*Quart. J. R. Met. Soc.* (1986), **112**, pp. 1177–1194 551.509.313

# Analysis methods for numerical weather prediction

By A. C. LORENC
*Meteorological Office, Bracknell*

(Received 4 November 1985; revised 13 March 1986)

SUMMARY

Bayesian probabilistic arguments are used to derive idealized equations for finding the best analysis for numerical weather prediction. These equations are compared with those from other published methods in the light of the physical characteristics of the NWP analysis problem; namely the predetermined nature of the basis for the analysis, the need for approximation because of large-order systems, the underdeterminacy of the problem when using observations alone, and the availability of prior relationships to resolve the underdeterminacy.

Prior relationships result from (1) knowledge of the time evolution of the model (which together with the use of a time distribution of observations constitutes four-dimensional data assimilation); (2) knowledge that the atmosphere varies slowly (leading to balance relationships); (3) other nonlinear relationships coupling parameters and scales in the atmosphere.

Methods discussed include variational techniques, smoothing splines, Kriging, optimal interpolation, successive corrections, constrained initialization, the Kalman–Bucy filter, and adjoint model data assimilation. They are all shown to relate to the idealized analysis, and hence to each other. Opinions are given on when particular methods might be more appropriate. By comparison with the idealized method some insight is gained into appropriate choices of parameters in the practical methods.

## 1. INTRODUCTION

The problem of determining the best initial conditions for numerical weather prediction (NWP) is of great practical importance, and has been the subject of many studies by people from different backgrounds. Not surprisingly these studies result in apparently rather different methods; however, on closer inspection they can usually be shown to be closely related. In this paper we derive a general form for the 'best' analysis, emphasizing the particular characteristics of the practical NWP analysis problem, and we relate this to other published methods. A thorough review of earlier work is not attempted in this introduction; the reader is referred to papers in Bengtsson *et al.* (1981) and Williamson (1982) (especially Wahba (1982), which relates many of the ideas discussed here), as well as the papers referenced in section 4.

One practical definition of the best analysis is that which gives the best subsequent forecast. However, forecast models are imperfect and subject to improvements, so it is better to define the best analysis so as to exclude the possibility of the analysis compensating for errors in the forecast model. It is necessary, however, to consider the particular forecast model to be used; this has a discrete basis for its model state which we will denote throughout this paper by a vector $\mathbf{x}$ with dimension $N_x$. Similarly we represent the observed data by a vector $\mathbf{y}$ of dimension $N_y$. The analysis required for NWP must be expressed in terms of this model basis. It must be consistent with the actual observed values $\mathbf{y}_o$. We need to know the relationship between the model basis $\mathbf{x}$ and the observations $\mathbf{y}$; we assume that this can be represented by an explicit operator $K_n$ such that

$$\mathbf{y} = K_n(\mathbf{x}) \tag{1}$$

is the best estimate of $\mathbf{y}$ for a given $\mathbf{x}$. The subscript n reminds us that $K_n$ might be nonlinear, although it is often just a simple interpolation operator. The precise meteorological interpretation of $\mathbf{x}$, $\mathbf{y}$ and $K_n$ is discussed further in section 3(*a*).

Our analysis problem is thus to find the 'best' $\mathbf{x}$ which inverts (1) for a given $\mathbf{y}_o$, allowing for observation errors and other prior information; this is a discrete inverse

problem (Menke 1984). In keeping with inverse theory nomenclature, the function $K_n$ is often referred to as the 'forward process' or 'forward model'. However, in our NWP context this can lead to confusion with the forecast model, so in this paper we will use the term 'generalized interpolation' instead. In section 2 we derive, using Bayesian methods, a generalized inverse to (1) in terms of probability distribution functions (PDFs) which are in principle estimatable. An explicit solution is derived for the case when $K_n$ is linearizable and the PDFs are Gaussian.

In section 3 the characteristics of the NWP analysis problem which distinguish it from other inverse problems in the physical sciences are discussed. These are (a) the use of a predetermined basis **x**, with very large order $N_x$; (b) the search for an inverse which is in general underdetermined by the observations alone; and (c) the availability of much useful prior knowledge which enables us to resolve the underdeterminacy. These features need to be considered in practical implementations of the equations of section 2. This we do in section 4, while comparing our analysis equations with those used in other published methods. We show that the methods are all related, and clarify some of their properties by regarding them as approximations to the idealized equations of section 2. We finish by deriving an iterative solution to the ideal generalized analysis problem, which is practically possible and which uses prior knowledge about the atmospheric state, its nonlinear evolution, and its balance. Finally in section 5 we summarize.

## 2. Derivation of analysis equations

In this section we proceed from a completely general Bayesian analysis equation, (3), expressed in terms of multi-dimensional probability distribution functions, through a fairly standard set of assumptions, to a variational equation for the 'best' analysis, (22). Linearized, this can be solved to give explicit formulae for the best analysis, (26, 27, 28), and its expected error variance, (29). The Bayesian approach is not original, nor are the equations which are derived. References to earlier work are given in section 4. Our objective is to show that one logical progression can lead to all the various equations given, showing what assumptions and approximations are implicit in equations which are often justified directly as a constrained variational best fit, (22), or as a minimum variance best estimate, (28). The details of the derivation of (22) are not essential for understanding the rest of this paper.

Bayes' theorem states that the posterior probability of an event A occurring, given that event B is known to have occurred, is proportional to the prior probability of A, multiplied by the probability of B occurring given that A is known to have occurred:

$$P(\mathrm{A}|\mathrm{B}) \propto P(\mathrm{B}|\mathrm{A})\, P(\mathrm{A}). \tag{2}$$

This is clearly applicable to our inverse analysis problem. If A is the event $\mathbf{x} = \mathbf{x}_t$ and B is the event $\mathbf{y} = \mathbf{y}_o$, then

$$P(\mathbf{x}{=}\mathbf{x}_t|\mathbf{y}{=}\mathbf{y}_o) \propto P(\mathbf{y}{=}\mathbf{y}_o|\mathbf{x}{=}\mathbf{x}_t)\, P(\mathbf{x}{=}\mathbf{x}_t). \tag{3}$$

Subscript t indicates the true value, and o indicates the observed value. (3) defines an $N_x$-dimensional PDF, which we shall call $P_a(\mathbf{x})$, specifying all we can know about the analysis. We will later decide that the 'best' estimate $\mathbf{x}_a$ is either the mean or the mode of $P_a(\mathbf{x})$, i.e. either

$$\mathbf{x}_a = \int \mathbf{x}\, P_a(\mathbf{x})\, \mathbf{dx} \tag{4}$$

or

$$\mathbf{x}_a = \mathbf{x} \text{ such that } P_a(\mathbf{x}) \text{ is maximum.} \tag{5}$$

These are, respectively, the minimum variance and the maximum likelihood estimates of $\mathbf{x}_a$. For a complete solution to the generalized inverse problem we need to know also the accuracy of $\mathbf{x}_a$; this information too is contained in $P_a(\mathbf{x})$.

The prior probability $P(\mathbf{x}=\mathbf{x}_t)$ encapsulates our knowledge about $\mathbf{x}$ before the observations are taken. We shall write this in terms of deviations from some known background $\mathbf{x}_b$ (remembering, however, that it might be a function of $\mathbf{x}_t$ if there is a nonlinear relationship between likely values of the elements of $\mathbf{x}$):

$$P(\mathbf{x}=\mathbf{x}_t) = P_b(\mathbf{x} - \mathbf{x}_b). \tag{6}$$

Any physical observation is subject to random observational errors. If we know that the true values of $\mathbf{y}$ are given by $\mathbf{y}_1 = \mathbf{y}_t$ then these errors can be expressed by

$$P(\mathbf{y}=\mathbf{y}_o|\mathbf{y}_1=\mathbf{y}_t) = P_o(\mathbf{y}_o - \mathbf{y}_t). \tag{7}$$

Again, $P_o$ might also be a function of $\mathbf{x}_t$, or even of $\mathbf{x}_b$. For (3) we need to know $P(\mathbf{y}=\mathbf{y}_o|\mathbf{x}=\mathbf{x}_t)$, which we shall write as $P_{of}(\mathbf{y}_o - \mathbf{y}_f)$ where $\mathbf{y}_f = K_n(\mathbf{x}_t)$. If the generalized interpolation $K_n$ is exact then $\mathbf{y}_f = \mathbf{y}_t$ and

$$P_{of}(\mathbf{y}_o - \mathbf{y}_f) = P_o(\mathbf{y}_o - \mathbf{y}_f). \tag{8}$$

