Real-time limited area numerical weather prediction in Australia: a historical perspective

L.M. Leslie and G.S. Dietachmayer
Bureau of Meteorology Research Centre, Australia

(Manuscript received April 1992; revised July 1992)

For more than a quarter of a century, Australian scientists have been active in developing numerical weather prediction models for operational usage. One of the earliest modellers was R.H. (Reg) Clarke, who in 1967 produced the first numerical weather prediction for the southern hemisphere using the primitive equations of motion.

In this paper we trace the development of limited area numerical weather prediction over the Australian region from the early days to the present. Our primary emphasis will be on the techniques used to solve the primitive equations; however, we also examine other important aspects of numerical weather prediction. These include the choice of lateral boundary conditions, the problem of making the best use of the unique mix of conventional and remotely-sensed data available over the Australian region, and the inherent predictability limitations of the atmosphere.

Finally, we discuss the directions limited area numerical weather prediction might take in Australia during the 1990s.

Introduction

In a colloquium held in Melbourne in 1967, Mr R.H. (Reg) Clarke of the CSIRO Division of Meteorological Physics stated that 'Australian meteorologists should consider the applicability of primitive equation models to current forecast problems' (Clarke 1967). At the time, Reg was attached to the Geophysical Fluid Dynamics Laboratory (GFDL), then in Washington, but which moved soon after to its current location in Princeton. Following his return to Australia in 1968 Reg presented the first multi-level numerical forecasts for the southern hemisphere. These had been carried out with the GFDL primitive equations model which Reg helped to develop, his major contribution being to the boundary-layer formulation as described in another paper in this issue (Hess 1992). Reg's work in this area eventually was published in Monthly Weather Review (Clarke and Strickler 1972). His words clearly were heeded in that the Australian Bureau of Meteorology issued the first southern hemispheric operational model forecasts in 1973 (Gauntlett and Hincksman 1971; Gauntlett et al. 1972).

Limited area (or regional) operational numerical weather prediction (NWP) over Australia, on which this paper concentrates, has an even longer history, with numerical forecasts (using one-level models) being carried out at the University of Melbourne in the early 1960s (Jessen and Radok 1960). The first real-time forecasts issued by the Bureau of Meteorology commenced in 1969 (Maine 1967). This quasi-geostrophic barotropic model was far less sophisticated than the hemispheric primitive equations model that Reg Clarke introduced to the Australian meteorological community. Subsequently, a multi-level filtered model replaced the barotropic model in operations (Maine 1972). However, as far as the authors are aware, the first multi-level primitive equations Australian region forecast was an experimental application of the British Meteorological Office's operational model (White 1971). It was not until 1977 that the Australian Bureau of Meteorology implemented its own operational limited area primitive equations NWP model (McGregor et al. 1978). Since then, two major operational upgrades have taken place: in 1986 significant improvements were made in both model resolution and the representation of physical processes (Leslie et al. 1985) and in 1989 an
intermittent data assimilation procedure was introduced (Mills and Seaman 1990).

At this stage it is necessary to define more precisely what we mean by the term ‘NWP’ in this paper. Our concept entails two primary components: firstly, an emphasis on integrating from initial states based on real data (‘forecasting’) rather than idealised initial states; and, second, a requirement that the forecasts must be produced in real time. The most important research goal of NWP practitioners is to improve the quality of these integrations, that is, to reduce the model forecast error (MFE). MFE usually is defined in terms of some measure of the difference between the analysis, defined as the specification of the state of the atmosphere at some point in time, and the corresponding model forecast. For real-time applications, model efficiency also obviously is of crucial importance, but gains in efficiency generally are funnelled back into model improvements directed at reducing the MFE. Indeed the objective of reducing MFE has been the preoccupation of NWP modellers since the first forecasts in 1950, and considerable success has been achieved since then. For example, Fig. 1 shows the six-monthly running-mean of the $S_1$ skill score (defined below) of the model forecast minus the $S_1$ skill score of persistence, defined over the Australian region, for the entire history of Australian limited area NWP. Significant milestones, such as the introduction of new models, are highlighted. The $S_1$ score (Teweles and Wobus 1954; Anthes 1983) is an accepted objective measure of forecast skill defined as

$$S_1 = 100 \frac{\Sigma e_j}{\Sigma \max (d_o, d_f)},$$

where $e_j$ is the error of the forecast pressure difference, $d_o$ is the observed pressure difference and $d_f$ is the forecast pressure difference between two points. Clearly, a lower $S_1$ score indicates a more skillful forecast.

This improvement in model skill has been hard-won, as there are many challenges in the development of NWP algorithms, some common to computational fluid dynamics (CFD) in general, others peculiar to the discipline itself. In very broad terms these may be split into three categories: numerical, physical and technological. The numerical integration of a set of partial differential equations may appear straightforward, but is in actuality beset with problems. One example is the numerical treatment of the artificial lateral boundaries present in limited area models. Another is the phenomenon of ‘non-linear instability’, first identified in the pioneering work of Phillips (1956, 1959). This instability arises when non-linear interaction within the model attempts to transfer energy down to length scales which the model is incapable of resolving. It responds by spuriously projecting this energy back up-scale to a resolvable frequency (‘aliasing’) and the resultant resolvable-scale energy growth leads eventually to instability. In nature the down-scale transfer of energy ends eventually in viscous dissipation. In numerical models non-linear instability can be controlled by a number of techniques such as the application of low-pass filters or through the finite-difference algorithms themselves (Arakawa 1966).

