A Practical Approximation to Optimal Four-Dimensional Objective Analysis

ANDREW C. LORENC*

Meteorological Office, Bracknell, England

(Manuscript received 7 May 1987, in final form 1 October 1987)

ABSTRACT

An iterative four-dimensional objective analysis scheme is described. The method is derived by approximating a variational algorithm which should give an optimal four-dimensional analysis. The complete set of operationally available observations, and operational analysis and forecast codes, are used. In this the scheme differs from most other studies of optimal four-dimensional analysis, which make fewer approximations in the algorithm, but use simplified models and data.

The scheme was developed using the optimal interpolation analysis, nonlinear normal-mode initialization, and nested-grid forecast model from the Regional Analysis and Forecast System of NMC. To these were added an approximate adjoint model of the forecast, and a code to implement a simple descent algorithm. Tests used the operational observation database.

The scheme was successful in producing a dynamically consistent four-dimensional analysis that fit the observations without totally impractical computer costs. However for the one test case studied, the forecast from the scheme's analysis was slightly worse than that from the operational analysis.

The tests highlighted some deficiencies of the current operational analysis, initialization, and forecast codes. They also indicated areas where further development of the scheme is desirable; in the adjoint forecast model and analysis error estimation.

1. Introduction

The objective of this work is to try to implement, in a practical environment, some of the four-dimensional analysis theory, which has been developed theoretically and tested in simple models. If errors introduced by a forecast over the period spanning the observations are neglected, then the "optimal" analysis is defined by the three-dimensional, balanced, state which simultaneously best fits a background state, and the observations over a period of time, when forecast using the standard model. The full solution to this variational problem has been derived and tested with a very simple model in Lorenc (1988).

The aim here is to construct an approximation to the theoretically correct method, using if possible existing programs for three-dimensional analysis (3DOI), forecasting, etc. By using the full observational dataset available for operational forecasting, and the operational forecast model, we hope to demonstrate that the scheme is capable of practical implementation. Compared to the "three-and-a-half" dimensional analysis given by currently operational data-assimilation schemes, a four-dimensional scheme should more correctly use tendency information in the observations, and be more readily adaptable to the use of asynoptic data. We also hope to indicate a way in which the current methods can be modified to become more nearly four-dimensional, without a disruptive sudden change to a conceptually different scheme. By using available programs as components in the new scheme, we will be able to switch to up-to-date versions as they become available, and to transfer the method easily to other models, allowing this research to proceed in parallel with other developments. We also minimize the amount of new code required. In particular all the handling, sorting, selection, and spatial interpolation of observations, which take the bulk of the effort in coding a practical analysis scheme, are kept unchanged in the 3DOI component.

An ideal four-dimensional scheme should be able to use the tendency information in observations, for instance observations indicating that a low is deepening should generate upper flows which would cause the model forecast fields to similarly deepen. As well as this, the method has potential in alleviating the practical "spinup" problem of current operational forecasts. If successful, the scheme will well fit the observations at the end of the 4D analysis period with a forecast from the beginning. This forecast will be consistent with the model's dynamics and physical parameterizations and can be extended into the future, avoiding

* This work was performed while the author was a visiting scientist at the National Meteorological Center, Washington D.C.

Corresponding author address: Mr. Andrew C. Lorenc, Met 0 11 (Forecasting Research), Meteorological Office, London Road, Bracknell RG12 2SZ, England.

some of the spinup problems, in parameters like convective rainfall, which occur when initializing a forecast from a three-dimensional analysis.

The scheme has not been developed for operational implementation in the near future; current computer constraints rule that out. One practical use might be to produce dynamically consistent four-dimensional analyses of research datasets, for diagnostic study.

A review of analysis methods for numerical weather prediction, which discusses the relationship between so-called optimal interpolation (OI), constrained variational minimization, the Kalman–Bucy filter, and the adjoint equation method, has been published by Lorenc (1986). We shall not repeat this review here, however it is appropriate to list research which actually attempted to perform a four-dimensional analysis constrained by nonlinear prediction equations. For a method to be truly four-dimensional, the use made of any observation must depend on whether there are similar observations at other times, defining tendencies, and on what these tendencies are. Lewis and Bloom (1978) used a variational technique on gridded fields. Ghil et al. (1981) used the Kalman–Bucy filter method for a simple one-dimensional example. Lewis and Derber (1985) and Courtier and Talagrand (1987) used the adjoint method. Hoffmann (1986) solved the minimization problem in a straightforward (computationally expensive) way for a very simple model. None of these examples used an observational database which approached in magnitude the operational database used for this work.

The method used here can be considered a combination of OI with the adjoint equation method. Equations for this were derived and tested in a simple one-dimensional model by Lorenc (1988). This work uses the same notation and basic equations, set out in section 2. Section 3 shows how, with some approximations, the 3DOI program which is used operationally can be modified to calculate some terms in the equations. Approximations are needed especially in estimating the analysis error covariances. Since we do not have available the adjoint of the operational forecast model, it must be approximated. Section 4 describes how this is done by integrating backwards an adiabatic perturbation form of the model. Many of the aforementioned approximations are only valid for "balanced" fields; the concept of balance also provides useful information for constraining the analysis. Section 5 discusses this aspect. Section 6 sets out the iterative procedure which results from all the preceding discussion. This procedure might indeed have been arrived at as a "common sense" reasonable thing to do; readers not interested in the mathematical justification may skip the intervening detailed equations. Section 7 describes the test scheme implemented using the operational analysis and forecast programs, gives results from test analyses and forecasts for a single case, and discusses how the approximations made have affected results. Finally in section 8 we give our conclusions.

2. Basic equations

a. Notation

- $\mathbf{x}$ — 4D analysis, represented as a single vector.
- $\mathbf{y}$ — observations, distributed in 4D, represented as a single vector.
- t — subscript indicating "true", hence:
- $\mathbf{x}_t$ — "true" value of $\mathbf{x}$, obtained by projecting the true atmospheric state onto our finite basis for $\mathbf{x}$.
- $\mathbf{y}_t$ — "true" value of $\mathbf{y}$, which would be obtained from hypothetical error free instruments, with the same resolution and averaging characteristics as the actual instruments.
- $\mathbf{y}_o$ — observed data values.
- n — subscript to function, indicating that it can be nonlinear.
- K_n — generalized interpolation from $\mathbf{x}$-representation to $\mathbf{y}$-representation, such that if we have an estimate $\mathbf{x}_i$ of $\mathbf{x}_t$, then $y_i = K_n(\mathbf{x}_i)$ is the best estimate of $\mathbf{y}_t$.
- $\mathbf{K}$ — Matrix of partial derivatives of K_n with respect to the elements of $\mathbf{x}$.
- $\mathbf{w}$ — 3-D analysis at the initial time covered by the 4D representation $\mathbf{x}$, with the same space-representation.
- G_n — forecast model, used as prognostic constraint on permitted values of $\mathbf{x}$, by the relationship $\mathbf{x} = G_n(\mathbf{w})$.
- $\mathbf{G}$ — Matrix of partial derivatives of G_n with respect to the elements of $\mathbf{w}$.
- $\mathbf{w}_t$ — "true" $\mathbf{w}$, as $\mathbf{x}_t$. Since we are assuming our forecast model to be perfect, in order to justify its use as a strong constraint, $\mathbf{x}_t = G_n(\mathbf{w}_t)$.
- $\mathbf{w}_b$ — background field for $\mathbf{w}$; the best available estimate of $\mathbf{w}_t$ given our prior knowledge, without using $\mathbf{y}_o$.
- $*$ — adjoint. Hermitian transpose.
- $\mathbf{B}$ — background error covariance matrix. $\mathbf{B} = \langle(\mathbf{w}_b - \mathbf{w}_t)(\mathbf{w}_b - \mathbf{w}_t)^*\rangle$
- $\mathbf{O}$ — observation error covariance matrix. $\mathbf{O} = \langle(\mathbf{y}_o - \mathbf{y}_t)(\mathbf{y}_o - \mathbf{y}_t)^*\rangle$
- $\mathbf{F}$ — representativeness error covariance matrix. $\mathbf{F} = \langle(K_n(\mathbf{x}_t) - \mathbf{y}_t)(K_n(\mathbf{x}_t) - \mathbf{y}_t)^*\rangle$
- $\mathbf{v}$ — transformed control variable. $\mathbf{v} = \mathbf{B}^{-1}(\mathbf{w} - \mathbf{w}_b)$
- i — iteration index; e.g. $\mathbf{w}_i$ is the best estimate of $\mathbf{w}_t$ at the ith iteration.
- $\mathbf{dy}_i$ — difference of current best estimate from observed values. $\mathbf{dy}_i = \mathbf{y}_i - \mathbf{y}_o = K_n(\mathbf{x}_i) - \mathbf{y}_o = K_n(G_n(\mathbf{w}_i)) - \mathbf{y}_o = K_n(G_n(\mathbf{w}_b + \mathbf{B}\mathbf{v}_i)) - \mathbf{y}_o$
- $J(\mathbf{w})$ — penalty function, whose minimum defines the "best" analysis.
- $L(\mathbf{v})$ — penalty function equivalent to $J(\mathbf{w})$, expressed in terms of the transformed variable $\mathbf{v}$.