But in general $K_n$ is inexact; we shall describe its error by

$$P(\mathbf{y}_1=\mathbf{y}_t|\mathbf{x}=\mathbf{x}_t) = P_f(\mathbf{y}_1 - \mathbf{y}_f). \tag{9}$$

Now $P(\mathrm{A}) = \int P(\mathrm{A}|\mathrm{B})\, P(\mathrm{B})\, db$, where B is the event that some parameter has a value between $b$ and $b + db$. So

$$P(\mathbf{y}=\mathbf{y}_o|\mathbf{x}=\mathbf{x}_t) = \int P(\mathbf{y}=\mathbf{y}_o|\mathbf{x}=\mathbf{x}_t \text{ and } \mathbf{y}_1=\mathbf{y}_t)\, P(\mathbf{y}_1=\mathbf{y}_t|\mathbf{x}=\mathbf{x}_t)\, \mathbf{dy}_1 \tag{10}$$

$$P_{of}(\mathbf{y}_o - \mathbf{y}_f) = \int P_o(\mathbf{y}_1 - \mathbf{y}_o)\, P_f(\mathbf{y}_1 - \mathbf{y}_f)\, \mathbf{dy}_1. \tag{11}$$

Substitution into (3) then gives

$$P_a(\mathbf{x}) \propto P_{of}(\mathbf{y}_o - K_n(\mathbf{x}))P_b(\mathbf{x} - \mathbf{x}_b) \tag{12}$$

$$P_a(\mathbf{x}) \propto \left\{ \int P_o(\mathbf{y}_1 - \mathbf{y}_o)\, P_f(\mathbf{y}_1 - K_n(\mathbf{x}))\mathbf{dy}_1 \right\} P_b(\mathbf{x} - \mathbf{x}_b). \tag{13}$$

In using this nomenclature we have anticipated the assumption that $P_b(\mathbf{x} - \mathbf{x}_b)$ and $P_{of}(\mathbf{y}_o - K_n(\mathbf{x}))$ are independent, i.e. that background errors and observational errors are uncorrelated. It is possible to work through the following derivation without this assumption, by considering their joint PDF.

To proceed further we need to specify the PDFs $P_b$, $P_o$ and $P_f$ of (13), and solve (4) or (5). A common assumption, which simplifies the solution, is that the PDFs are multi-dimensional Gaussian functions:

$$P_b(\mathbf{x} - \mathbf{x}_b) \propto \exp[-\tfrac{1}{2}(\mathbf{x} - \mathbf{x}_b)^{\mathrm{T}}\mathbf{B}^{-1}(\mathbf{x} - \mathbf{x}_b)] \tag{14}$$

$$P_o(\mathbf{y} - \mathbf{y}_o) \propto \exp[-\tfrac{1}{2}(\mathbf{y} - \mathbf{y}_o)^{\mathrm{T}}\mathbf{O}^{-1}(\mathbf{y} - \mathbf{y}_o)] \tag{15}$$

$$P_f(\mathbf{y} - \mathbf{y}_f) \propto \exp[-\tfrac{1}{2}(\mathbf{y} - \mathbf{y}_f)^{\mathrm{T}}\mathbf{F}^{-1}(\mathbf{y} - \mathbf{y}_f)]. \tag{16}$$

It can be shown that for these Gaussian PDFs $\mathbf{B}$, $\mathbf{O}$, $\mathbf{F}$ are covariance matrices, so that

$$\mathbf{B} = \langle (\mathbf{x}_b - \mathbf{x}_t)(\mathbf{x}_b - \mathbf{x}_t)^T \rangle \tag{17}$$

$$\mathbf{O} = \langle (\mathbf{y}_o - \mathbf{y}_t)(\mathbf{y}_o - \mathbf{y}_t)^T \rangle \tag{18}$$

$$\mathbf{F} = \langle \{\mathbf{y}_t - K_n(\mathbf{x}_t)\}\{\mathbf{y}_t - K_n(\mathbf{x}_t)\}^T \rangle \tag{19}$$

also that $P_{of}$, defined in (11), is given by

$$P_{of}(\mathbf{y}_o - \mathbf{y}_f) \propto \exp[-\tfrac{1}{2}(\mathbf{y}_o - \mathbf{y}_f)^T(\mathbf{O} + \mathbf{F})^{-1}(\mathbf{y}_o - \mathbf{y}_f)]. \tag{20}$$

Furthermore the minimum variance and maximum likelihood estimates (4) and (5) are identical. For the maximum likelihood estimate we wish to maximize

$$\begin{aligned} P_a(\mathbf{x}) &= P_{of}(\mathbf{y}_o - K_n(\mathbf{x}))P_b(\mathbf{x} - \mathbf{x}_b) \\ &\propto \exp[-\tfrac{1}{2}\{\mathbf{y}_o - K_n(\mathbf{x})\}^T(\mathbf{O} + \mathbf{F})^{-1}\{\mathbf{y}_o - K_n(\mathbf{x})\} - \tfrac{1}{2}(\mathbf{x} - \mathbf{x}_b)^T\mathbf{B}^{-1}(\mathbf{x} - \mathbf{x}_b)]. \end{aligned} \tag{21}$$

Maximizing $P_a$ is equivalent to minimizing $-\ln(P_a)$, i.e. to minimizing

$$J = \{\mathbf{y}_o - K_n(\mathbf{x})\}^T(\mathbf{O} + \mathbf{F})^{-1}\{\mathbf{y}_o - K_n(\mathbf{x})\} + (\mathbf{x} + \mathbf{x}_b)^T\mathbf{B}^{-1}(\mathbf{x} - \mathbf{x}_b). \tag{22}$$

If we write $\mathbf{K}$ for the matrix of partial derivatives of $K_n$ with respect to the components of $\mathbf{x}$, then

$$K_n(\mathbf{x} + d\mathbf{x}) = K_n(\mathbf{x}) + \mathbf{K}\, d\mathbf{x} \tag{23}$$

and that $\mathbf{x}_a$ which minimizes $J$ is given by

$$0 = \mathbf{K}^T(\mathbf{O} + \mathbf{F})^{-1}\{\mathbf{y}_o - K_n(\mathbf{x}_a)\} + \mathbf{B}^{-1}(\mathbf{x}_b - \mathbf{x}_a). \tag{24}$$

If the linearization of $K_n$ is valid in the entire range of probable values of $\mathbf{x}$ in the region of $\mathbf{x}_b$, then an explicit solution can be found by evaluating $\mathbf{K}$ at $\mathbf{x} = \mathbf{x}_b$ and writing

$$K_n(\mathbf{x}_a) = K_n(\mathbf{x}_b) + K(\mathbf{x}_a - \mathbf{x}_b). \tag{25}$$

Substitution in (24) gives an explicit solution which can be manipulated into several forms, three of which we will be referring to later:

$$\mathbf{x}_a = \mathbf{x}_b + \{\mathbf{B}\mathbf{K}^T(\mathbf{O} + \mathbf{F})^{-1}\mathbf{K} + \mathbf{I}\}^{-1}\mathbf{B}\mathbf{K}^T(\mathbf{O} + \mathbf{F})^{-1}\{\mathbf{y}_o - K_n(\mathbf{x}_b)\} \tag{26}$$

$$\mathbf{x}_a = \mathbf{x}_b + \mathbf{B}\mathbf{K}^T(\mathbf{O} + \mathbf{F})^{-1}\{\mathbf{K}\mathbf{B}\mathbf{K}^T(\mathbf{O} + \mathbf{F})^{-1} + \mathbf{I}\}^{-1}\{\mathbf{y}_o - K_n(\mathbf{x}_b)\} \tag{27}$$

$$\mathbf{x}_a = \mathbf{x}_b + \mathbf{B}\mathbf{K}^T(\mathbf{K}\mathbf{B}\mathbf{K}^T + \mathbf{O} + \mathbf{F})^{-1}\{\mathbf{y}_o - K_n(\mathbf{x}_b)\}. \tag{28}$$

The expected analysis error covariance for this linearized Gaussian case is given by

$$\langle (\mathbf{x}_a - \mathbf{x}_t)(\mathbf{x}_a - \mathbf{x}_t)^T \rangle = \mathbf{B} - \mathbf{B}\mathbf{K}^T(\mathbf{K}\mathbf{B}\mathbf{K}^T + \mathbf{O} + \mathbf{F})^{-1}\mathbf{K}\mathbf{B} \tag{29}$$

and the solution $\mathbf{x}_a$ also minimizes this.