The sheer number and complexity of physical processes present in the atmosphere pose further difficulties for NWP. Clouds play a vital role in transporting water, heat and momentum throughout the atmosphere but are very difficult to model as they encompass, over a vast range of scales, many physical processes most of which are poorly understood. Similarly, turbulence is a difficult problem even in the relatively ‘clean’ context of simple laboratory experiments, but for atmospheric applications there are many additional complicating factors (for example, effects due to density stratification and heterogeneous boundary-layer forcing). Representation of radiative heating and cooling in a cloudy atmosphere and the interaction of the free flow with the planetary boundary layer also must be accounted for. A description of the representation of physical processes used by the Bureau of Meteorology Research Centre (BMRC) operational model is given by Hart et al. (1988). Moreover, even if these processes could be modelled exactly, there are still inherent limits to model forecast accuracy. In general, future weather states are sensitive to small variations in the initial state. The chaotic nature of non-linear dynamical systems, such as the primitive equations, was first discussed by Lorenz (1963) and the estimate of a predictability limit of about two weeks for global forecasts (Lorenz 1969) still is the generally accepted value.

The implications for limited area models appear mixed, and are possibly dependent on the scale of the phenomenon simulated. In the re-
gional-scale studies of Errico and Baumhefner (1987) and Vukicevic and Errico (1990), it is shown that predictability can be enhanced, in comparison to global forecasts, by the use of one-way lateral boundary conditions (in which 'perfect' large-scale data are continually fed to the model through the inflow boundaries) and by systematic topographic forcing. Of course, enhanced model predictability does not necessarily imply enhanced atmospheric predictability. At very high resolutions, such as those required for the simulation of individual thunderstorms, the predictability problem can become severe, with relatively small changes in the initial conditions (or relatively minor changes in the physical parameterisations) leading to rapid solution divergence (Lilly 1990). On the positive side, the a priori classification of weather systems in terms of their degree of predictability has profound implications for operations, in that it can be used to provide the forecaster with error bounds on the numerical forecast (Leslie and Holland 1991).

Finally, technology has driven NWP development in two main ways; by providing ever-increasing computer power and through the steady introduction of new instruments for data acquisition. The increases in computer power allow the use of higher resolutions, larger domains, and more sophisticated numerical algorithms and physical parameterisations. The associated hardware changes (from scalar to vector to massively parallel architectures) can have a profound impact on the structure of the model code. For example, the 'on/off' nature of many physical parameterisations is difficult to implement efficiently on massively parallel computers, particularly on SIMD (single-instruction multiple-data) machines. Of course, improvements in computer technology are not just applied to models, but are also used in processing the output to provide better diagnostics and facilities such as animated flow visualisation. Also it is clear that forecast accuracy is related to the adequacy of the data base, and in Australia data sparsity has long been a major problem. Advances in observing systems such as satellites, aircraft-borne instruments and automatic weather stations will enhance the data base significantly in the 1990s.

Within a broadly historical framework this paper documents the evolution of limited-area numerical modelling over the Australian region, focussing on the following areas: the finite-difference numerical algorithms used to perform the model integrations; formulation of lateral boundary conditions; methods for assimilating available data; and schemes for the a priori estimation of model forecast error. Particular emphasis is placed on techniques tailored to meet operational requirements, such as model reliability and forecast timeliness. Therefore, major areas such as particular methods of meteorologi-
cal data analysis, global modelling, Galerkin methods (spectral and finite-element schemes), and parameterisation of subgrid-scale processes (usually referred to as the model's 'physics'), will either not be covered or only touched upon. Finally, we look at future directions that limited area NWP may take in the next decade.

**Numerical algorithms**

In common with more general CFD, numerical techniques for integrating the governing equations are judged on the basis of their robustness, accuracy and efficiency. NWP differs from CFD in setting such a high priority on robustness, as it is considered vital that some numerical forecast guidance be available for operational use at all times. In contrast, an occasional model failure is tolerated in a research environment, and may in fact be used as an indication that the problem under consideration is simply too difficult for the current model configuration to deal with successfully (Gresho and Lee 1981).

**The primitive equations**

The satisfaction of boundary conditions at the ground is facilitated by the use of a terrain-following vertical coordinate. The most common choice for hydrostatic models is the 'sigma' representation (Phillips 1957), for which the vertical coordinate is \( \sigma = \rho / \rho_s \). The governing equations can then be written as

\[
\frac{du}{dt} - \nu(f + f_s) = -m \left( \frac{\partial \phi}{\partial x} + \frac{1}{p_s} \frac{RT \partial p_s}{\partial x} \right), \quad \ldots \quad 1
\]

\[
\frac{dv}{dt} + u(f + f_s) = -m \left( \frac{\partial \phi}{\partial y} + \frac{1}{p_s} \frac{RT \partial p_s}{\partial y} \right), \quad \ldots \quad 2
\]

\[
\frac{\partial \phi}{\partial \sigma} = -RT \sigma, \quad \ldots \quad 3
\]

\[
\frac{\partial p_s}{\partial t} = -m^2 \left[ \frac{\partial}{\partial x} \left( \frac{p_s u}{m} \right) + \frac{\partial}{\partial y} \left( \frac{p_s v}{m} \right) \right] - p_s \frac{\partial \omega}{\partial \sigma}, \quad \ldots \quad 4
\]

and

\[
\frac{\partial \theta}{\partial t} = 0, \quad \ldots \quad 5
\]

where

\[
\frac{d}{dt} = \frac{\partial}{\partial t} + m \left( \frac{u}{\partial x} + \frac{v}{\partial y} \right) + \frac{\partial \theta}{\partial \sigma}, \quad \ldots \quad 6
\]

and

\[
f_s = m^2 \left[ \frac{v}{\partial x} \left( \frac{1}{m} \right) - \frac{u}{\partial y} \left( \frac{1}{m} \right) \right], \quad \ldots \quad 7
\]

In Eqns 1 to 7, \( u \) and \( v \) are the horizontal velocity components, \( \omega \) is the vertical velocity component, \( w = \sigma \partial \sigma / \partial t \), \( \phi, T, \theta, p, p_s \) are the geopotential, temperature, potential temperature, pressure and
surface pressure respectively, $f$ is the Coriolis parameter, and $m$ is the map factor for the chosen projection. One of the advantages of the Lambert conformal map projection is that local derivatives of the map factor are very small so that the $f$ term can be neglected, and in operational Australian region models using this projection the term has always been discarded.