$_f$ subscript for penalties J and L denoting the component of the penalty measuring the fit to observations. (NB in Lorenc, 1988, the double subscript $_{of}$ was used for this, to indicate its origin from the convolution of instrumental and representativeness error distributions.)

$_b$ subscript which when applied to **xw** or **v** denotes the background prior estimate of the best analysis, and when applied to penalties J or L denotes the component of the penalty measuring the fit to the background.

$'$ differentiation of function by argument.

$_m$ subscript distinguishing time-slice. (It is convenient for manipulation of the 4D representations **x** and **y**, to partition them into a finite number (N) of time-slices denoted by subscript m).

S smoothing operator for the diagonal matrix of normalized analysis errors.

T_m nominal validity time for observations in time-slice m.

t, o, n, b, f are mutually exclusive subscripts, and always precede i (iteration) and m (time-slice), which are added in that order.

b. Penalty function and derivatives

Lorenc (1988) showed that, if Gaussian error statistics are assumed, and errors in the forecast model during the time spanned by the current observations are ignored, the optimal analysis can be obtained by minimizing a penalty function J with respect to the three-dimensional field (**w**) at the beginning of the current period. The penalty function measures the deviation **dy** from the observations of a forecast from **w**, plus the deviation of **w** from the background information. (In section 5 below we discuss an additional constraint that **w** be balanced.)

$$\mathbf{dy} = K_n(G_n(\mathbf{w})) - \mathbf{y}_o \tag{1}$$

$$J(\mathbf{w}) = \mathbf{dy}^*(\mathsf{O} + \mathsf{F})^{-1}\mathbf{dy}/2 + (\mathbf{w} - \mathbf{w}_b)^*\mathsf{B}^{-1}(\mathbf{w} - \mathbf{w}_b)/2. \tag{2}$$

We use an iterative descent algorithm to search for this minimum. Subscript i is used to indicate values at a particular iteration. As explained in Lorenc (1988), if we are to use only a few iterations of a simple descent algorithm, and not deviate too far from the background, it is better to use a descent algorithm derived in terms of a transformed variable **v**, rather than the basic variable **w**. This makes differences between the analysis and the background smoother, rather than generating sharp spikes to fit closely the observations. In the appendix we set out the equations defining the terms which will be used in such a descent algorithm. Note that, despite the use of **v** conceptually to derive the equations, the fields need never be represented in terms of **v**. Indeed this would be impracticable since we do not have a convenient representation for the background error covariance matrix **B** or its inverse, which enters in the transformation from **w** to **v**:

$$\mathbf{v}_i = \mathsf{B}^{-1}(\mathbf{w}_i - \mathbf{w}_b). \tag{3}$$

We call the penalty function in terms of this transformed variable L. Expressions for L and its derivatives are given in the appendix.

c. Basic descent algorithm

For the approximately quadratic penalty functions that we assume above, if all the terms above are known, the best algorithm for finding the minimum of the penalty function is that of Newton: If $\mathbf{v}_i$ is an approximate minimum of L, then a better approximation is given by

$$\mathbf{v}_{i+1} = \mathbf{v}_i - \{L''(\mathbf{v}_i)\}^{-1}L'(\mathbf{v}_i). \tag{4}$$

For our basic variable **w** this gives

$$\mathbf{w}_{i+1} = \mathbf{w}_i - \mathsf{B}\{L''(\mathbf{v}_i)\}^{-1}L'(\mathbf{v}_i). \tag{5}$$

If this is accurately evaluated, the transformation from **w** to **v** has no effect. We shall however introduce some approximations below, which make $\mathsf{B}\{L''(\mathbf{v}_i)\}^{-1}$ diagonal. This can only be justified for the transformed equations, and gives the iteration some of the properties of a preconditioned steepest-descent algorithm, so the transformation of variables is important.

3. Use of the three-dimensional analysis program 3DOI

a. Use of increments, error, and fit of observations to guess

The four-dimensional analysis problem is difficult to handle in practice for operational resolutions. However there is considerable experience with a reasonable approximation to the equivalent three-dimensional problem, the so called optimal interpolation method (OI). At time T_m OI gives an approximate minimum for:

$$J_m(\mathbf{x}_m) = \mathbf{dy}^*_{im}(\mathsf{O}_m + \mathsf{F}_m)^{-1}\mathbf{dy}_{im} + (\mathbf{x}_m - \mathbf{x}_{bm})^*\mathsf{B}^{-1}(\mathbf{x}_m - \mathbf{x}_{bm}). \tag{6}$$

Here K_{nm} is assumed linear, so the minimizing field $\mathbf{x}_{am}$ is given explicitly by

$$\mathbf{x}_{am} - \mathbf{x}_{bm} = -J''_m(\mathbf{x}_{bm})^{-1}J'_m(\mathbf{x}_{bm}) \tag{7}$$

$$\mathbf{x}_{am} - \mathbf{x}_{bm} = (\mathsf{K}^*_m(\mathsf{O}_m + \mathsf{F}_m)^{-1}\mathsf{K}_m + \mathsf{B}^{-1})^{-1} \times \mathsf{K}^*_m(\mathsf{O}_m + \mathsf{F}_m)^{-1}(\mathbf{y}_{om} - K_{nm}(\mathbf{x}_{bm})). \tag{8}$$

We call the programs for doing this the 3DOI. They actually use the equivalent form (Lorenc 1986):

$$\mathbf{x}_{am} - \mathbf{x}_{bm} = \mathbf{BK}_m^*(\mathbf{O}_m + \mathbf{F}_m + \mathbf{K}_m^*\mathbf{BK}_m)^{-1}(\mathbf{y}_{om} - K_{nm}(\mathbf{x}_{bm})). \quad (9)$$

They approximate this large matrix inverse problem by many smaller ones, each for a small selection of the data, and each providing only a few elements of the analysis increment vector $\mathbf{x}_{am} - \mathbf{x}_{bm}$.

3DOI programs also usually calculate the estimated analysis error variance; the diagonal of $\mathbf{A}_m$:

$$\mathbf{A}_m = \langle(\mathbf{x}_{am} - \mathbf{x}_{tm})(\mathbf{x}_{am} - \mathbf{x}_{tm})^*\rangle = (\mathbf{K}_m^*(\mathbf{O}_m + \mathbf{F}_m)^{-1}\mathbf{K}_m + \mathbf{B}^{-1})^{-1}. \quad (10)$$

Or else they calculate the normalized error variance; the diagonal of $\mathbf{A}_m$ divided by the diagonal of $\mathbf{B}$.

Third, since they calculate all the observation increments $\mathbf{dy}_m$ [defined as $\mathbf{y}_{om} - K_{nm}(\mathbf{x}_{bm})$], they can readily calculate the observation penalty J_{fm} for the background field $\mathbf{x}_{bm}$.