Iterative solution methods for (24) can also be found, usually related in form to (26) or (27). By reevaluating $K_n(\mathbf{x})$ and if necessary $\mathbf{K}$ each iteration, nonlinearity in $K_n$ is allowed for. By varying (for instance) $\mathbf{O}$, non-Gaussian errors can be allowed for. By approximating the large matrix inverse in (26) or (27) computationally practical methods can be found. These methods are discussed further in section 4.

Throughout this derivation we have assumed that matrices such as $\mathbf{B}$ and $\mathbf{B}^{-1}$ are non-singular. If one of these is singular a similar derivation can be achieved by the expedient of replacing it by a modified matrix with the same eigenvectors and with zero eigenvalues replaced by $\varepsilon$, and letting $\varepsilon$ tend to zero as the final step. Note that (28) and (29) do not require $\mathbf{B}$ to be non-singular if $\mathbf{O}$ is non-singular, as is usually the case for random observational errors. So, except when there are perfect, redundant observations, the limiting process converges.

## 3. Characteristics of the NWP analysis problem

Equations like those of section 2 have been used to analyse data in many branches of the physical sciences, for example in the inversion of satellite-observed radiances to derive temperature profiles (Rodgers 1976), and in geophysics and oceanography (Menke 1984; Marshall 1985). We now discuss the aspects of the NWP problem which distinguish it from other applications.

### (*a*) *Pre-defined discrete basis*

The first important assumption in our approach is that a discrete basis for the model state has been pre-defined independently of the analysis process. Note that we require a precise definition; we assume the existence of a unique $\mathbf{x}_t$ determined by the state of the real atmosphere. (Some grid-point models do not have a precise definition; values being used sometimes as grid box averages and sometimes as spot values.) $\mathbf{x}_t$ is the vector of coefficients obtained by projecting the true state of the atmosphere onto the model basis. The definition of the basis has implications for the generalized interpolation $K_n$ even in the simplest case when $K_n$ is just a spatial interpolation from grid points to observation positions. In the general case any basis can be used, e.g. components of a spectral model, with $K_n$ representing the evaluation of values at observation positions from the coefficients. Also any observations can be used; for satellite radiance observations $K_n$ is the radiative transfer equation from the model temperature profile.

The externally specified discrete basis (on criteria of accuracy of NWP, and computing economy) removes the need to discuss trade-offs between resolution and accuracy, an important consideration in other fields. It also enables us to use simpler matrix methods, rather than methods based on function theory.

Having distinct model and observational bases clarifies questions of resolution and representativeness which arise in some methods (for instance optimum interpolation, see section 4(*c*)). Because of the definition of $\mathbf{x}_t$, the background error covariance of $\mathbf{x}_b - \mathbf{x}_t$ (**B**) does not contain errors of representativeness. Similarly the observational error covariance of $\mathbf{y}_o - \mathbf{y}_t$ (**O**) contains only instrumental errors. The errors of representativeness which are often included in the observational error are now separated into the errors **F** of the generalized interpolation $K_n$.

Since an NWP model uses the same basis for initial conditions and forecast, the generalized interpolation $K_n$ is appropriate for both. For instance if surface air temperature and wind forecasts are derived from a NWP forecast whose basis does not contain them, then for consistency the same $K_n$ should be used when analysing observations of the same parameters.

### (*b*) *Need for approximation*

Having derived suitable analysis equations, we might in many problems obtain solutions using standard computer techniques. However, in the NWP case the best forecasts usually come from models for which the storage and processing of the vector **x** of dimension $N_x$ uses up most of the available computer resources. One cannot envisage processing matrices such as **B** with its $N_x^2$ elements, and approximations must be made. In fact many successful analysis schemes have been constructed empirically, without reference to idealized solutions. However, an understanding of how approximations relate to the ideal can guide us in their development and shorten lengthy empirical tuning.

### (*c*) *Underdeterminacy: the need for prior information*

With the continuing increase in computer power, there has been a corresponding

increase in the size and complexity of NWP forecast models. Important weather phenomena such as fronts and severe storms need very high resolutions to forecast them well. It seems that for some time to come the size of models ($N_x$) will continue to increase as fast as computing considerations allow. Thus despite progress in remote sensing and other automatic observing systems, which is increasing the number of data ($N_y$), it seems likely that the NWP analysis problem will remain underdetermined, with $N_x > N_y$. Certainly we will never have observations for every model value, and the problem will be at least partially underdetermined.

Formally in section 2 we resolve this underdeterminacy by assuming prior knowledge to justify constraining **x** on or near a surface or point in the $N_x$-dimensional space of **x**. This is expressed generally by (6), and for Gaussian PDFs by (14). We shall see in section 4(*a*) how any linear constraints and prior knowledge about **x** can be incorporated into $\mathbf{x}_b$ and **B** and how nonlinear constraints and relationships can be handled. In the rest of this section we discuss the physical sources of prior knowledge for our NWP problem, in addition to the assumptions of smoothness and closeness to the long-term norm (climatology) that are commonly used for this sort of problem.

### (*d*) *Time evolution*

Since the atmosphere is being regularly observed, one way of increasing $N_y$ is to consider data from a longer time period. The equations of section 2 are general; the model basis of **x** is not restricted to an instantaneous three-dimensional picture. We could consider the basis to be a four-dimensional array in space and time. In this case to increase the time period of useful observations would lead to a corresponding increase in $N_x$, and the underdeterminacy problem would not be resolved. However, the fact that we are attempting NWP implies that we know relationships for the time evolution of model fields. These can be used as constraints on the possible four-dimensional values of **x**. Thus use of a space and time distribution of observations together with knowledge of evolution equations can resolve the underdeterminacy; the process of achieving this has become known as four-dimensional data assimilation (Bengtsson 1975).

### (*e*) *Balance*

It is an observed fact that atmospheric fields are usually slowly varying, considering the magnitudes of the individual forces involved; i.e. that the forces are in approximate balance. The relationships this implies can be used to constrain **x** and reduce the underdeterminacy. Techniques to achieve such a balance after the analysis and before a forecast are called initialization. Ideally the analysis should use all prior information, fitting balance relationships simultaneously with fitting the data. Ways of doing this using initialization techniques as part of the analysis process are discussed in section 4(*g*). A separate initialization step after the analysis is useful only if the analysis method is imperfect in its use of prior information about balance.

### (*f*) *Nonlinear coupling*

Interactions exist between different scales of motion and parameters of the atmosphere, which means that some scales and parameters are not completely independent but to a large extent determined by forcings from boundaries and other scales and parameters. For instance the boundary layer structure depends largely on the surface heating and the synoptic situation, the high resolution detail of a front depends on frontal dynamics and the large-scale forcing, and the humidity field depends on past evaporation and vertical motion. The pioneering experiment of Charney *et al.* (1969) showed that wind and surface pressure could be deduced from temperature data; Lorenc and Tibaldi

(1980) showed that realistic frontal humidity fields could be deduced from height and wind data; and Errico and Baumhefner (1986) showed that smaller scales could be deduced from surface forcing and larger scales. These relationships reduce the effective number of degrees of freedom to less than $N_x$. They are becoming more important now that models have higher resolutions and more sophisticated representations of physical processes. Although these relationships are in principle not very different from those discussed in sections 3(*d*) and (*e*), they are in practice quite different since their nonlinear effects actually aid the analysis process. Analysis based on a coarser resolution basis, followed by interpolation to finer resolution, cannot fully use them, and it is not possible, for instance, for a linear analysis method, given observational data for the larger scales, to analyse consistent smaller-scale detail for features such as fronts.