**Early models**

Methods for the time-integration of partial differential equations commonly are categorised as either 'explicit' or 'implicit' (Richtmyer and Morton 1967; Roache 1982). Explicit methods suffer from a severe restriction on the allowable size of the time-step (with a concurrent loss in efficiency) when the governing equations support very fast modes, as is the case with the full primitive equations. In contrast with the simplicity of the explicit approach, implicit methods require the solution of a set of simultaneous equations at each time-step, however the restriction on the time-step is relaxed. Major centres around the world with sufficient computer capacity (for example, National Meteorological Center (NMC), Washington, the UK Meteorological Office, Japan Meteorological Agency) chose to use explicit schemes such as the leap-frog, Lax-Wendroff and Matsuno methods respectively (see Mesinger and Arakawa 1976) for details of explicit time-differencing schemes. However, the relatively limited computing resources available in the early 1970s meant that regional NWP in Australia was restricted to filtered models. Following the path set by the classical work of Charney et al. (1950), the earliest forecast models assessed for operational usage were filtered barotropic models of various kinds (Maine 1967). As mentioned above, a quasi-geostrophic barotropic model was the first operational limited area NWP model in Australia (Maine 1967). The domain over which the forecasts were made is shown in Fig. 2 and it has remained as the operational forecast domain for the Bureau of Meteorology for over twenty years. (It has been expanded recently to minimise the impact of error propagation from the lateral boundaries into the forecast region.) Several years later a multi-level filtered model was introduced operationally (Maine 1972).

**The 1970s**

The 1970s saw the emergence of efficient semi-implicit and split-explicit approaches (Robert et al. 1972; Gadd 1978). Both recognise that the fastest modes are associated only with certain terms in the governing equations (the RHS of Eqs 1, 2 and 4), but adopt different strategies for treating these terms. The semi-implicit technique linearises these terms and then time-averages them over two or three time levels, thereby rendering them implicit. This implicit treatment of the linearised gravitational mode terms results in the allowable time-step being restricted only by the slower advective and Rossby modes and typically allows the time-step to be increased by a factor of between four and six. The split-explicit method uses an operator-splitting approach (Marchuk 1971) to separate the fast and slow mode components of the equations, and the fast mode terms can then be integrated separately using a series of small time-steps. Efficiency of the method as a whole is achieved as the fast mode terms are relatively inexpensive to calculate.

In Australia the semi-implicit path was the only one considered for the production of operational 24-hour predictions in the available time. Following several years of model development (see, for example, Gauntlett et al. (1976)) the filtered equations model finally was replaced in operations by a semi-implicit primitive equations model in 1977 (McGregor et al. 1978). The limitations imposed by the computing environment of the time considerably restricted the model configuration to 250 km horizontal resolution and six levels in the vertical. Representation of physical processes in the model also was very simple. It consisted of a constant flux surface layer with a drag coefficient of 0.004 and 0.001 over the land and oceans respectively; mixing-length theory above the surface layer; and precipitation processes comprised of large-scale ascent based on a relative humidity threshold. A version of the Arakawa-Schubert convection scheme was tested initially but not used in the operational configuration of the model.
The 1980s
The 1980s was both a period of maturation, in which the split-explicit and semi-implicit techniques became established in NWP centres around the world, and a period of active research which saw the introduction and subsequent rapid development of the semi-Lagrangian method. As the name suggests, the method results from a change in viewpoint from an Eulerian to Lagrangian perspective, which simplifies the governing equations, but introduces an additional complication in that trajectories must now be calculated. For the simplest case of passive advection of a scalar quantity $Q$, governed by the equation,

$$
\frac{dQ}{dt} = 0,
$$

a Lagrangian calculation for the value of $Q$ at grid point $(x_n, y_n)$ at time $t + \Delta t$ yields

$$
Q(x_n, y_n, t + \Delta t) = Q(x_n, y_n, t),
$$

where $(x^*, y^*, t)$ is the beginning of the trajectory which ends at $(x_n, y_n, t + \Delta t)$. The numerical implementation of Eqn 9 requires two calculations, the evaluation of $x^*$ and $y^*$, and an interpolation from known grid-point values to find the value of $Q$ at that point. There are many ways to estimate the trajectories and perform the interpolations and this has led to a proliferation of publications on the topic, beginning with the fundamental paper by Robert (1981), which extended earlier similar types of schemes to a formulation valid for Courant numbers greater than unity. Further important contributions were also made by Bates, McDonald, Ritchie, Staniforth, Temperton and Côté, and a comprehensive summary of their work and others is given in the recent review article by Staniforth and Côté (1991). The improvements introduced include ‘non-interpolating’ schemes, global applications, general improvements to the accuracy of both the trajectory and interpolation procedures and, most importantly, the extension of the approach to the barotropic and full primitive equations in combination with both split-explicit and semi-implicit schemes.

In Australia, operational NWP concerned itself mainly with improvements to the existing semi-implicit model, in particular increases in resolution to 150 km in the horizontal and 12 levels in the vertical, and the introduction of a much more sophisticated physics package (Leslie et al. 1985). By the end of the eighties the Australian region analysis and forecast systems were integrated into a six-hour cycle intermittent data assimilation system referred to as ‘RASP’ (regional assimilation and prognosis). The RASP system, which has 16 levels in the vertical and is described in detail in Mills and Seaman (1990), is discussed further in a later section.