If, instead of the background field $\mathbf{x}_{bm}$, we feed such a 3DOI program with the current best estimate $\mathbf{x}_{im}$ from an iteration of a four-dimensional analysis, then it will give us estimates of the observation penalty for $\mathbf{x}_{im}$, and expressions involving its first and second derivatives. Using $\mathbf{dx}_m$ to denote the analysis increment from the 3DOI program, we get:

$$J_{fm}(\mathbf{x}_{im}) = \mathbf{dy}_{im}^*(\mathbf{O}_m + \mathbf{F}_m)^{-1}\mathbf{dy}_{im} \quad (11)$$

$$\mathbf{dx}_m = -(\mathbf{K}_m^*(\mathbf{O}_m + \mathbf{F}_m)^{-1}\mathbf{K}_m + \mathbf{B}^{-1})^{-1} \times \mathbf{K}_m^*(\mathbf{O}_m + \mathbf{F}_m)^{-1}\mathbf{dy}_{im} \quad (12)$$

and the (normalized) diagonal of $\mathbf{A}_m$.

b. Approximations to covariances

Our intention is to use a 3DOI program to calculate the values just described, as a component of an iterative procedure for finding an approximate minimum to $L(\mathbf{v})$, our four-dimensional penalty function. We need first to justify some further approximations in our handling of $\mathbf{BA}_m$ and $\mathbf{G}_m$, since we cannot store and manipulate these matrices for a full resolution NWP model. Let us first nondimensionalize $\mathbf{x}$ by a diagonal normalization matrix $\mathbf{Z}$, such that $(\mathbf{Zx})^*\mathbf{Zx}$ is a measure of energy. The normalized background error covariance is then given by:

$$\langle\mathbf{Z}(\mathbf{x}_b - \mathbf{x}_t)(\mathbf{x}_b - \mathbf{x}_t)^*\mathbf{Z}^*\rangle = \mathbf{ZBZ}^* = \mathbf{EbE}^*$$

where $\mathbf{E}$ is a matrix whose columns are the normalized eigenmodes of the background error covariance, and $\mathbf{b}$ is a diagonal matrix of error energies for each mode; $\mathbf{E}$ is self-inverse. If both the truth and the background are balanced, then a linearization of the balance relationship leads to some of the elements of $\mathbf{b}$ being near zero. The same, unbalanced, modes should also then be near zero in results from the 3DOI program, both in $\mathbf{A}_m$ and $\mathbf{dx}_m$. We neglect these modes, and concentrate on the balanced modes. Applying the normalization to the linearized forecast model $\mathbf{G}_m$, and expressing it as a sequence of time-steps, gives

$$\mathbf{G}_m = \mathbf{Z}^{-1}\mathbf{M}_m\mathbf{M}_{m-1}\cdots\mathbf{M}_2\mathbf{M}_1\mathbf{Z}. \quad (13)$$

It seems plausible to assume that the error structure of the forecast background can be described as a random distribution of energy among modes of the model. Phillips (1986) suggested equipartition of energy among Rossby modes, with random phases. If this is so, then the modes in $\mathbf{E}$ must be expressible as simple sums and differences of the modes of $\mathbf{M}$, in complex conjugate pairs. If the partition of energy in $\mathbf{b}$ is equal within each pair, then $\mathbf{M}$ and $\mathbf{b}$ will commute:

$$\mathbf{bM} = \mathbf{Mb}. \quad (14)$$

Hence $\mathbf{B}$ and $\mathbf{G}_m$ also commute:

$$\mathbf{BG}_m = \mathbf{G}_m\mathbf{B}. \quad (15)$$

These assumptions are only valid for spatially homogeneous error distributions. (Otherwise the random phase assumption above is not correct.) If $\mathbf{B}$ has some spatial variability, for instance from variations in observation density at earlier times, then we are neglecting the advection of this structure in the errors during the forecast.

We need also to make some assumptions about $\mathbf{A}_m$, since the OI program only provides information about its diagonal. The correlation structure of the analysis errors is a function of observation distribution. Where there are no observations, analysis errors are identical to background errors. Where there are observations, the spatial correlation of analysis errors will tend to drop to zero at about the observation separation distance. We make the convenient, but rather gross, assumption that the structure of $\mathbf{A}_m$ is similar to that of $\mathbf{B}$, so that

$$\mathbf{BA}_m^{-1} = \mathbf{a}_m^{-1} \quad (16)$$

where $\mathbf{a}_m$ is a diagonal matrix made from the normalized error variances calculated by the OI program. The validity of this assumption depends on the representation chosen for $\mathbf{x}$. We are not at liberty to allow any arbitrary values for the elements of $\mathbf{a}$. For instance $\mathbf{B}$ is often modelled (see section 7 below) using the geostrophic assumption and a fixed horizontal correlation structure. This determines the local relationship between height and wind error variances. Since we are assuming that the analysis is similarly balanced, and has similar error correlations, we must ensure that the analysis error variances implied by $\mathbf{a}$ obey the same relationship. Thus locally, the elements of $\mathbf{a}$ for height and wind must be approximately equal. The 3DOI estimates of normalized analysis error variance have no such constraint; they allow large differences between the implied correlation structures of the analysis errors and the background errors. Thus before use in this approximate scheme as the elements of $\mathbf{a}$, these normalized analysis errors should be locally averaged over height and wind, and spatially smoothed.

c. Use of 3DOI results in four-dimensional scheme

Using the above approximations, we can now use the results from the three-dimensional analysis program in the four-dimensional scheme. Note that we use the current best estimate $\mathbf{x}_{im}$ as "guess" for the 3DOI, rather than the background field $\mathbf{x}_{bm}$, so the actual three-dimensional analysis produced will not be an optimal combination of background and observations. For each time-slice m the 3DOI program gives us the contribution to the observation penalty J_{fm}, the analysis increments $\mathbf{dx}_{im}$ and the normalized analysis error variance $\mathbf{a}_m$. Using these we get

$$\begin{aligned} L_{fm}(\mathbf{v}_i) &= J_{fm}(\mathbf{x}_{im}) \\ &= \mathbf{dy}^*_{fm}(\mathbf{O}_m + \mathbf{F}_m)^{-1}\mathbf{dy}_{im}. \end{aligned} \tag{17}$$

The first derivative of this is

$$L'_{fm}(\mathbf{v}_i) = \mathbf{B}\mathbf{G}^*_m\mathbf{K}^*_m(\mathbf{O}_m + \mathbf{F}_m)^{-1}\mathbf{dy}_{im} \tag{18}$$

which with our approximations is given by

$$L'_{fm}(\mathbf{v}_i) = -\mathbf{G}^*_m\mathbf{a}_m^{-1}\mathbf{dx}_{im}. \tag{19}$$

Similar approximations give an expression for the second derivative:

$$L''_{fm}(\mathbf{v}_i) = \mathbf{G}^*_m(\mathbf{a}_m^{-1} - \mathbf{I})\mathbf{G}_m\mathbf{B}. \tag{20}$$

This is still impracticable for computation for large operational models; we have to make a further approximation as to the effect of operating on the normalized error matrix by the linearized forecast model $\mathbf{G}$. We will make the gross approximation that this can be modelled as a simple spatial smoothing similar to that used in obtaining an estimate of $\mathbf{a}$ from the normalized analysis error variances. We denote this by $\mathbf{S}_m$. So finally we get

$$L''_{fm}(\mathbf{v}_i) = \mathbf{S}_m(\mathbf{a}_m^{-1} - \mathbf{I})\mathbf{B}. \tag{21}$$

These expressions can be substituted into a Newton iteration to find the minimum of $L(\mathbf{v})$, our four-dimensional analysis problem. (Strictly, because we have ignored second derivatives of G_n in our expression for L''_{fm}, we are using a Gauss-Newton algorithm). If $\mathbf{v}_i$ is an approximate minimum, a better estimate $\mathbf{v}_{i+1}$ is given by

$$\begin{aligned} \mathbf{v}_{i+1} = \mathbf{v}_i + \mathbf{B}^{-1}\{(\sum_{m=1}^{N} \mathbf{S}_m(\mathbf{a}_m^{-1} - \mathbf{I})) + \mathbf{I}\}^{-1} \\ \times \{(\sum_{m=1}^{N} \mathbf{G}^*_m\mathbf{a}_m^{-1}\mathbf{dx}_m) - \mathbf{B}\mathbf{v}_i\}. \end{aligned} \tag{22}$$

For our basic variable $\mathbf{w}$ the matrix $\mathbf{B}$ conveniently cancels, giving:

$$\begin{aligned} \mathbf{w}_{i+1} = \mathbf{w}_i + \{(\sum_{m=1}^{N} \mathbf{S}_m(\mathbf{a}_m^{-1} - \mathbf{I})) + \mathbf{I}\}^{-1} \\ \times \{(\sum_{m=1}^{N} \mathbf{G}^*_m\mathbf{a}_m^{-1}\mathbf{dx}_m) + \mathbf{w}_b - \mathbf{w}_i\}. \end{aligned} \tag{23}$$

4. Adjoint model G*

We have already assumed that the analysis we require can be expressed in terms of balanced, slowly varying, modes. Furthermore, in the arguments used to justify the commutability of $\mathbf{B}$ and $\mathbf{G}$ we have implicitly assumed that these modes are normal, and the same for all the forecast timesteps $\mathbf{M}_m$. Each complex normal mode of $\mathbf{M}_m$ has associated with it a complex frequency $\omega - il$, so that multiplication of a state by $\mathbf{M}_m$ is equivalent to multiplication of each of its modes by $\exp((i\omega + l)dt)$. Multiplication by the adjoint $\mathbf{M}^*_m$ is thus equivalent to multiplying each mode by $\exp[(-i\omega + l)dt]$. This can be thought of as running energy conserving parts of the model backwards, while retaining diffusion and damping terms. The nonlinear model M_{nm} can easily be modified to do this for its dynamical terms, and for simple physical parameterizations such as diffusion and friction. We denote this modified model by H_{nm}. Multiplication by the adjoint of the linearized model, $\mathbf{M}^*_m$, can thus be approximated by

$$\mathbf{M}^*_m\mathbf{a}_m^{-1}\mathbf{dx}_m = \{H_{nm}(\mathbf{x}_{im} + k\mathbf{a}_m^{-1}\mathbf{dx}_m) - H_{nm}(\mathbf{x}_{im})\}/k. \tag{24}$$