## 4. Relationship to other methods

### (*a*) *Variational methods: constraints*

Equations similar to (22) appear in many variational approaches to the analysis of data. The first term measures the fit to the data, the second the fit to additional prior knowledge and constraints. Any relationships which are linear in **x** can be imposed to any desired amount by a penalty function of the form used for the second term of (22), which was called by Sasaki (1970) a weak constraint. We showed in section 2 that Gaussian PDFs (14), (15), (16) lead to a variational problem with $L_2$ norms (22), which has a linear solution (28). Other forms of PDF give other norms and nonlinear solutions. A useful property of $L_2$ norms is that any number of weak constraints can be combined into a single $L_2$ penalty. There is no formal need to consider individually in the analysis information from separate sources, such as climatology $\mathbf{x}_c$ and a forecast $\mathbf{x}_b$; the two can be combined to form a new background $\mathbf{x}_d$, and corresponding error covariance **D**:

$$(\mathbf{x}-\mathbf{x}_b)^{\mathrm{T}}\mathbf{B}^{-1}(\mathbf{x}-\mathbf{x}_b)+(\mathbf{x}-\mathbf{x}_c)^{\mathrm{T}}\mathbf{C}^{-1}(\mathbf{x}-\mathbf{c}_c)=(\mathbf{x}-\mathbf{x}_d)^{\mathrm{T}}\mathbf{D}^{-1}(\mathbf{x}-\mathbf{x}_d)+\text{constant} \quad (30)$$

$$\mathbf{x}_d=\mathbf{D}(\mathbf{B}^{-1}\mathbf{x}_b+\mathbf{C}^{-1}\mathbf{x}_c) \quad (31)$$

$$\mathbf{D}^{-1}=\mathbf{B}^{-1}+\mathbf{C}^{-1}. \quad (32)$$

Other prior information, such as the linearized balance constraint of section 4(*g*), can be similarly merged.

We mentioned at the end of section 2 a technique for coping with singular matrices, by using instead a small value $\varepsilon$, and letting it tend to zero. The same method can be used here. If this is done for matrix **C** then $\mathbf{x}-\mathbf{x}_c$ is constrained to be orthogonal to the corresponding eigenvectors; the same constraint is imposed by using instead $\mathbf{x}_d$ and **D**.

The imposition of constraints on the analysis through the background covariances was discussed by Lorenc (1981), who pointed out that (using our notation) the analysis equation (28) can be rewritten

$$(\mathbf{x}_a-\mathbf{x}_b)^{\mathrm{T}}=\{(\mathbf{y}_o-K_n(\mathbf{x}_b))^{\mathrm{T}}(\mathbf{KBK}^{\mathrm{T}}+\mathbf{O}+\mathbf{F})^{-1}\}(\mathbf{KB}).$$

Thus changes made to the background field are linear combinations of the rows of **KB**, i.e. of the background error covariances between each observation and the model field. Any linear relationship satisfied by these covariance vectors is thus satisfied by the analysis increments. The imposition of balance constraints, and the need (or lack of it) for a separate initialization step was discussed by Phillips (1982) and Wahba (1982). A comprehensive discussion of the relationship between the covariance matrix and weak

and strong constraints is given by Ikawa (1984a and b). We discuss balance constraints further in 4(*g*).

Other methods can be used in variational analysis to impose a relationship exactly (a strong constraint in Sasaki's (1970) terminology). One is called by Le Dimet and Talagrand (1986) reduction of the control variable. If constraints exist which enable certain elements (**v**) of vector **x** to be calculated from the rest (**u**):

$$\mathbf{x}^{\mathrm{T}} = (\mathbf{u}^{\mathrm{T}}, \mathbf{v}^{\mathrm{T}}) \tag{33}$$

$$\mathbf{v} = F(\mathbf{u}) \tag{34}$$

then the constrained analysis problem for **x** reduces to an unconstrained problem for **u**. The generalized interpolation $K_n$ needs to be modified to contain the constraint equation

$$K_n(\mathbf{x}) = K_n\{\mathbf{u}, F(\mathbf{u})\} = K'_n(\mathbf{u}). \tag{35}$$

Methods which analyse the mass field only, calculating the wind field by a balance equation, are using this technique.

An important application is to the data assimilation problem of section 3(*d*). If we assume that we know the prognostic equations sufficiently well to use them as a strong constraint, we can take the general four-dimensional analysis equations and reduce the control variable to a three-dimensional field at one time (the beginning of the period if the NWP forecast model is not reversible in time). The modified generalized interpolation $K_n$ thus consists of two stages—a forecast to the time of observations followed by an interpolation in space only. The four-dimensional assimilation problem is reduced to determining a three-dimensional field; note, however, that the three-dimensional solution together with the NWP forecast model define a four-dimensional analysis which is equally good at other times. This technique is used in section 4(*h*).

### (*b*) *Smoothing splines and Kriging*

A smoothing spline is usually defined as a function $g(t)$ which minimizes a penalty $J$ given by

$$J = (1/N_y) \sum_{i=1}^{N_y} w_i \{y(t_i) - g(t_i)\}^2 + \lambda \int J_s(g)\, dt \tag{36}$$

where $y(t_i)$ is the observational datum at position $t_i$ and $J_s(g)$ is a penalty on smoothness, and $\lambda$ is the relative weight given to the smoothness penalty (Reinsch 1967). For a one-dimensional case, with smoothness related to the $m$th derivative of $g$

$$J_s = \{g^{(m)}(t)\}^2 \tag{37}$$

this leads to polynomial splines of order $2m - 1$. Wahba (1978) showed that this is equivalent to Bayesian estimation with diffuse prior estimates of polynomials of degree less than $m$, i.e. assuming that the mean and first $m - 1$ derivatives of $g$ are entirely determined by the observations, with no useful prior estimates or constraints. If we replace differentiation by the equivalent matrix operation on our finite basis $x$, then it can be seen that (36) is a special case of (22). Wahba and Wendelberger (1980) related $m$ to the rate of fall-off of the power spectrum of the field. Similar relationships can be made for the general form (22), since the covariance is the Fourier transform of the power spectrum.

In the above mentioned paper, and others by Wahba and her collaborators, the variational method is extended to the case where we know the form of the constraints

embodying prior knowledge, but not the coefficients. This procedure, known as generalized cross validation, can for instance be used to find the $\lambda$ and $m$ from (36) and (37) most consistent with the data. It is not limited to splines; for instance Seaman and Hutchinson (1985) used it for calculation of correlation coefficient parameters in statistical (optimal) interpolation.

Splines are an aesthetically pleasing method of estimating a field from noisy data, using as few prior assumptions as possible. For our NWP analysis problem, when we have a predetermined discrete basis and useful prior information, they are less appropriate. In particular, from forecasts and climatology, we often have fairly accurate prior estimates of the mean and mean gradient of the field, while some observing systems can have correlated errors such that the mean is poorly observed. Deriving the mean and mean gradient from the observational data alone can cause a spline-fitting analysis to give poor values when extrapolating beyond the edge of areas with data (e.g. Seaman and Hutchinson 1985).

A common analysis technique used in the mining industry is Kriging (Matheron 1963). This is a minimum variance estimate of deviations from a trend field, which in Universal Kriging (Journel 1977) can be a low-order polynomial obtained by a minimum variance best fit to the observational data. Despite its apparently different derivation this technique can be shown to be equivalent to the use of a polynomial spline.

### (c) *Optimal interpolation, observational errors, and indirect observations*

A technique very commonly used for meteorological analysis is optimal interpolation (OI). As first used (Gandin 1963) it was very similar to Kriging except that some prior knowledge of a climatological background was assumed, and explicit account was taken of observational errors. Generalized to use a forecast background, three-dimensional multivariate data, and constraints such as non-divergence and geostrophy (e.g. Lorenc 1981), it can be written in a matrix form very similar to (28). The important difference is that the background, and hence potentially the analysis, is assumed to be continuous. No account is taken of the generalized interpolation $K_n$, a continuous covariance function replaces **B**, and $\mathbf{KBK}^{\mathrm{T}}$ is replaced by a matrix of this function evaluated at observation positions. This has three effects: (i) If the number of observations used is limited by selection, then the order of matrices to be processed ($N_y$) is much less than that of section 2 ($N_x$). This makes the solution of the equations a practical possibility, although the data selection has other implications, discussed in 4($d$). (ii) The continuous approximation used for $\mathbf{KBK}^{\mathrm{T}}$ imposes differential forms of constraints such as geostrophy, rather than the forms appropriate to the model basis. (iii) Errors in the generalized interpolation (**F**) are not explicitly considered; their effect is obtained by including in **O** 'errors of representativeness'.