The 1990s
During the late 1980s and early 1990s a considerable amount of research was carried out within BMRC on model formulation in an attempt to develop a NWP model that would replace the RASP model and fit within the multi-processor supercomputer environment of the 1990s. There was a number of key research findings that resulted from this research effort and they are summarised as follows. Firstly, contrary to the almost universal opinion (Haltiner and Williams 1980) that the non-staggered grid (or A-grid in the notation of Arakawa and Lamb (1977)) was unsuitable for NWP modelling, it was demonstrated that the A-grid was a very attractive one (Purser and Leslie 1988). The argument against the A-grid was that it was inaccurate relative to the other, staggered, grids, due to the problem of two grid-length noise associated with grid separation. Purser and Leslie showed that simply by using high-order differencing operators, which are introduced very easily, finite-difference approximations could be made at least as accurate as those on any of the other grids. In addition, they pointed out that none of the grids handle features at the highest wave numbers and the noise generated in the A-grid could be removed simply by the application of well-designed low-pass filters (Purser 1987).

The arguments presented above for the horizontal non-staggered grid were extended to the vertical, where it was shown further that the use of high-order operators in the vertical significantly improved the accuracy of the forecasts and, once again, demonstrated that the non-staggered grid was a suitable choice (Leslie and Purser 1991). Of particular interest is the use of ‘compact’ schemes for high-order differencing in the vertical (Leslie and Purser 1992). Such schemes both reduce truncation error (by significantly reducing the size of the coefficient of the leading truncation error term), and help to avoid the stencil-width problems which occur near the ground and the top boundary.

Given the research findings described above, and the acquisition of a vector supercomputer by the Australian Bureau of Meteorology, the choice of models for the 1990s was reduced essentially to either a semi-implicit, semi-Lagrangian scheme (with no time-step restriction) or an explicit scheme. Both approaches are very efficient on vector machines. The choice finally settled upon was a two time-level explicit scheme with third or higher-order differencing on a non-staggered grid. This choice was made basically because of the simplicity of the code, the ease of introducing high-order differencing, and because the semi-Lagrangian scheme still has some unresolved issues concerning the calculation of the departure point, at least for very small-scale features with strong deformation on a length-scale comparable to the model grid spacing (Seibert and Moraiu
1991). However, the very recent work by McGregor (1992) goes some way towards addressing this problem by incorporating information about the spatial variability of the velocity field into the trajectory calculation.

The new model has been completed and currently is being optimised on the Cray Y-MP/2E vector supercomputer recently acquired by the Australian Bureau of Meteorology. It is anticipated that a horizontal resolution between 50 and 75 km will be possible, and that the number of levels in the vertical can be increased to at least 20. Model resolutions of this order are comparable with those presently used at other NWP centres around the world.

**Lateral boundary conditions**

Within a given computing environment, limited area models enjoy the advantage of operating at considerably higher resolution than their global counterparts. This advantage has a cost, namely, the introduction of lateral boundaries which have no counterpart in the real atmosphere. Because the boundaries are artificial, simple physically-based boundary conditions are not available and a variety of techniques have been experimented with. A further complication is introduced by the governing equations themselves, as Oliger and Sundström (1978) have argued that the primitive equations are ill-posed for any choice of local boundary conditions. With no rigorous construction available, attention has focussed on more pragmatic approaches.

We assume that large-scale boundary data are available from either observational data or other model forecasts, and that this large-scale flow essentially is independent of the small-scale features that the limited area model is intended to capture. The latter assumption is required if one-way nesting is used in preference to two-way nesting, which is the usual situation in operational systems. The lateral boundary conditions must then be formulated in such a way as to prevent the reflection of outgoing small-scale information, whilst simultaneously allowing large-scale features to propagate into the limited area domain. For simple equation sets, boundary conditions meeting these requirements can be found using the theory of characteristics. This approach is particularly useful for avoiding the problem of overspecified boundary conditions, which otherwise usually result in grid-scale noise propagating inwards from the boundaries. This source of error was demonstrated by Shapiro and O'Brien (1970) in the context of a limited area barotropic model. When the vorticity was specified at all boundary grid-points, two grid-length noise in the vorticity field was generated on outflow boundaries (Fig. 3(a)). This noise was removed by allowing the limited area model to determine its own outflow vorticity values using a Lagrangian method (Fig. 3(b)). Although characteristic theory can be applied to construct exact outgoing radiation conditions valid in the presence of dispersive waves, the boundary conditions so obtained are intractable owing to excessive time and storage requirements (Bennett 1976). Despite this rather negative result, approximate radiation conditions have been used successfully, and steadily improved upon by a number of workers (Orlanski 1976; Miller and Thorpe 1981; Raymond and Kuo 1984). Application of these approximate radiation conditions has been limited to research models, presumably due to two factors. Firstly, the perceived inability of such conditions to incorporate smoothly large-scale data external to the fine-grid model. Indeed, although Carpenter (1982) demonstrated that external boundary forcing may be included within the radiative boundary condition formulation, his ideas have been implemented only recently in the work of Chen (1991). Second, current radiation conditions treat each model variable independently, ignoring the coupling between the fields enforced by the primitive equations.