Here k is a small scaling factor chosen so as to improve the approximation we are making in using a perturbation to a nonlinear model instead of the linearized model. Theoretically it should be infinitesimal, but in practice, because of numerical truncation errors in the computation of H_n, a small finite value is used. This multiplication by $\mathbf{M}^*_m$ gives weighted increments valid at time T_{m-1}, further multiplications by $\mathbf{M}^*_{m-1}$ etc. are required to give the equivalent of multiplication by $\mathbf{G}^*_m$; weighted increments valid at the initial time of $\mathbf{w}$. However, because we are approximating a linear adjoint model, it is valid to combine these further multiplications with those necessary for the weighted increments from time T_{m-1}, and so on, so that all the adjoint model integrations $\mathbf{G}^*_m$ can be implemented by a single series of integrations of $\mathbf{M}^*_N, \cdots \mathbf{M}^*_m, \mathbf{M}^*_{m-1}, \cdots \mathbf{M}^*_1$.

5. Initialization

We have based many of the preceding approximations on the assumption that both the background and the estimates to the "best" analysis should be balanced. The 3DOI program, although it attempts to maintain a linear geostrophic balance to the increments, does not necessarily ensure full balance. Neither will our approximated adjoint integration, nor the descent algorithm for finding the state which minimizes $L(\mathbf{v})$. It is necessary therefore to include in the procedure a step which explicitly ensures balance, either by nonlinear normal-mode initialization, or some equivalent means.

In an ideal optimal analysis scheme our prior knowledge that the atmosphere is balanced should be used in the analysis. This can be done linearly, through the eigenmodes of $\mathbf{B}$, or nonlinearly, through an additional penalty in the variational minimization, for instance adding a factor proportional to the mean square change during the first time-step of the forecast to the penalty function J. This would then lead to an additional term containing the adjoint of the forecast model in the iterative analysis. However since we do not have the adjoint of the model operator, we cannot include such a nonlinear penalty. Instead, between iterations, we initialize the new estimate $\mathbf{w}_{i+1}$ using an existing nonlinear normal-mode initialization program. This method of combining initialization with 3DOI in an iterative analysis is a generalization of the unified analysis-initialization technique of Williamson and Daley (1983). Their technique did not include the background field except to start the first iteration, so that iterated indefinitely it tends to the balanced field which fits closest to the observations, independent of the initial background. Our new scheme does include the background, so that in the degenerate case of a single time-period, it will tend to the balanced field which fits closest the observations and the background, with the relative weight for each determined by the 3DOI.

6. Organization of the iterative four-dimensional analysis

a. Basic iteration

The preceding equations and approximations lead to a procedure for the iterative search for an approximate minimum to $J(\mathbf{w})$ as follows:

1) Initialize the current best estimate $\mathbf{w}_i$, to ensure balance.

2) Forecast $G_{nm}(\mathbf{w}_i)$ to obtain the estimates $\mathbf{x}_{im}$ at each time-period $m = 1, N$.

3) Clear accumulators for the weighted increments, and weights. Loop back through the time-periods $m = N, 1, -1$:

(i) Run the 3DOI program using $\mathbf{x}_{im}$ as guess. Calculate the observational penalty $L_{fm}(\mathbf{v}_i)$, the analysis increment $\mathbf{dx}_m$, and smooth the normalized analysis error variance to give $\mathbf{a}_m$.

(ii) Weight the analysis increment by the inverse of the normalized error variance, to give $\mathbf{a}_m^{-1}\mathbf{dx}_m$.

(iii) Add this to the accumulated weighted increments.

(iv) Add $(\mathbf{a}_m^{-1} - \mathbf{I})$ to the accumulated sum of weights. Smooth this with a spatial filter to model the effect of pre- and post-multiplying by the adjoint model matrix.

(v) Add the accumulated weighted increments to $\mathbf{x}_{im}$, initialize, and integrate using the dynamically backward model H_{nm} to get a field valid at time T_{m-1}.

(vi) Similarly initialize and backcast $\mathbf{x}_{im}$, and subtract from the results of 5, to get accumulated weighted increments valid at time T_{m-1}.

4) Add the forcing towards the background, $\mathbf{w}_b - \mathbf{w}_i$, to the accumulated weighted increments valid at time T_1.

5) Add the weight ($\mathbf{I}$) given to the background to the accumulated sum of weights.

6) Divide the accumulated weighted increments by the accumulated sum of weights, and add to $\mathbf{w}_i$ to give a new estimate.

b. First iteration

It seems natural to start the iteration procedure just outlined from the background value $\mathbf{w}_b$. However, in order to save time, it is desirable to reduce the number of iterations required by starting from the best available estimate. If the nominal time T_1 of the first time-period of observations is the initial time at which $\mathbf{w}_b$ is valid, then $\mathbf{x}_{i1}$ is identical to $\mathbf{w}_i$ and a better starting estimate can be obtained by a conventional 3DOI of the time-period 1 data, using the background $\mathbf{w}_b$ as guess. The 3DOI used in this way finds the $\mathbf{w}_1$ which minimizes $J_{f1}(\mathbf{w}_1) + J_b(\mathbf{w}_1)$, or equivalently minimizes $L_{f1}(\mathbf{v}_1) + L_b(\mathbf{v}_1)$. Hence we have the relationship:

$$L'_{f1}(\mathbf{v}_1) + L'_b(\mathbf{v}_1) = 0. \qquad (25)$$

Using this enables us during the first iteration to omit the calculation of these terms. That is, steps 3(i), 3(ii), 3(iii), for the first time-period of observations, and step 4, which would cancel with them, can be omitted. The normalized analysis error $\mathbf{a}_1$ is still needed, but this is available from the 3DOI done to make $\mathbf{w}_1$. Thus, by performing a zeroth iteration which consists solely of the 3DOI, we get a better guess for the first iteration at no net computational cost, since we can omit that 3DOI from the first iteration.

c. Alternative descent algorithms

Our use of the 3DOI program, and the approximate adjoint model, has given us the following information about the components of $L(\mathbf{v}_i)$ and their derivatives:

$L_f(\mathbf{v}_i)$ given by adding the terms $L_{fm}(\mathbf{v}_i)$ from each 3DOI.

$L_b(\mathbf{v}_i)$ unknown. However it was shown by Lorenc (1988) that this should remain small for a few iterations with $\mathbf{v}$ as control variable.

$L_f'(\mathbf{v}_i)$ given approximately from the adjoint model integration.

$L_b'(\mathbf{v}_i)$ given by $\mathbf{w}_i - \mathbf{w}_b$.

$L_f''(\mathbf{v}_i)$ approximated by the diagonal matrix of the accumulated sum of weights, multiplied by $\mathbf{B}$.

$L_b''(\mathbf{v}_i)$ the identity $\mathbf{I}$, multiplied by $\mathbf{B}$.

The Newton descent algorithm is optimal if the penalty function is near quadratic, and if the Hessian is accurately known. We have some gross approximations in our estimation of L_f'', and the forecast model is nonlinear, making the penalty function non-quadratic. The simplest modification to the method is the addition of a step-length s, which has to be determined so as to ensure that the method is converging.

$$\mathbf{v}_{i+1} = \mathbf{v}_i - s\{L''(\mathbf{v}_i)\}^{-1}L'(\mathbf{v}_i). \tag{26}$$

This "damped" Newton method is globally convergent even for nonquadratic penalty functions (Gill et al. 1981). Ideally the step-length should be chosen, using an iterative search, to ensure that the penalty function decreases each main iteration. In the example given in the next section, we have used this method, but with a fixed value for s initially chosen in preliminary tests to be 0.5.

If we were to regard our approximation to L_f'' as worthless, and instead used the identity matrix $\mathbf{I}$ as Hessian, then the above algorithm would become the method of steepest descent, which is known to converge slowly. An improvement on this, which uses information about L and L' remembered from previous iterations, is the method of conjugate gradients (Navon and Legler 1987). This method is related to limited-memory quasi-Newton methods (Gill et al., 1981), which take our approximation to L'' as a first estimate of the Hessian, and refine it in subsequent iterations using differences between L' at different iterations in a finite-difference approximation to a second derivative.