These errors of representativeness are justified by redefining 'truth' to contain only those scales that we wish to analyse (e.g. Lorenc 1981). Observational errors quoted for use in OI analysis schemes are thus dependent on the predetermined basis, and should not be compared directly between schemes of different resolution. Equation (28) is more satisfactory in avoiding the somewhat circular definition of 'truth', and clearly separating errors into a part (**O**) dependent on the observing process and a part (**F**) dependent on the generalized interpolation $K_n$ and the resolution of the model basis. When evaluating the covariance function for OI, observations are treated as point values. This is an approximation, since all data are actually averages over sometimes large regions of space and time. Explicit consideration of $K_n$ as in (28) allows for this; **O** is the observational error for these averages, and **F** is the error of the generalized interpolation in estimating these averages. Because OI has in principle a continuous form for the analysis, its prior

estimate (the background), and the background error covariances, it is not immediately apparent which spectral scales should be included in the background error. This is important because the atmosphere has a continuous energy spectrum with significant components at scales unresolved by NWP models. To be consistent we must redefine the background error in terms of the redefined 'truth' used to justify observational errors of representativeness. The forms, in (28), of $\mathbf{KBK}^{\mathrm{T}}$ and $\mathbf{BK}^{\mathrm{T}}$ for the covariance functions show clearly that only scales resolved by the model basis should be included.

Some so-called 'indirect' observations in fact result from an inversion identical to that of section 2. Examples are temperature soundings, derived from satellite radiances, and low-level winds over oceans derived from satellite microwave scatterometers. If satellite radiances are used as data then $K_n$ includes the radiative transfer equation (e.g. Rodgers 1976; O'Sullivan and Wahba 1985). It is theoretically advantageous to combine the inversion and analysis stages if $K_n$ is nonlinear or the errors are non-Gaussian (Lorenc *et al.* 1985; Wahba 1985). For linear $K_n$ and Gaussian errors the sequential inversion and analysis can give the same result as simultaneous analysis and inversion of the basic observational data, in a manner analogous to the combination of penalties from different prior constraints discussed in section 4(*a*), and to the Kalman–Bucy filter discussed in section 4(*h*). Note that this requires a matrix of statistics to be passed from the inversion step to the analysis step, as well as the retrieved indirect observation. The background information used in the inversion also has to be passed, if it is different from that used in the analysis. Many of the awkward error properties of indirect observations, such as the spatial correlations of errors for satellite temperature soundings (Schlatter 1981), are due to errors in the prior information used in the inversion process and to errors in the forward model, rather than to errors in the basic observed quantity. The errors and covariances of indirect observations can in principle be calculated during a statistical inversion procedure using equations like (29), and used in a subsequent analysis.

OI in its general multivariate form can thus be considered as a continuous approximation to our ideal discrete generalized inverse analysis. Not surprisingly therefore, Wahba and Wendelberger (1980) have shown that OI can also be expressed in terms of the solution of a variational problem.

### (*d*) *Data selection*

OI, by modelling $\mathbf{KBK}^{\mathrm{T}}$ directly, reduces the order of matrices to be processed from $N_x$ to $N_y$, and enables a further reduction by data selection. This is achieved in practice by dividing $\mathbf{x}$ into sections, and for each section selecting only a subset of $\mathbf{y}$. This approximation removes two features of the full optimal solution. Firstly, that the analysis $\mathbf{x}_a$ also gives directly the optimal, minimum variance, estimate of any linear combination of the elements of $\mathbf{x}$. For instance the spatial gradient of an analysed field may not be a good estimate, if it is calculated using a difference of elements of $\mathbf{x}$ which were estimated from different subsets of $\mathbf{y}$. Secondly, that linear constraints can be imposed through use of appropriate background and covariances. If these constraints are a good embodiment of the sort of useful prior information discussed in section 3, then this is clearly a loss. However, if the covariance model is itself approximate, as is always the case in practice, then it can be advantageous not to enforce its implications too strongly. For instance Lorenc (1981) showed that because of data selection, large-scale divergences could be analysed even when using a covariance model which assumed non-divergent errors. For simple climatological backgrounds, which contribute little useful prior information, and with no useful prior relationships between analysis variables, it is advantageous to restrict the number of data selected to less than about 10 (Gandin, personal communication) in

order that inappropriate covariances should not adversely affect the analysis. However, with improved covariance models, more can be used (Seaman and Hutchinson 1985).

### (*e*) *Successive corrections*

The successive correction method (Bergthorsson and Doos 1955; Cressman 1959; Barnes 1973) is a very successful empirically based iterative analysis method. The $(u + 1)$th iteration can be expressed

$$\mathbf{x}[u+1] = \mathbf{x}[u] + \mathbf{QW}\{\mathbf{y}_o - K_n(\mathbf{x}[u])\} \tag{38}$$

$\mathbf{W}$ being a $N_x \times N_y$ matrix of weights which depend on the distance between observations and grid points and on the observational errors. $\mathbf{Q}$ is a $N_x \times N_x$ matrix of normalization factors. This is very similar to an iterative solution to (24), related to (26), which can be written

$$\mathbf{x}[u+1] = \mathbf{x}[u] + \mathbf{Q}(\mathbf{W}\{\mathbf{y}_o - K_n(\mathbf{x}[u]) + \mathbf{x}_b - \mathbf{x}[u]\}) \tag{39}$$

where

$$\mathbf{W} = \mathbf{BK}^T(\mathbf{O} + \mathbf{F})^{-1} \tag{40}$$

$$\mathbf{Q} = (\mathbf{WK} + \mathbf{I})^{-1}. \tag{41}$$

The original version (38) omits a term $(\mathbf{x}_b - \mathbf{x}[u])$, forcing the analysis towards the background, it will thus converge towards an exact fit to the observations. Some later implementations have included this. Replacing $\mathbf{BK}^T$ by a continuous function, as in OI, we see that (40) is indeed like the empirically derived scheme. If the observations are regularly distributed and $\mathbf{K}$ is a simple interpolation operator then a fair approximation to $\mathbf{Q}$ is a diagonal matrix such that

$$q_{kk} = \left(\sum_i w_{ki} + 1\right)^{-1}. \tag{42}$$

Here the subscript $k$ indicates a particular element of $\mathbf{x}$ (e.g. one grid point), and the sum over $i$ covers those data selected as influencing it. This is like some versions of the empirical scheme; others omit the 1 (effectively the weight given to the background). In the original empirical schemes, the weighting function used to calculate the elements $w_{ki}$ of $\mathbf{W}$ contained little useful prior information, so in keeping with section 4(*d*), data selection was limited, and furthermore the weighting function was changed for each iteration.

Because of the approximate form necessary in practice for $\mathbf{Q}$, (38) does not converge to the ideal solution as $u$ tends to infinity, nor does it impose any constraints implicit in $\mathbf{B}$. Franke and Gordon (1983) and Bratseth (1986) have shown that a modified form can do so:

$$\mathbf{x}[u+1] = \mathbf{x}[u] + \mathbf{WQ}'\{\mathbf{y}_o - K_n(\mathbf{x}[u])\}. \tag{43}$$

Here $\mathbf{Q}'$ is a $N_y \times N_y$ matrix of normalization factors whose exact form should be

$$\mathbf{Q}' = (\mathbf{KW} + \mathbf{I})^{-1} \tag{44}$$

as can be seen by comparison with (27). A suitable approximation for $\mathbf{Q}'$ will make (43) converge to a solution which imposes any constraints implicit in $\mathbf{B}$ and which fits the data exactly. A solution which allows for observational errors may be found by treating the estimation problem as a two-stage process; firstly finding the best estimate of the true

values for the observed parameters, then interpolating these estimates as if they were exact. (Smoothing splines and OI can also be considered in this way.) The modified successive correction method then becomes

$$\mathbf{y}[u+1] = \mathbf{y}[u] + (\mathbf{O}+\mathbf{F})^{-1}\mathbf{Q}'\{\mathbf{y}[u] - K_n(\mathbf{x}[u])\} \tag{45}$$

$$\mathbf{x}[u+1] = \mathbf{x}[u] + \mathbf{W}\mathbf{Q}'\{\mathbf{y}[u] - K_n(\mathbf{x}[u])\}. \tag{46}$$