Faced with the above difficulties of the characteristic-based approaches, some workers have opted to specify the value of all (or most) boundary variables, and then introduce a modification to the governing equations to prevent the more deleterious effects of over-specification of the boundary conditions. Davies (1976) added dynamical relaxation terms, similar to ‘Newtonian law’ adjustments, to the prognostic equations to continually force the predicted variables back towards their external values. The coefficients of these relaxation terms are maximised at the model boundaries and decrease rapidly with distance into the model domain. The spatial variation of the coefficients necessitated the introduction of additional terms to prevent the introduction of a spurious source term of vorticity. Outgoing waves are strongly damped by the relaxation terms, and hence, as far as the interior solution is concerned, these waves appear to exit freely the model domain. In the scheme of Perkey and Kreitzberg (1976) the tendency of model variables near the boundaries is set as a weighted average of the model-predicted tendency and the tendency associated with the large-scale external flow. They used a linear fall-off in the weights so that the tendencies were entirely model-generated four or more grid-points away from the boundaries. This method does not damp outgoing waves, but instead slows them, with the degree of retardation increasing with decreasing distance away from the boundary. This differential slowing
Fig. 3(a) A 12-hour barotropic forecast for 1200 UTC 8 December over North America (continental outline not shown). Solid lines show the streamfunction field, dashed lines the absolute vorticity, and shaded areas denote regions in which the absolute vorticity is greater than $1.6 \times 10^{-4}\text{sec}^{-1}$. Boundary conditions on the streamfunction and vorticity obtained by interpolation from a coarse-mesh hemispheric forecast. Note the noise in the vorticity field, particularly in the northeast corner.

Fig. 3(b) As in 3(a), but with the outflow boundary conditions on vorticity specified by a pseudo-Lagrangian scheme. Note that the noise in the northeast corner has been eliminated. Figures 3(a) and (b) adapted from Shapiro and O'Brien (1970).

leads to a compaction of the waves, which can then be efficiently removed by filtering. Again, the effect is the same as that of waves being transmitted through the boundary.

The scheme that has been implemented in Australian NWP models primitive equation models up to and including RASP is that of Kallberg (1977), which can be interpreted as a simplified version of the Davies scheme, applied in an operator-split manner. It is likely that schemes of the Davies or Perkey/Kreitzberg type will continue in operational use for some time as they are robust and simple to use. However, their use of buffer zones near model boundaries is wasteful and thus in the future they could face some competition from the characteristic-based schemes should the latter be demonstrated to be effective in handling large-scale external forcing.

The view that high resolution can be achieved only within the context of a limited area model implicitly assumes that model resolution essentially is constant over the entire model domain. Courtier and Geleyn (1988) and Staniforth et al. (1991) have relaxed this assumption by introducing global models with high resolution over a limited area through the use of coordinate transformations which allow the resolution to vary significantly around the globe. Although there are potential difficulties with this approach, its great advantage is that lateral boundaries are not present, and hence the difficulties discussed above cease to apply.

Data assimilation

Data analysis
The specification of the initial state of the atmosphere, also known as objective analysis, is usually achieved through one of four means: surface-fitting techniques, empirical interpolation methods, statistical interpolation schemes, and variational analysis approaches. An excellent summary of the different methods is given by Schlatter (1988) along with a description of their individual weaknesses and advantages. At present most centres have adopted the statistical interpolation analysis methods, usually the multivariate optimal interpolation technique. For Australian operational use, cost effectiveness has always been a major issue owing to the limited computer resources available, and the more computationally expensive statistical interpolation techniques only recently have become standard procedure. Currently, multivariate optimal interpolation analysis is replacing the previous univariate scheme in operations.
Data assimilation schemes

In the past, NWP systems were dominated by approaches in which the analysis was prepared at time $t = 0$ and a forecast of specified duration then was made with the NWP model. In this way the analysis and the forecast were regarded as independent parts of the total analysis/forecast system. As described below, a number of problems arose from this demarcation. Methods aimed at integrating the analysis and model forecast into a single process with the general aim of improving the forecast are referred to as ‘data assimilation’ techniques. In the context of operational NWP there are three primary problems posed by the interaction of data and model. The first is that there simply have not been enough data available to completely specify the state of the atmosphere at a given time. Figure 4 shows the data available at two consecutive data assimilation times. Note the lack of data at 1800 UTC and also the large variation in data coverage that can occur between successive assimilation times. The second is that although the model produces a complete solution we are interested mainly in the slow manifold (for the atmosphere, the Rossby modes) component, and hence we wish to use the available data in such a way as to avoid exciting the fast inertia-gravity modes. The third, commonly referred to as ‘spin-up’, is that the models need to generate a self-consistent distribution of those model variables (such as vertical velocity, or precipitation rate) which are not readily provided from conventional analyses. This problem is further exacerbated by inadequacies in the model physics. The consequences of the spin-up problem are most keenly felt in the hydrological variables, leading to model forecasts in which initially the precipitation is too low and the evaporation is too high.

Early NWP systems used statistical methods to construct a complete initial condition for the model, and then applied a process of ‘initialisation’ (see, for example, Haltiner and Williams (1980), Chapter 11) which modified this data by filtering out the fast modes. A model forecast was then performed using the modified data set. In this approach the data and the model were viewed as separate, independent entities. In its broadest sense, the term ‘data assimilation’ refers to the coupling of both model and data, with the goal of improving the quality of both. Improvements in the quality of data sets are important for research applications. Here we are concerned with operational NWP and hence we focus on improvements to model forecast accuracy. At present most NWP centres use the same method for achieving this linking of model and data, namely, ‘intermittent’ data assimilation. This is a cycling analysis/forecast procedure in which the forecast model provides a background field, which is interpolated to analysis locations to act as the required ‘guess’ field. Differences between the guess field and the observations within a time window centered on the observation time are calculated, interpolated in space, and then added back to the first guess to provide the new analysis. The procedure is repeated, usually in six or even three-hourly cycles. The actual cycle time depends upon both the nature and availability of the data. Intermittent data assimilation, which is inherently three-dimensional in design, cannot directly treat the initialisation problem and hence a separate initialisation procedure is required before the model forecast commences. The spin-up problem remains, but can be reduced by using a scheme in which only analysis increments are added to the first-guess forecast. The current Australian regional model, RASP, uses such an intermittent six-hourly data assimilation cycle with the initialisation problem treated using the vertical mode initialisation (VMI) scheme of Bourke and McGregor (1983). An example of the reduction in the spin-up problem obtained using RASP is given in Fig. 5.