7. Experimental test

a. *Details of method*

As discussed in the Introduction, one objective of this work was to use operational datasets and programs as much as possible, in order to test the four-dimensional ideas in a practical environment. The scheme which we chose to adapt was the Regional Analysis Forecast System (RAFS) of the National Meteorological Center (NMC), as operational during March 1987. This consists of an optimal interpolation (OI) analysis, nonlinear normal-mode initialization (NNMI), and Nested-Grid Model (NGM) forecast. The background field for the analysis comes from a 6-hour forecast from the global data-assimilation system; it is interpolated from a rhomboidal-40 spectral representation to the 180 × 60 longitude–latitude, 16 sigma-level grid used for the hemispheric analysis. Details of the analysis are given by DiMego (1987). It is an OI scheme, multivariate in geopotential height, and wind components. Humidity is analyzed univariately. Height, wind, and humidity data are used to calculate analysis increments at the forecast model's sigma levels, but on a longitude–latitude analysis grid. The height increments are converted to equivalent temperature and surface pressure increments, and the increments are added to the background, which has been interpolated to the same grid. The background error variances used in the analysis are estimated in a simple fashion from the data distributions at the previous analysis in the global data assimilation scheme, using its estimated analysis errors. They thus vary significantly between data-sparse and data-dense areas. The height error correlation is modelled as a function of separation in horizontal distance and pressure. Wind error correlations are calculated to be geostrophically consistent with this model. These estimated error variances and correlations define our background error covariance matrix $\mathbf{B}$. The programs which perform this analysis are referred to collectively as the 3DOI.

The analysis is converted to a rhomboidal-80 spectral representation, and initialized in a hemispheric nonlinear normal-mode initialization (Sela, 1980). The initialized field is interpolated horizontally to the nested polar stereographic grids of the NGM, which have resolutions varying from 366 km for the hemispheric outermost grid to 91.5 km for the grid covering North America.

To use the 3DOI in our scheme, it was modified slightly to provide, in convenient form, the observation penalty $J_{fm}(\mathbf{x}_{im})$, the analysis increment $\mathbf{dx}_{im}$, and the normalized analysis error variance $\mathbf{a}_m$. The observational penalty included deviations from the guess of all height and wind data, normalized by the observational error variance assigned to them in the OI scheme. Data which were rejected by the quality control scheme, which is part of the 3DOI, were assumed to give a constant contribution to the penalty, equal to that of data on the borderline of rejection.

The basic control variable $\mathbf{w}$ was taken to be the vector of the analysis variables on the latitude-longitude-sigma grid. In order to do a forecast from this using the NGM, it was interpolated horizontally to the model's grid points, via the spectral representation used for initialization. Three programs not used operationally complete our iterative scheme:

(1) a backcast model made by removing all physical parameterizations except the diffusive filter and dry adiabatic adjustment from the NGM,

(2) a bi-linear horizontal interpolation from the NGM grids to the analysis grid,

(3) a new program to process the fields, increments, and errors, and implement the descent algorithm.

Since our scheme assumes "balance" in its derivation, and forces balance by incorporation of the NNMI each iteration, an additional step was included to ensure that the background was balanced according to the same criterion. This was achieved by interpolating the background on the analysis grid, as obtained from the global data-assimilation system, via the spectral initialization, to the NGM grid, and then immediately back to the analysis grid.

The observations used in our experiments were also taken from the operational RAFS, which runs every 12 hours. They were thus partitioned into sets nominally valid at the main synoptic hours 00Z and 12Z, including asynoptic observations from up to 3 hours before and about 2 hours after these times.

The computational cost of the scheme was dominated by that of the constituent 3DOI, initialization, and forecasts. The cost per iteration was between one and two times the cost of a conventional forward analysis-forecast cycle assimilation of the same data.

b. Experiments performed

A series of experiments were performed to test the scheme, and its sensitivity to changes in some of its components. Those presented here used the observations for 00Z and 12Z 27 February 1987. They are listed in Table 1.

Experiment A was the basic scheme, with a steplength chosen on the basis of an earlier experiment to be 0.5. This was run for four iterations. The steplength was then halved to 0.25, and a further four iterations performed.

Experiments B and C were three-dimensional analyses incorporating iteratively a nonlinear normal-mode balance relationship, as discussed in section 5. The background for experiment C was interpolated from the global data-assimilation system background valid at the second time: 12Z 27 February 1987. The first iteration of experiment B or C was almost identical to the current (March 1987) operational regional analysis system. Further iterations should improve the nonlinear balance, while maintaining the fit to the observations.

Each iteration of experiment B actually performed a forecast to T_2, and measured the fit to the data, to calculate the observational penalty function. Thus, experiment B can be regarded as identical to A except that the observations at the T_2 were given zero weight. Experiment D was the other extreme from this; the observations at T_1 were given zero weight. It can be thought of as attempting to find the field valid at T_1 which, when forecast, best fits the observations at T_2, subject to constraints on balance and fit to the background at T_1. After two iterations of experiment D it was found that the iteration was not converging, so the steplength s was halved to 0.25.

Experiments E and F were identical to A except that one aspect of the scheme was replaced by a simple dummy version. Thus experiment E used persistence instead of the backcast model in the adjoint calculation, and F omitted all nonlinear normal-mode initializations.

Experiments G, H and I were actually by-products of experiment A. They are included in the forecast results, presented in section e below, to provide comparisons simulating possible alternative practical schemes. Experiment G tested a scheme (called the "periodic spinup"), which is currently being investigated as a compromise between having a completely independent data-assimilation cycle for the regional model, and the operational system which performs each 3DOI with a background interpolated from the lower resolution global model. The UK Meteorological Office (Bell, 1986), and the US Navy (Barker, personal communication) have implemented such schemes with success. Because the basic experiment A had a "zeroth" iteration using only T_1, the analysis produced while calculating the 3DOI increments for iteration 1 T_2 was the product of: 3DOI (T_1) − nonlinear normal-mode initialization − forecast − 3DOI (T_2). This T_2 analysis, from the first iteration of experiment A, is thus the "periodic spin-up" analysis of experiment G. It is the closest analog in this work of an analysis from a conventional data assimilation scheme which performs an indefinitely repeated analysis–forecast cycle. The similar analysis, from the eighth iteration of experiment A, was called experiment H. This experiment can be thought of as using our four-dimensional scheme to provide an improved background, consistent with observed tendencies, to a conventional three-dimensional analysis of the data at the final time. Experiment I tested whether the balance achieved by the iterative scheme was sufficiently good to permit the omission of the nonlinear normal-mode initialization of the final anal-

TABLE 1. Experiments performed.

A	basic scheme: iterative four-dimensional analysis.
B	as A, but only using T_1 observations. Equivalent to iteration of 3DOI and NNMI at T_1.
C	as B, iteration of 3DOI and NNMI at T_2.
D	as A, but only using T_2 observations.
E	as A, with persistence replacing backward NGM in adjoint model.
F	as A, without NNMI.
G	"periodic spinup": 3DOI(T_1) − NNMI − forecast − 3DOI(T_2).
H	as A, then NNMI − forecast − 3DOI(T_2).
I	as A, without NNMI of final analysis.

ysis from A, before the forecast described in section d below.

c. Fit to observations

A necessary property of a good analysis is that it should fit the observations used, within a tolerance governed by the expected observational error. In our formalism this property is measured by the observational penalty J_f. The wind observational penalties for experiments A–F (see Table 1) are plotted in Fig. 1 as a function of iteration. The penalty plotted is that of the current best estimate at the beginning of the iteration; for iteration 0 it is that of the background field. The penalty function was evaluated during each iteration, so that for the latest estimate at the end of the last iteration is not shown. The behavior of the geopotential height observational penalties was similar to that of the wind observations; unfortunately because of a coding error not all values are available for plotting.

If our analysis were the true field, then, because of the definition of the observational errors, the mean observational penalty per datum should be $N_y/2$, where N_y is the number of data. It is easy to show for the OI equations that this is an upper limit; an analysis which has available less than the perfect observation set necessary to analyze the true field, should fit the observations used in the analysis more closely than the true field would. This result only holds if the statistical estimates of observational error variances used in the calculation of J_f are correct. The observational penalties are plotted in Fig. 1 scaled by $2/N_y$, so the values should be between zero and one. This is clearly not true. In an effort to obtain a close fit to the data by the analysis, the observational errors specified to the 3DOI program have been reduced below the theoretically correct values, and the resulting observational penalties are higher than expected. However for our purposes this is not important, since the values are approximately correct, and we are more interested in the relative reduction in penalty than the actual value.