With $\mathbf{y}[0] = \mathbf{y}_o$, $\mathbf{x}[0] = \mathbf{x}_b$, and $K_n$ linear, this iteration converges to a solution of (24) for any approximation to $\mathbf{Q}'$ which satisfies the condition that the eigenvalues of $\{\mathbf{I} - (\mathbf{O}+\mathbf{F}+\mathbf{K}\mathbf{B}\mathbf{K}^T)\mathbf{Q}'\}$ must have modulus less than one. This may be proved by considering (45) minus $\mathbf{K}$ times (46), which gives

$$\mathbf{y}[u+1] - \mathbf{K}\mathbf{x}[u+1] = \{\mathbf{I} - (\mathbf{O}+\mathbf{F}+\mathbf{K}\mathbf{B}\mathbf{K}^T)\mathbf{Q}'\}\{\mathbf{y}[u] - \mathbf{K}\mathbf{x}[u]\}. \tag{47}$$

Repeated substitution of (47) into the equation obtained by linearizing $K_n$ in (46), summation of the series, and taking the limit as $u$ tends to infinity, shows that the approximation does converge to (28).

### (*f*) *Non-Gaussian PDFs and quality control*

Purser (1984) showed how simple non-Gaussian error distributions could be handled in an iterative maximum likelihood solution of (12), by using the Gaussian solution, and making the observation error variances used in each iteration depend on the last estimate of $\mathbf{x}_a$. Many meteorological observations have error distributions which can be modelled by Gaussian distributions with enlarged tails due to occasional gross errors. Lorenc (1984) showed that in this case a reasonable approximation is to perform an observation quality control, rejecting extreme data, and treating accepted data as if their error distribution was Gaussian.

### (*g*) *Initialization and balance constraints*

Most early analysis schemes for NWP, designed for extra-tropical latitudes, applied a geostrophic or similar relationship as a strong constraint by a reduction of the control variable, analysing only the mass field. More recently other variational techniques have been used to impose more general balance relationships, such as that implied by nonlinear normal mode initialization (see for example papers in Bengtsson *et al.* (1981) and Williamson (1982)). Since the empirical justification of slowly varying fields and balance in the atmosphere is not strong for all scales of motion, and since numerical methods of representing this balance are only approximate, it seems appropriate to apply our prior knowledge of balance as a weak rather than a strong constraint. If $\mathbf{E}^T\mathbf{x}$ is a projection of $\mathbf{x}$ onto normal modes, and if $\mathbf{x}_e$ is a nonlinearly balanced state near $\mathbf{x}_t$, then with the assumptions usual in normal mode initialization, $\mathbf{E}^T(\mathbf{x} - \mathbf{x}_e)$ should have near-zero elements for fast gravity modes. Thus we can add a penalty term $(\mathbf{x} - \mathbf{x}_e)^T\mathbf{E}\mathbf{W}\mathbf{E}^T(\mathbf{x} - \mathbf{x}_e)$ to the penalty function $J$ in (22). $\mathbf{W}$ is a diagonal matrix with $1/\varepsilon$ for the unwanted gravity modes and $\varepsilon$ elsewhere. Combining this with the term in $\mathbf{x}_b$ and $\mathbf{B}$, using (30), (31) and (32) gives a new background and error covariance matrix. If we let $\varepsilon$ tend to zero then it is easy to show that the new background must satisfy the constraint, and the new covariance matrix must have zero eigenvalues for the unwanted gravity modes. Thus a constraint on balance is roughly equivalent, when linearized, to a combination of two steps which are standard practice in many operational schemes: (1) ensuring that the background is balanced by incorporation of nonlinear normal mode initialization into the analysis–forecast cycle, or by using a damping time integration scheme; (2) modelling the covariances used with geostrophic and non-divergence assumptions, or in some other

way ensuring that deviations of the analysis from the background are in approximate linear balance. It is interesting to note that empirical studies of the error covariances for such a balanced background state (Hollingsworth and Lonnberg 1986) are similar to those obtained assuming equipartition of energy amongst Rossby (i.e. balanced) normal modes (Phillips 1986).

### (*h*) *Data assimilation: the Kalman–Bucy filter and adjoint models*

In section 3(*d*) we defined data assimilation as the resolution of underdeterminacy by the use of observations for a period of time together with knowledge of the equations governing evolution of model states with time. In section 4(*a*) we discussed how this time evolution constraint could be achieved in a generalized 4-dimensional analysis framework by a reduction of the control variable, making $K_n$ consist of a forecast to the observation time, followed by the interpolation of the observed parameters. Let us represent the forecast model as a sequence of discrete forecasting steps:

$$\mathbf{x}_{i+1} = M_{ni}(\mathbf{x}_i) \tag{48}$$

and divide the observations into corresponding groups

$$\mathbf{y}_o{}^T = (\mathbf{y}_{o0}{}^T, \mathbf{t}_{o1}{}^T \ldots \mathbf{y}_{oi}{}^T, \ldots \mathbf{y}_{on}{}^T) \tag{49}$$

each with its own $K_{ni}(\mathbf{x}) = L_{ni}(\mathbf{x}_i)$, where the additional subscripts denote values at a particular time. If we linearize the $M_{ni}$ and $L_{ni}$ in the same way as for (23) then the product rule for partial differentials gives us

$$\mathbf{K}_i{}^T = \mathbf{M}_0{}^T\mathbf{M}_1{}^T\mathbf{M}_2{}^T \ldots \mathbf{M}_{i-1}{}^T\mathbf{L}_i{}^T. \tag{50}$$

These are the discrete adjoint models of our forward process (Le Dimet and Talagrand 1986; Lewis and Derber 1985). Now for our forecast model to be a suitable strong constraint its errors must be small, so we assume that any errors in the forward processes are restricted to the $L_{ni}$, and that observation errors at different times are uncorrelated. (22) can then be written as

$$J = \sum_{i=0}^{n} \{\mathbf{y}_{oi} - K_{ni}(\mathbf{x})\}^T(\mathbf{O}_i + \mathbf{F}_i)^{-1}\{\mathbf{y}_{oi} - K_{ni}(\mathbf{x})\} + (\mathbf{x} - \mathbf{x}_b)^T\mathbf{B}^{-1}(\mathbf{x} - \mathbf{x}_b). \tag{51}$$

For $K_{ni}$ linear it is easy to show, in a manner analogous to that of (30), (31) and (32), that the terms in $J$ can be sequentially combined in any order. In particular the term for $i = 0$ can be combined with the background to give a new background valid at $i = 0$ and associated error covariance incorporating the information from $i = 0$ observations. These can in turn be multiplied by $\mathbf{M}_0$ and combined with the $i = 1$ observations and so on; at each stage the combination is equivalent to a three-dimensional analysis using (28) and (29) followed by a forecast of a new background and covariance. Thus the optimal four-dimensional analysis reduces to a sequence of three-dimensional analyses with appropriate forecasts using $\mathbf{M}_i$ of the background and background error covariance; this is the Kalman–Bucy (K–B) filter (Ghil *et al.* 1981). The K–B filter method can be extended to imperfect and nonlinear models; it is then no longer precisely equivalent to an optimal four-dimensional analysis with a strong constraint on time-evolution.