The current trend in assimilation work is towards ‘four-dimensional’ or ‘continuous’ assimilation in which data are inserted at their true (usually asymptotic) time instead of being grouped
at pre-specified synoptic times. The time-continuous nature of this approach further diminishes the spin-up problem. There are two main types of four-dimensional data assimilation: the 'nudging' technique (see for example Anthes and Hoke (1974)) in which the model variables are continuously forced (through a simple modification of the governing equations) to target values; and the 'variational' approach in which the formal objective is to find the state of the atmosphere which provides a best fit to the data, boundary conditions and, significantly, to the equations of motion themselves. It is this inclusion of the equations of motion that is the distinctive feature of the present-day variational data assimilation methods. The criterion for best fit is a weighted quadratic functional that includes the data, dynamical boundary condition, and initial condition misfits. There is now a large amount of literature in the field, but for a clear and elegant exposition of the theory of variational data assimilation the reader is referred to the relatively early paper by Talagrand and Courtier (1987).

In Australia the nudging method is now being used in a real-time NWP system operating over a domain centered on the tropics, the spin-up problem has been greatly diminished by the use of proxy moisture data derived from satellite imagery. This work is described in Puri et al. (1992), and Davidson and Puri (1992). An example of how well the tropical cyclone track is forecast using this system is shown in Fig. 6 (Puri and Davidson, personal communication). Some work also is being carried out within BMRC using the variational method and is described below in the section on future directions.

### Atmospheric predictability

As was mentioned in the Introduction, NWP has regarded the reduction of MFE as its major goal, and has been very successful in doing so. However, there have been a number of disturbing features that have remained despite the achievements. One problem is the large variation in model forecast skill on a day-to-day basis, as illustrated for RASP in Fig. 7. Such variations erode the level of confidence in the skill of a forecast on a given day. Another problem is that such variations in skill occur not only in a domain averaged sense, but also on a geographical basis. A particular forecast may be very good in some parts of the domain, but poor over other parts of the domain. Finally, in recent years, NWP centres around the world have had available forecasts from more than one centre. On many occasions the forecasts are very similar, particularly the first day or two. However, major differences frequently occur between the forecasts, and on these occasions the duty forecasters are faced with the difficult task of choosing between these forecasts. An example is given in Fig. 8 which shows forecasts over the Australian region from two different models. The initial analyses are very similar, but the forecasts have diverged dramatically by day five. These variations in MFE of individual NWP models, and between NWP models, are a manifestation of the fact that there are inherent limits to the predictability of the deterministic system that comprises the primitive equations set. Moreover, the predictability of the atmosphere varies from day to day and place to place.

Meteorologists have been aware for a long time that the equations governing atmospheric motion are chaotic in the sense that they are very sensitive to changes (errors) in the initial state. Lorenz (1963) showed nearly thirty years ago that a simplified version of the equations of motion exhibited chaotic behaviour and made the first estimate of the limit of deterministic predictability for the atmosphere. Lorenz's estimate was about two weeks and this still is the generally accepted limit, beyond which forecasts using the primitive equations are of little value, unless statistically averaged in some sense.

However, until the 1980s NWP modellers appeared not to be particularly concerned with the existence of the predictability limit. There were two main reasons for this apparent lack of interest in predictability considerations: firstly, until the 1980s forecasts in most centres were restricted to 48 hours or less, particularly in the case of regional modelling, owing largely to limitations in data coverage, computer power and sophistication of the NWP systems. Second, the necessary mathematical tools did not exist for analysing the behaviour of non-linear sets of equations like the primitive equations. It was only with the develop-
Fig. 6  Tropical cyclone forecasts out to 48 hours for tropical cyclone *Fran* using the tropical NWP model of Puri et al. (1992), together with the corresponding verifying analyses.
Fig. 7 Illustration of the large day-to-day variability of RMS forecast errors for RASP over the sixty-day period.

Fig. 8 (a)–(b) Initial analyses and 120-hour forecasts for the Australian global model (GASP), (c)–(d) Initial analyses and 120-hour forecasts for the ECMWF global model (EC). Note the large divergence between the forecasts particularly over southeastern Australia.

ment of non-linear systems analysis (see, for example, Eckman and Ruelle (1985)) that techniques such as the Grassberger–Procaccia theory (Grassberger and Procaccia 1984) appeared. Regional studies such as the predictability of tropical cyclones in the Australian region (Fraedrich and Leslie 1989) were then made possible. Many further predictability studies need to be done using the long observational time series available to meteorologists. However, as shown below, much simpler procedures also have been developed for providing estimates of the predictability characteristics of NWP systems, many of which are based on the ideas contained in the work of Kalnay and colleagues (see, for example, Kalnay and Dalcher (1987)).

Leslie et al. (1989) showed that a priori estimates of MFE in the Australian Bureau of Meteorology’s regional NWP forecasts could be produced with a sufficient level of accuracy and reliability to provide, along with the forecast, a prediction of the skill of that forecast. This was achieved by correlating the observed MFE with several predictors based on the initial analysis and
the model forecast itself. A correlation of 0.71 was found between the MFE of the 36-hour operational forecasts and an optimal linear combination of the predictors, thereby explaining approximately 50 per cent of the variance in the MFE. Furthermore, the predictors were able to discriminate very well between good and poor model forecasts. An example of the performance of the technique for forecasting forecast skill is given in Fig. 9, on this occasion for a good forecast over southeastern Australia. Therefore along with the forecast itself (Fig. 9(a)), an estimate of the skill of that forecast is provided (Fig. 9(b)). An alternative use of the *a priori* estimate of the MFE is to correct the model forecast itself. This procedure was used operationally in the late 1980s (see Fig. 1) with considerable success. It was removed when the assimilation procedure was introduced, but is under consideration again for the new model.