The zeroth iteration of experiments A, B and E were identical; the observations for T_1 were used to update the (initialized) background field. Figure 1a shows at iteration 0, the fit of these observations to the background, and at iteration 1, their fit to the resulting initialized analysis. The effect of the NNMI can be seen by comparing these with the values for experiment F; without a balance constraint a closer fit to the observations is possible. Figure 1b shows the fits of the observations at T_2. At iteration 1 we can compare the values for experiments A, B and E, for a forecast from the iteration 0 analysis, with that for experiment F, which omitted initialization. The NNMI, which degraded the fit to the T_1 data, slightly improved that to the T_2 data. Fit to data which have been used in the analysis is often, as in this case, a poor measure of the likely accuracy of a forecast from the analysis. We can also compare with experiment D iteration 1, a forecast from the initialized background. As we would expect, the T_1 observations do improve the subsequent fore-

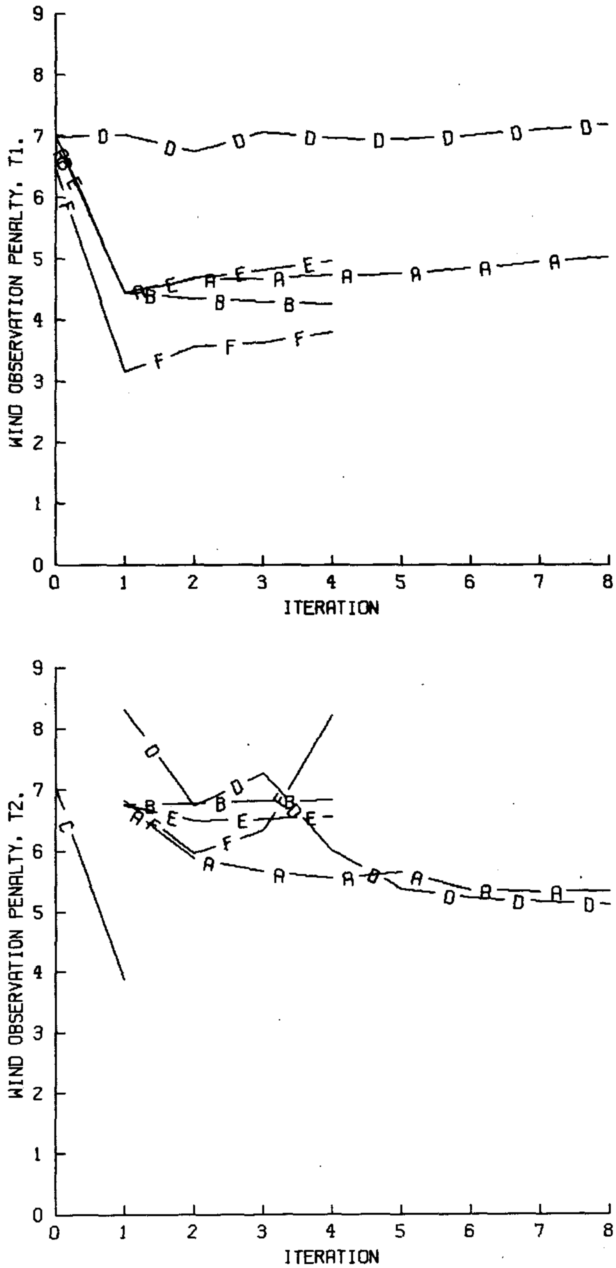


FIG. 1. Observational penalties for wind observations, for the analysis fields valid at times T_1 and T_2, plotted against iteration for the experiments listed in Table 1. The penalty is scaled by $2/N_y$, where N_y is the number of data included, and is thus the mean square deviation of the fields from the observations, normalized by the estimated observational error variance, as used in the 3DOI. Values plotted are for the fields at the beginning of each iteration. See Table 1 for details of each experiment.

cast. Experiment C only used T_2; its iteration 0 value measures the fit of the appropriate six-hour forecast from the global data assimilation cycle, and its iteration 1 value that of a 3DOI analysis at T_2.

Let us now consider the improvements in fit gained by iterating. We can see from both experiments A and D in Fig. 1b that the scheme is managing to find a state at T_1 which, when forecast, better fits the observations at T_2. Hence in a basic way the iteration is working, although the reduction in steplength at iteration 5 of A and iteration 3 of D was necessary for this. Some initialization is necessary, as evidenced by experiment F. Our approximate adjoint of the model integration was also beneficial, as compared to simple persistence used in experiment E. This is particularly true for the wind data penalties shown in Fig. 1, which reflect smaller scales than the geopotential height penalties (not shown). However the success is only partial; the four-dimensional analyses were not as good as the 3DOI of experiment C at fitting the data at T_2. There was no demonstrable benefit at T_1 from the use of the T_2 observations; experiment D fields at T_1 did not fit the observations (which it never used) any better than did the background. The improvement in fit to the T_2 observations, seen in experiment A Fig. 1b, was achieved at the expense of the fit to the observations at T_1 (Fig. 1a), so that the total observation penalty (not shown) for experiment A stayed almost constant. That for experiment E slightly increased. We can at present only speculate on the improvement in these results which might be achieved by a better approximation to the adjoint of the forecast model.

d. Fit to background

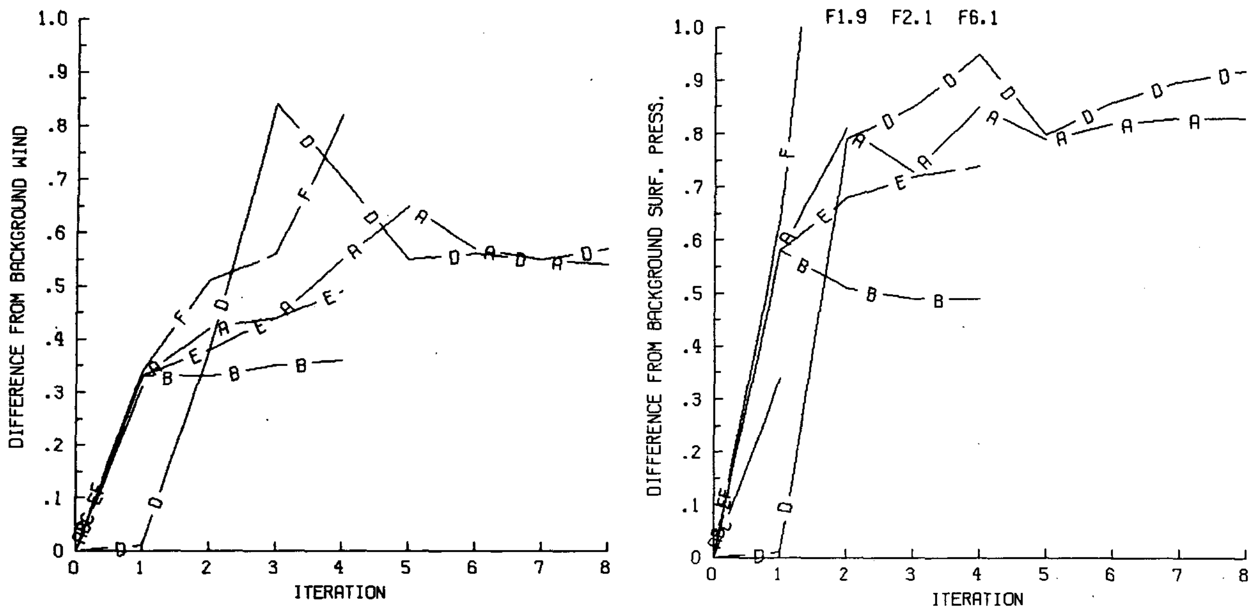
Our prior knowledge about the true state $\mathbf{w}_t$ can be expressed by $\mathbf{w}_b$, the most likely state, and (for a Gaussian system) by $\mathbf{B}$, the error covariance matrix of $\mathbf{w}_b$, defining which modes are more likely to be in error. We have $\mathbf{w}_b$ from a forecast from the global data assimilation system, but we do not have an explicit definition of $\mathbf{B}$. Instead we have an estimate of the prediction error variance, the diagonal of $\mathbf{B}$, and a correlation model used in the 3DOI which implicitly defines the rest of $\mathbf{B}$. Thus we cannot easily calculate the background penalty function. We can however calculate the mean-square deviation from $\mathbf{w}_b$, normalized at each gridpoint by the background error variance; this is plotted in Fig. 2. The correlation model used in the 3DOI is based on assuming smoothness and approximate linear balance in the background errors. We can get a measure of balance from the changes made during the NNMI. These are shown in Fig. 3, also normalized at each gridpoint by the background error variance. Since the background is made to be balanced in these experiments by applying the NNMI to it, imbalance in the analysis implies an imbalance in the deviations from $\mathbf{w}_b$. As for Fig. 1, the values plotted are for the estimate at the beginning of the iteration. The NNMI is applied as the first step in each iteration; Fig. 3 shows the changes made during this NNMI. For iteration zero Fig. 3 shows the changes made during the NNMI of the background. These changes were largely due to imbalances introduced when changing the orography in the background representation, since a forecast field from the global data assimilation system (which has its own NNMI) should otherwise be reasonably balanced.