The big practical disadvantage of the K–B filter method is that it requires knowledge of and processing of very large matrices such as **B**. Approximations such as retaining only a number of diagonals of the matrix can make the method feasible for simple models (Parrish and Cohn 1985), but it seems unlikely that it will be feasible for the very high resolution models with sophisticated physical parametrizations which are needed to

model the effects discussed in section 3($f$). Le Dimet and Talagrand (1986), and Lewis and Derber (1985) described an iterative method of minimizing (51) which allowed $K_n$ to be nonlinear, and which by ignoring the term in $\mathbf{B}$ avoids large matrix manipulations. The basic advantage of their method is that the forecast model $M_{ni}$ and its derivatives $\mathbf{M}_i$ must have been factorized into simple very sparse steps in order for $M_n$ to be a practicable forecast model. The same factorization can be used to implement multiplication by $\mathbf{M}_i^{\mathrm{T}}$; this is the adjoint model. Le Dimet and Talagrand's method ignored completely the background term $(\mathbf{x} - \mathbf{x}_b)^{\mathrm{T}}\mathbf{B}^{-1}(\mathbf{x} - \mathbf{x}_b)$ in (51), a valid approximation only if the assimilation period is sufficiently long to contain enough observations to determine $\mathbf{x}$ without the prior information implicit in $\mathbf{x}_b$ and $\mathbf{B}$. Here we derive a more general iteration. Starting from (51), and linearizing about the current best estimate of its minimizer $\mathbf{x}[u]$, we get a new estimate

$$\mathbf{x}[u+1] = \mathbf{x}[u] + (\sum_{i=1}^{n} \mathbf{K}_i^{\mathrm{T}}(\mathbf{O}_i + \mathbf{F}_i)^{-1}\mathbf{K}_i + \mathbf{B}^{-1})^{-1} \times$$
$$(\sum_{i=1}^{n} \mathbf{K}_i^{\mathrm{T}}(\mathbf{O}_i + \mathbf{F}_i)^{-1}\{\mathbf{y}_{oi} - K_{ni}(\mathbf{x}[u])\} + \mathbf{B}^{-1}\{\mathbf{x}_b - \mathbf{x}[u]\}). \quad (52)$$

Le Dimet and Talagrand proposed integrating the prediction model forward in time from $\mathbf{x}[u]$, remembering the model fields at times $i = 1$ to $n$, then integrating the adjoint model backward in time. The remembered model fields are used to calculate the $\mathbf{M}_i^{\mathrm{T}}$, and the vector $\Sigma\, \mathbf{K}_i^{\mathrm{T}}(\mathbf{O}_i + \mathbf{F}_i)^{-1}\{\mathbf{y}_{oi} - K_{ni}(\mathbf{x}[u])\}$ is accumulated by taking out the common factors $\mathbf{M}_0^{\mathrm{T}} \ldots \mathbf{M}_{i-1}^{\mathrm{T}}$ from $\mathbf{K}_i^{\mathrm{T}}$ using (50). Assuming $\mathbf{B}^{-1}$ to be negligibly small, this gives the direction of descent in $N_x$-dimensional space towards the minimum of (51), but not the distance, which they had to obtain from some descent algorithm. The full equation (52) gives also the distance, but only at the cost of manipulating very large matrices in the first summation term.

If $\mathbf{B}^{-1}$ is not negligible, (53) is a more appropriate form:

$$\mathbf{x}[u+1] = \mathbf{x}[u] + \{\mathbf{B}\sum_{i=1}^{n} \mathbf{K}_i^{\mathrm{T}}(\mathbf{O}_i + \mathbf{F}_i)^{-1}\mathbf{K}_i + \mathbf{I}\}^{-1} \times$$
$$(\mathbf{B}\sum_{i=1}^{n} \mathbf{K}_i^{\mathrm{T}}(\mathbf{O}_i + \mathbf{F}_i)^{-1}\{\mathbf{y}_{oi} - K_{ni}(\mathbf{x}[u])\} + \mathbf{x}_b - \mathbf{x}[u]). \quad (53)$$

We saw in section 4($g$) how multiplication by $\mathbf{B}$ could be approximated by projection onto normal modes, multiplication by a diagonal matrix of weights for each mode, and projection back to the model basis. These operations are practically achievable. It remains to estimate $\{\mathbf{B}\,\Sigma\, \mathbf{K}_i^{\mathrm{T}}(\mathbf{O}_i + \mathbf{F}_i)^{-1}\mathbf{K}_i + \mathbf{I}\}^{-1}$; calculation of this is impracticable, but guidance may be obtained by comparing (53) with (38). Equation (53) is just a generalization of the successive correction method, and the term we have to estimate is the normalization factor $\mathbf{Q}$. Unfortunately the adjoint model integration which is used to calculate the other summation in (53) avoids explicit calculation of $\mathbf{W}$, so an approximation like (42) is difficult to find. Moreover, just as for the successive correction method, an approximation for $\mathbf{Q}$ will not retain any constraints on the increments imposed by the multiplication by $\mathbf{B}$. The modified successive correction method (43) suggests one method for avoiding this latter problem, by using

$$\mathbf{x}[u+1] = \mathbf{x}[u] + \mathbf{B}\sum_{i=1}^{n} \mathbf{K}_i^{\mathrm{T}}(\mathbf{O}_i + \mathbf{F}_i)^{-1}\mathbf{Q}'_i\{\mathbf{y}_{oi} - K_{ni}(\mathbf{x}[u])\}. \quad (54)$$

This does not allow for observation errors; the method used for (45) and (46) might do this, but cannot be shown to give the same convergence limit as (52) for $K_n$ nonlinear.

## 5. Discussion and conclusions

In section 2 we derived idealized analysis equations for NWP using Bayesian probabilistic arguments, and showed how a maximum likelihood best estimate could be found as the solution of a variational problem. For the practical implementation of these equations, the physical characteristics of the NWP analysis problem need to be considered. These were discussed in section 3 and are as follows:

a. Considerations of forecast model accuracy and economy predetermine the model basis, rather than data resolution.

b. Approximations are necessary because of the very large order of the model and number of data.

c. Prior information or assumptions are essential; the analysis is underdetermined using observations alone.

d. A time history of observations, plus knowledge of predictive equations constraining possible four-dimensional solutions, can help resolve the underdeterminacy; this is the basis of four-dimensional data assimilation.

e. The empirical knowledge that atmospheric fields are usually slowly varying, implying a balance between the forces acting on them, can be used to reduce the underdeterminacy. Techniques to achieve this are related to initialization.

f. Relationships exist between different atmospheric fields and different scales of motion, which mean that some of these are largely determined by others. In contrast to the simple linearizable relationships such as geostrophy, many potentially useful relationships require complex nonlinear models to represent them. It is difficult to incorporate these into a linear analysis method.

Most of the methods which have been proposed for analysis for NWP can be related to the idealized method of section 2, and hence also to each other. These include variational methods, splines, Kriging, optimal interpolation, successive corrections, constrained initialization, Kalman–Bucy filters, and data assimilation using adjoint models. In section 4 these methods were discussed in the light of the physical characteristics of the NWP analysis problem. None of these methods is ideal in all respects; which is preferred depends on which of the characteristics are considered more important. Some possible cases follow.

(i) If the observations are most important, with little weight given to the prior relationships, then methods such as the traditional successive correction method, smoothing splines, Kriging, and OI with a limited data selection are appropriate. This might be the case if there are plentiful data, for instance from satellites, and relatively less accurate forecast models, as in nowcasting and very-short-period forecasting. These methods are also appropriate for analyses for other purposes than NWP, e.g. diagnostic studies.

(ii) If an accurate NWP forecast model is available, then this and other prior information such as balance must be effectively and efficiently used. Methods linearized about this accurate background are appropriate. There are relatively fewer effective data, so balance relationships should be used, either through the background error covariance **B** (and in OI a large selection of data), or explicitly by setting constraints on the increments. Specification of **B** is difficult, requiring either extensive data studies, or theoretical approaches such as the Kalman–Bucy filter. A practical choice of method here needs to take account of the cost-effective use of available computer power.

(iii) As NWP models get more complex, with high resolution and sophisticated representations of convection, cloud, rain, radiation, and the planetary boundary layer, then their degrees of freedom increase and the nonlinear relationships discussed in section 3($f$) become more important. At present our best method of expressing such relationships is in terms of the NWP model, which, if it is well designed, tends to satisfy them. Thus a data assimilation scheme designed around the complete NWP model is needed. For such NWP models we now have quite good representations of the generalized interpolation $K_n$ for observations like cloud amount, rainfall rate, outward long-wave radiation, convective activity, and surface pressure tendency. However, these are not easily linearized in order to do a linear explicit analysis. Iterative techniques such as that of Eq. (53) give the hope that such observations can nevertheless be effectively used.

