operational 3D-Var in 1991, the replacement of the satellite retrievals of temperature with the direct assimilation of cloud-cleared radiances in 1995, and the use of "raw" (not cloud-cleared) radiances in 1998. The model resolution was increased in 1987, 1991, and 1998. The first operational ensemble system was implemented in 1992 and enlarged in 1994. The resolution of the ensembles was increased in 2000.

Table 1.5.3 contains a summary of the regional systems used for short-range forecasts (up to 48 h). The RAFS (triple nested NGM and OI) were implemented in 1985. The Eta model, designed with advanced finite differences, step-mountain coordinates, and physical parameterizations, was implemented in 1993, with the same 80-km horizontal resolution as the NGM. It was denoted "early" because of a short data cut-off. The resolution was increased to 48 km, and a first "mesoscale" version with 29 km and reduced coverage was implemented in 1995. A cloud prognostic scheme was implemented in 1995, and a new land-surface parameterization in 1996. The OI data assimilation was replaced by a 3D-Var in 1998, and at this time the early and meso-Eta models were unified into a 32-km/45-level version. Many other less significant changes were also introduced into the global and regional operational systems and are not listed here for the sake of brevity. The Rapid Update Cycle (RUC), which provides frequent updates of the analysis and very-short-range forecasts over