Lorenc (1988) showed how the transformation to control variable $\mathbf{v}$, from the model variables $\mathbf{w}$, means that during the first few iterations of a descent algorithm the background penalty should remain small. This is partly borne out by Fig. 2 and Fig. 3, however by iteration 5 of experiment A and iteration 3 of experiment D values have got quite large. This deviation from our prior assumptions about the atmosphere was also visible in the corresponding plotted fields. There was a very sharp trough in a strong upper westerly flow at 50°N off the west coast of Canada, with associated maxima and minima in vorticity and vertical motion. This pattern looked very "unmeteorological", indeed without the halving of descent steplength to 0.25 in experiments A and D the NGM forecast failed in the next iteration. With the halved steplength, most of the extremes were removed in the subsequent iterations. The position of the anomalous feature was such that it was probably associated with advection, by the strong upper flow, of 3DOI increments which at T_2 were caused by coastal observations. The approximate adjoint model used in this work would advect these back along the flow to the oceanic position at T_1, where there were few other data. Our neglect of model advection in the corresponding approximate adjoint forecast of the error covariances meant that the increments would then be given inappropriate weights. Experiment E, which did not use the approximate adjoint, instead using a persistence approximation consistent with that used for the error variances, had no such feature. Note however that despite this shortcoming, the net effect of the approximate adjoint was positive.

Experiment F had no NNMI; the small changes plotted in Fig. 3 were caused by the spectral grid transformations. Without its controlling effect, "unmeteorological" features such as that discussed above grew each iteration, adversely affecting the fit to observations (Fig. 1) and the fit to the background (Fig. 2).

e. Forecast results

The ultimate test of any analysis scheme for NWP must be the accuracy of the subsequent forecasts. The RAFS system was developed for short range forecasts for the USA. Our analyses have therefore been tested by running the NGM for 48 hours. For all except ex-



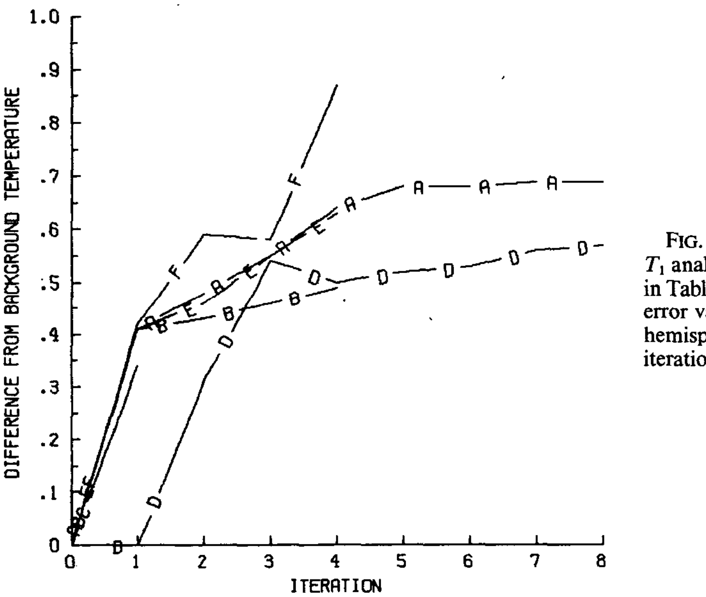
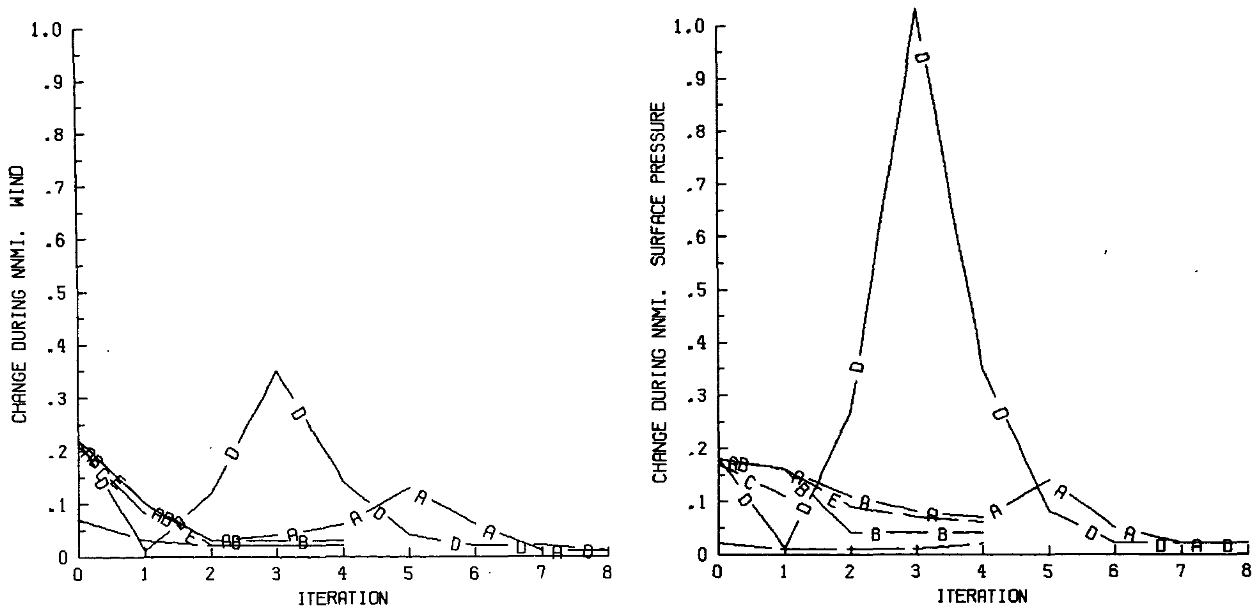


FIG. 2. Mean square deviation from the background field, for the T_1 analysis fields, plotted against iteration for the experiments listed in Table 1. The deviations are normalized by the estimate background error variance, as used in the 3DOI, and averaged for the northern hemisphere. Values plotted are for the fields at the beginning of each iteration. See Table 1 for details of each experiment.

periments F and I, nonlinear normal-mode initialization was performed on the analyses before integrating the forecast model. The forecasts were verified against available observations from the 850, 500, 250 and 100 mb levels, from a standard set of 110 stations in North America. Results, averaged for these levels, are plotted in Fig. 4. In keeping with operational nomenclature, the nominal time of the latest observations available to the analysis (our T_2) is called 0 hours. Curves are labeled with the experiment and cycle number. In keeping with our nomenclature on earlier figures, the cycle number refers to the estimate at the beginning of the cycle. Thus curve A1 is for a forecast from the output field from the zeroth iteration of our basic experiment, which was the input field for iteration 1. Apart from an extra NNMI of the background field, this is equivalent to the operational RAFS analysis valid at T_1 (-12 hours). Curve A9 is from the basic experiment after 8 iterations of the four-dimensional analysis scheme. This analysis has used the 0 hours observations, and hence verifies better against them. The improvement is maintained throughout the forecast. Curve D9 is from a similar four-dimensional analysis only using the 0 hours observations. This forecast is almost as skilful at later times as A9, indicating that the -12 hours observations are in this case adding little



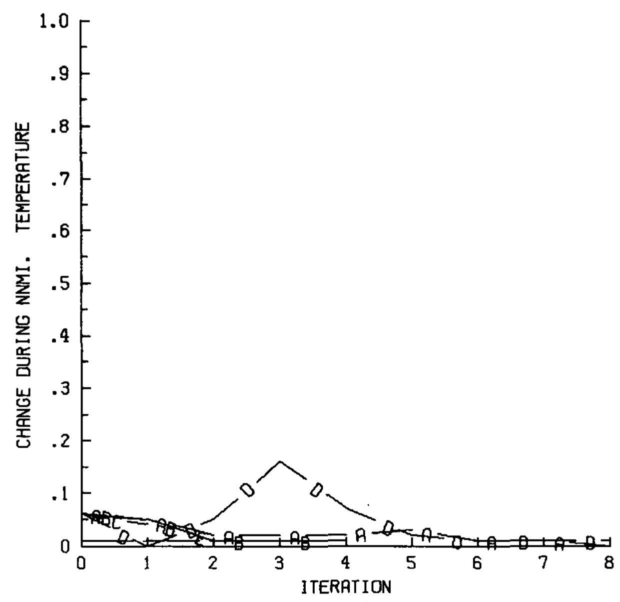


FIG. 3. As in Fig. 2 but for the mean square change during the NNMI.

skill to the forecast, except at the earliest time. (It should be remembered that the four-dimensional analysis is defined by a forecast from the field at the initial time, so these forecast experiments run from −12 hours.)