### Future directions

Attempting to predict the future directions of NWP in Australia is fraught with difficulties, given the large number of unknowns present. A simple glance at previous decades, the 1960s, 1970s and 1980s, reveals immediately how different one decade is from its predecessor. In that sense, future directions might best be expressed as a set of aims, tempered by predictions of some of the major forces acting on the development of operational NWP systems.

We have identified a number of areas which we would like to address, and we do so immediately below.

### Advances in numerical algorithms

Research into numerical algorithms suitable for application to NWP models is an on-going process leading to improvements in both model accuracy and efficiency. We examine below methods which are currently in the process of being implemented in, or may have application to, NWP models as they evolve through the nineties.

Although elliptic equations are not present in the primitive equation set, they often arise when semi-implicit techniques are applied. Their fast solution is vital for overall model efficiency and recent work has demonstrated the efficacy of the conjugate-gradient (see, for example, Kapitza and Eppel (1987)) and multi-grid (see, for example, Fulton et al. (1986)) approaches for this work. The accuracy of finite-difference discretisations has always been a concern and high-order schemes (mentioned above) are beginning to be introduced in several models around the world (e.g. Bureau of Meteorology, NMC Washington). Traditional numerical advection algorithms produce spurious over or under-shoots in the advected quantity when strong gradients are present. These false extrema may result in the advected field taking on spurious negative values which is, of course, particularly damaging for quantities such as water vapour. To eliminate this problem sophisticated 'positive-definite' and 'monotone' advection algorithms have been developed for both Eulerian (see, for example, Smolarkiewicz and Clark (1986)) and semi-Lagrangian applications (see, for example, Williamson and Rasch (1989)). An alternative means for improving the efficiency or accuracy of NWP models concerns itself not with the methods used to approximate the equations, but rather with the best choice of the distribution of grid-points in both space and time. These 'adaptive-grid' methods have been applied to idealised meteorological modelling by Skamarock et al. (1989) and Dietachmayer and Droegemeier (1992). NWP application of these schemes awaits the solution of several difficult problems, for example the question of the design of physical parametrisations on grids which possess spatial variations in resolution.
Rationalisation
One of the key problems facing NWP centres around the world is the need to achieve as much as possible from the available resources. In the case of operational NWP systems, rationalisation of resources has a number of forms. Referring to the models themselves, clearly it is desirable to have as few models as possible at a given institution. Some centres have achieved this aim very well. For example, the UK Meteorological Office uses the same model for its global and limited area forecasts. On the other hand, the US National Meteorological Centre has a spectral global model and a finite-difference regional model.

In Australia we now have three different NWP systems running in real time. These are the two operational models, the global spectral model (GASP) and the regional finite-difference model (RASP), and the tropical finite-difference model (Puri et al. 1992). Clearly it is desirable to avoid a proliferation of limited area models and this will be achieved by the development of a generalised limited area model. A further goal is to unify where possible the global and limited area systems, for example, by using the same analysis scheme and physics package.

Finally, in order to rationalise interaction with other institutions, as well as to simplify transitions to alternate hardware, model code should move towards being written in highly portable form so that it may be transferred readily between different models. Attempts already have begun to set standards for code so that interchange between institutions and models is facilitated (for example, see Kalnay et al. (1989)).

Model resolution
It generally has been regarded that model skill increases as model resolution increases. In turn, model resolution has been linked inextricably with advances in computer power and numerical algorithms designed to make best use of the increases in computational speed. Hydrostatic regional NWP systems are now running in real time at resolutions down to 30 km in the horizontal (Mesinger 1991), and there is a clearly established need for model resolutions as small as 2 km or less for problems such as real-time pollution or fog and low-cloud forecasting. At such high resolutions, questions arise such as whether non-hydrostatic models should replace hydrostatic models. This question will be addressed in a separate section below.

While noting that model skill is a function of model resolution, it should be noted also that the increase in skill can be realised only if the data base is adequate. Apart from some particular kinds of flow patterns, NWP models cannot be expected to produce forecasts at resolutions beyond those supported by the resolution of the analysis scheme. Predictability experiments on the mesoscale are needed to determine the exact data requirements for the prediction of mesoscale systems.

In Australia, as mentioned above, the resolution of the regional NWP model will be improved from the present 150 km, 16 levels to at least 75 km, 20 levels in 1992/1993 and with future computer upgrades and/or advances in code development, should improve further in the 1990s to at least 20 km horizontally and more than 30 levels in the vertical. Further increases in horizontal resolution to 2 km horizontally will require the relaxation of current assumptions in the model, and are planned to occur as discussed below in the outlook for mesoscale modelling.

Observations
Operational NWP has been restricted throughout its history by limitations of data-base adequacy. In fact, as the accuracy of regional NWP models has improved, the importance of the initial conditions has become more noticeable. In Australia, the data sparsity and distribution problems have been acute. Together with possible intrinsic predictability differences between the hemispheres, the data problem has seen the skill of Australian NWP models lag behind the skill exhibited by NWP models in the northern hemisphere.

However, observing technology is emerging in the 1990s that gives support to hopes for steady improvement in data assimilation specifications of the initial state for NWP systems in the 1990s. There will be a data explosion in the 1990s following major advances in observing systems such as satellite and aircraft-borne instruments, Doppler radars, automatic weather stations and wind profilers.