Whichever practical analysis technique is chosen, comparison with the idealized equations of section 2 can give insight into the suitable choice of parameters in the approximate method. The following points were made as each method was discussed in section 4:

(1) The penalty function used in variational analysis is related to the PDFs of background and observations and their error covariances. Any linearizable constraints can be combined with other background fields and covariances and need not be explicitly included as separate terms. Accurate prior relationships can be applied as strong constraints by a reduction of the control variable.

(2) Smoothing spline approximation is a special case of the idealized method; the penalty on smoothness is related to the assumed spectral error characteristics of the background (or for a zero background, of the field).

(3) Observational 'errors of representativeness' as used in OI are shown to be the errors in the generalized interpolation $K_n$. Only scales resolved by the model basis should be included in the background error covariance function of OI.

(4) The number of data selected in OI governs the extent to which prior information implicit in the background and its error covariance is used.

(5) The successive correction method can be modified, by appropriate specification of its weighting function and normalization factor, to converge towards a solution of the ideal analysis equation.

(6) Observation rejection criteria in quality control procedures can be related to the non-Gaussian form of their error distributions.

(7) Nonlinear balance constraints can be treated explicitly if they can be linearized above some model state. To achieve this it is necessary that the background used is in nonlinear balance and that the covariance matrix and analysis increments satisfy the linear relationship.

(8) Four-dimensional data assimilation is covered by the same equations if the interpolation $K_n$ is generalized to include time extrapolation. A descent algorithm for a nonlinear variational analysis using the adjoint model method can be suggested by analogy with the successive correction method.

## REFERENCES


Barnes, S. L. 1973 'Mesoscale objective analysis using weighted time-series observations'. NOAA Tech. Memorandum ERL NSSL-62

Bengtsson, L. 1975 '4-dimensional assimilation of meteorological observations'. ICSU/WMO GARP Pub. Series No. 15

Bengtsson, L, Ghil, M. and Kallen, E. (E ;) 1981 *Dynamic meteorology: Data assimilation methods.* Springer-Verlag, New York

Bergthorsson, P. and Doos, B. R. 1955 Numerical weather map analysis. *Tellus*, **7**, 329–340

Bratseth, A. M. 1986 Statistical interpolation by means of successive corrections. *ibid.*, **38A**, (to appear)

Charney, J., Halem, M. and Jastrow, R. 1969 Use of incomplete historical data to infer the present state of the atmosphere. *J. Atmos. Sci.*, **26**, 1160–1163

Cressman, G. P. 1959 An operational objective analysis scheme. *Mon. Wea. Rev.*, **87**, 367–374

Errico, R. and Baumhefner, D. 1986 Predictability experiments using a high-resolution limited-area model. *ibid.*, **114**, (to appear)

Franke, R. and Gordon, W. J. 1983 'The structure of optimum interpolation functions'. Naval Postgraduate School, Monterey, Cal. Report NPS-53-83-0005

Gandin, L. S. 1963 'Objective Analysis of Meteorological Fields'. Gidrometeorologicheskoe Izdatel'stro, Leningrad. Translated from Russian, Israeli Program for Scientific Translations, Jerusalem, 1965

Ghil, M., Cohn, S. E., Tavantzis, J., Bube, K. and Isaacson, E. 1981 'Applications of estimation theory to numerical weather prediction'. Pp. 139–284 in *Dynamic meteorology: Data assimilation methods*. Eds. Bengtsson L, Ghil, M. and Kallen, E. Springer-Verlag, New York

Hollingsworth, A. and Lonnberg, P. 1986 The statistical structure of short-range forecast errors as determined from radiosonde data. Part I: The wind field. Part II: the covariance of height and wind error. *Tellus*, **38A**, 111–136 and 137–161

Ikawa, M. 1984a An alternative method of solving weak constraint problem and a unified expression of weak and strong constraints in variational objective analysis. *Papers in Met. and Geophys.*, **35**, 71–79

1984b Generalization of multivariate optimum interpolation method and the roles of linear constraint and covariance matrix in the method. Part I: in discrete form. Part II: in continuous form. *Papers in Met. and Geophys.*, **35**, 81–102 and 169–180

Journel, A. G. 1977 Kriging in terms of projections. *Math. Geology*, **9**, 563–586

Le Dimet, F-X. and Talagrand, O. 1986 Variational algorithms for analysis and assimilation of meteorological observations. *Tellus*, **38A**, 97–110

Lewis, J. M. and Derber, J. C. 1985 The use of adjoint equations to solve a variational adjustment problem with advective constraints. *Tellus*, **37A**, 309–322

Lorenc, A. C. 1981 A global three-dimensional multivariate statistical analysis scheme. *Mon. Wea. Rev.*, **109**, 701–721

1984 'Analysis methods for the quality control of observations', in proceedings of ECMWF workshop on The use and quality control of meteorological observations for numerical weather prediction, 6–9 Nov. 1984

Lorenc, A. and Tibaldi, S. 1980 'The treatment of humidity in ECMWF's data assimilation scheme'. Pp. 497–512 in *Atmospheric Water Vapor*, Ed. Deepak A., Wilkinson, T. and Ruhnke L. H. Academic Press, New York

Lorenc, A. C., Adams, W. and Eyre, J. 1985 'The analysis of high-resolution satellite data in the Meteorological Office', in Proceedings of ECMWF workshop on high-resolution analysis, 24–26 June 1985

Marshall, John C. 1985 Determining the ocean circulation and improving the geoid from satellite altimetry. *J. Phys. Oceanogr.*, **15**, 330–349

Matheron, G. 1963 *Traité de géostatistique appliquée.* Editions Technip., Paris

Menke, W. 1984 *Geophysical data analysis: Discrete inverse theory.* Academic Press

O'Sullivan, F. and Wahba, G. 1985 A cross-validated Bayesian retrieval algorithm for non-linear remote sensing experiments. *J. Comput. Phys.*, **59**, 441–455

Parrish, D. E. and Cohn, S. E. 1985 'A Kalman filter for a two-dimensional shallow-water model'. Proc. Seventh Conf. Numerical Weather Prediction, June 17–20, 1985, Montreal, Canada. American Meteorological Society

Phillips, N. A. 1982 On the completeness of multi-variate optimum interpolation for large-scale meteorological analysis. *Mon. Wea. Rev.*, **110,** 1329–1334

1986 The spatial statistics of random geostrophic modes and first-guess errors. *Tellus*, **38A,** 314–332

Purser, R. J. 1984 'A new approach to the optimal assimilation of meteorological data by iterative Bayesian analysis'. Pp. 102–105 in preprints, 10th conference on weather forecasting and analysis. American Meteorological Society

Reinsch, C. H. 1967 Smoothing by spline functions. *Numer. Math.*, **10,** 177–183

Rodgers, C. D. 1976 Retrieval of atmospheric temperature and composition from remote measurements of thermal radiation. *Rev. Geophys. Space. Phys.*, **14,** 609–624

Sasaki, Y. 1970 Some basic formalisms on numerical variational analysis. *Mon. Wea. Rev.*, **98,** 875–883

Schlatter, T. W. 1981 An assessment of operational TIROS-N temperature retrievals over the United States. *ibid.*, **109,** 110–119

Seaman, R. S. and Hutchinson, M. F. 1985 Comparative real data tests of some objective analysis methods by withholding observations. *Aust. Met. Mag.*, **33,** 37–46

Wahba, G. 1978 Improper priors, spline smoothing and the problem of guarding against model errors in regression. *J. R. Statist. Soc.*, **B, 40,** 364–372

1982 'Variational methods in simultaneous optimum interpolation and initialization'. Pp. 178–185 in *The interaction between objective analysis and initialization: Proc. 14th Stansted seminar*. Ed. Williamson, D. (Publications in Meteorology, **127**), McGill University, Montreal

1985 'Variational methods for multidimensional inverse problems'. Pp. 385–410 in *Advances in remote sensing retrieval methods*. Eds. H. E. Fleming and M. T. Chanine. A. Deepak.

Wahba, G. and Wendelberger, J. 1980 Some new mathematical methods for variational objective analysis using splines and cross validation. *Mon. Wea. Rev.*, **108**, 1122–1143

Williamson, D. (Ed.) 1982 *The interaction between objective analysis and initialization. Proc. 14th Stanstead seminar* (Publications in Meteorology, **127**), McGill University, Montreal