Data assimilation methods
Along with the opportunities presented by the anticipated dramatic increase in data comes the challenge to make best use of that data. In recent years, meetings on NWP systems have been dominated by developments in data assimilation techniques. The two methods receiving most attention at present are, as mentioned above, the nudging and the variational (adjoint) techniques. The adjoint technique is seen as the possible next-generation data assimilation scheme. Real-time implementation has not yet occurred, mainly because of issues of complexity and computational expense. However, it is anticipated that in the 1990s a number of major NWP centres, such as the ECMWF and NMC, will indeed introduce variational methods for continuous data assimilation. At this point in time, adjoint versions of the ECMWF and NMC adiabatic code already have been produced (respectively: M. Miller, personal communication; J. Derber, personal communication). The adjoint of the physical parametrisations also is underway, but poses some important theoretical problems that have
not yet been solved, for example, 'on-off' physical processes.

**Massively parallel computing**

It is generally accepted that significant increases in computer processing power will not be achieved by simply scaling up current architectures, but rather by the introduction of machines with many thousands of processors, the so-called 'massively parallel' computers. Although they promise eventual teraflop performance, this speed can only be achieved if a sufficiently close mapping can be made between the model software and the computer hardware. Unfortunately, this mapping may be difficult to achieve for two reasons. Firstly, owing to their size and complexity, NWP models have tended to develop in an evolutionary fashion, each new version often based closely on the code from the previous one. Whilst this approach has the enormous advantage of conserving limited man-hours of development time, it also 'ages' the code and acts to preserve those coding practices designed to maximise efficiency on old slow/small-memory machines at the expense of simplicity of the model formulation as a whole. Codes such as these hide the inherent parallelism present in the finite-difference calculations and thus can render automatic parallelising by the compiler ineffective. Second, even if a model was rewritten from scratch, the on-off nature of many of the physical parametrisations implies that different calculations are to be performed on different parts of the grid, which would appear to violate the parallel computing principle of 'computing on all the data at once'. This problem is reduced but not eliminated on MIMD (multiple-instruction multiple-data) machines.

One way to avoid these problems may be to replace traditional FORTRAN with a language that encourages good coding practices amongst the modelling community, and is designed for implementation on parallel machines. Experiments using the SISAL language applied to both spectral and finite-difference barotropic models have yielded encouraging results (see, for example, Chang and Egan (1990a, 1990b)), and similar work using the new DPML language is underway using models from the CSIRO Division of Atmospheric Research (Dix et al. 1992). The advantage of the DPML language in particular is that it transfers the burden of adjusting to changes in architecture from the modellers, to the computer scientists, where it more properly belongs. At this stage however, no major NWP centre has shown any interest in moving away from FORTRAN based modelling systems. The New Zealand Meteorological Service has demonstrated the performance that can be achieved when a model is designed from the outset with parallel applications in mind. Timing experiments on their model show a performance increase proportional to the number of processors to within two per cent over a range of one to nine processors (Purnell and Revell, personal communication). With such a small overhead, they expect that their model could effectively use hundreds of processors. However, they have yet to duplicate this performance in a 'complete' model; for example, the current version is without a physics package.

**Small-scale atmospheric modelling**

Although the operational version of the regional prediction system has, as yet, been run only at relatively coarse resolutions, its utility in modelling smaller-scale atmospheric phenomena has been examined by a number of workers using research versions of the code. Using a modified physics package together with 50 km horizontal resolution, McInnes (1991) and McInnes and McBride (1991) applied the model to the simulation of the well known 'Southerly Baster' that affects the coastal regions of NSW. They found that not only was the model able to predict whether or not a given synoptic-scale environment would lead to a Baster event, it was able also to capture much of the detailed structure of the surge. Another problem area is that, in common with most operational systems of relatively coarse resolution, RASP tends, in general, to underpredict rainfall amounts. Hess (1990) demonstrated that although 150 km resolution suffices to model the initial formation of east-coast cyclones (in agreement with Leslie et al. (1987) and Lynch (1987)), mesoscale processes are important to the subsequent intensification and amount of rainfall produced by the storm. Using a 15 km grid Hess obtained much higher rainfall amounts. Furthermore, a recent study by Mills and Russell (1992) of a flood event over eastern Australia also emphasised the importance of increased horizontal resolution, in both the model and the data, with significant improvements in forecast accuracy achieved by the use of enhanced topography and a high resolution sea-surface temperature analysis.

From an operational perspective, the above results indicate that RASP will be a very important forecasting tool for the foreseeable future. Indeed, as time progresses and increases in computer power allow higher and higher resolutions, RASP, in steadily improved form, will provide qualitatively different forecast guidance as it acquires the ability to predict weather elements directly. Nevertheless, the RASP model resolution cannot be increased indefinitely as the hydrostatic approximation breaks down for grid spacings of around 2 km or less (Physick 1988). With models at this sort of resolution predicted to be running operationally early in the next century (McPherson 1991), there is now considerable interest in the development of nonhydrostatic models. Moreover, Tanguay et al. (1990) have demonstrated that nonhydrostatic models need be no more expensive than their hydrostatic counter-
parts, and hence the drive towards model rationalisation may mean that, eventually, nonhydrostatic models will be responsible for both regional and local NWP. Golding (1992) has applied the UKMO nonhydrostatic mesoscale model at about 5 km grid-spacing to a variety of Australian weather systems. His results show that features with length scales of the order of 20 km were qualitatively correct, but the finer scales are very sensitive, in particular to details of the diabatic forcing. A nonhydrostatic model is currently being developed within BMRC. At this stage the adiabatic component of the code has been tested (Dietachmayer 1992), and work is beginning on the incorporation of physics packages for assessment.

Acknowledgments

The authors would like to thank Michael Manton for his careful reading and constructive criticisms of the original manuscript, and to Brian Golding, Dale Hess, John McBride and Graham Mills for useful discussions. The authors also would like to express their gratitude to John McGregor for his close reviewing of the manuscript and for the improvements that resulted from his comments.

References