We saw in Fig. 1b that neither of these experiments achieved as good a fit to the time T_2 (0 hours) wind observations as could be achieved by a simple 3DOI. This is borne out by curves C1, G1 and H8 in Fig. 4b. These were all forecasts from 3DOI analyses at 0 hours, using various backgrounds. The C1 used the initialized 6-hour forecast from the global data assimilation system; it was thus equivalent, 12 hours later, to A1. G1 used the NGM forecast valid at 0 hours from A1. The H8 used the field valid at 0 hours from A8. The better wind verification scores are maintained throughout the forecast; by this criterion the four-dimensional analyses were not as good as the "traditional" three-dimensional ones.

Another comparison that can be made in Fig. 4 is between the "traditional" 3DOI followed by NNMI, and an iteratively balanced three-dimensional analysis as described in section 5. Experiment B iterated the 3DOI and NNMI using the background from the global data assimilation system, and the observations, valid at T_1 (−12 hours). Experiment C did the same for T_2 (0 hours). Scores for the forecast after four iterations

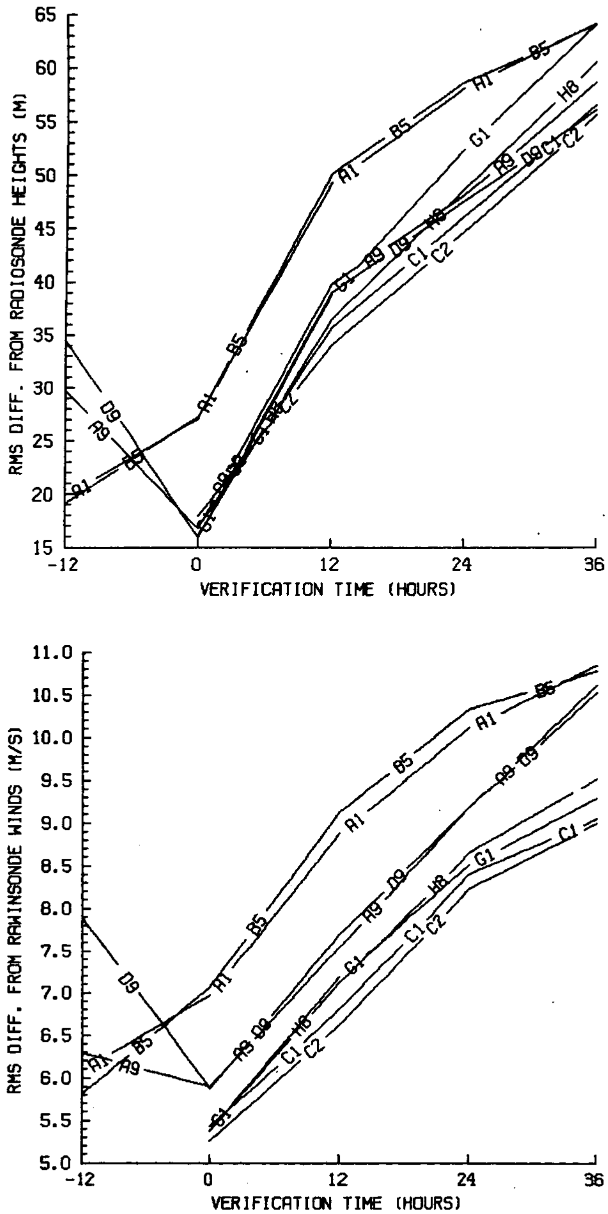


FIG. 4. Root mean square forecast verification statistics averaged for 850, 500, 250 and 100 mb, against radiosondes over North America. Curves are labeled with the experiment letter, as given in Table 1, and the iteration number.

of B are shown as B5, this can be compared with A1, the forecast from the 3DOI analysis of the zeroth iteration of experiment A. Scores after the zeroth and the first iteration of experiment C can similarly be compared. Differences are marginal, and contradictory for the two cases; there is no indication that the iterative analysis is better.

It appears from Fig. 4 that the best forecasts were from experiment C. The distinguishing feature of this experiment was its use of the six-hour forecast background from the global data assimilation system valid at T_2 (0 hours). This has had the benefit of the observations valid at −6 hours. Another difference was the global forecast model. A direct measure of the quality of this background is shown in Fig. 1b, as the fit to the observations at iteration 0 of experiment C. This can be compared with that of experiment B in the same figures. Experiment B only used the T_1 observations, so the T_2 fits measure the accuracy of the resulting 12-hour NGM forecast. The global forecast is better for height, but slightly worse for wind. Note that Fig. 1 uses all observations in the hemispheric RAFS domain, while Fig. 4 only uses North American radiosondes.

Finally in this section, we can mention that the scores for the forecast from experiment I were very little different from those for experiment A. At −12 hours, the time of the observations most directly used in the analysis, the fit to the heights of forecast I9 was 3 meters better than that of forecast A9. The fit of the winds was 0.2 m s^{-1} better. At other times the scores were indistinguishable, so I9 is not plotted in Fig. 4.

8. Effect of approximations

In this section we discuss in turn the effect of the approximations made in deriving a practicable method, as demonstrated in the results of the previous section's experimental test. One type of approximation is the use of operational programs (3DOI, NNMI, and NGM) as if they are perfect. Another type is in the evaluation of the penalty function, its gradient, and their use in a descent algorithm.

a. 3DOI

We assume that the analysis increments given by the 3DOI program are truly an optimal weighting of observations and background. We saw in section 7a that the observational error variances used are far from the theoretically correct values. On the other hand the normalized mean square deviations from the background (Fig. 2) are between zero and one, indicating that the background errors variances are probably more nearly correct. Since the relative weight given to the observations and background depends on the ratio of the assumed error variances, it follows that the 3DOI gives too much weight to the observations. This might partly explain why there is little difference between the scores of forecasts from the analyses A9, G1, D9, H8, C1 and C2 in Fig. 4a. These were all 3DOI analyses using the same data, the only difference was in the background fields used for each.

b. NNMI and NGM

The nonlinear normal-mode initialization is included as a strong constraint in our scheme for two reasons:

(1) because of the observation that the atmosphere is usually slowly varying. NNMI is an approximate way of including this extra knowledge into our scheme, which would otherwise allow rapidly varying solutions.

(2) because the approximations in our handling of the forecast model's adjoint can only be justified for balanced slow modes.

The latter effect is demonstrated in Fig. 1. Experiment F, without NNMI, did not converge to fit the observations. The former effect is shown in Fig. 5. Forecast F5 had a large scale height oscillation of an amplitude not seen in reality. Figure 5 shows the mean verification against North American radiosonde 250 mb heights for forecasts from the experiments related to NNMI. There was a large scale upper trough covering North America on 27 February 1987. The NNMI filled this slightly, as can be seen by the difference between the −12 hours mean errors of forecast F1 and A1, and of forecasts I9 and A9. (The latter difference is probably smaller because the NNMI was used while making the experiment I analysis, it was only omitted before the final forecast). The subsequent forecasts tended to deepen the trough again. This seems to indicate that the balance achieved by the NNMI was not that required by the NGM forecast model. A similar behavior of a similar NMC NNMI scheme in large-scale troughs was noted by Hollingsworth et al. (1985). Note that the −12 hours mean error in experiment D was larger than that of A, because D was only trying to fit the 0 hours observations.

The oscillations in forecast F5 in Fig. 5 were in the external mode; they were as large in the 850 mb height (not shown). In contrast the large change between −12 hours and 0 hours in forecasts from experiments A D and I were mostly in the 850–250 mb thickness. Forecasts initialized at 0 hours (C1, C2, G2, G1 and H8, not shown on Fig. 5) showed a similar decrease in 250 mb height in their first 12 hours. However in these forecasts, most of the change occurred in the 850 mb height. It is unclear how much of the bias error in 250 mb height was due to NNMI, and how much to the NGM forecast model. It is possible that differences in calibration between satellite derived height observations over the oceans and radiosondes over the land also contributed. There is evidence however that part of the bias is due to the NGM. All the forecasts, including the uninitialized ones, showed a steady cooling of the model's lowest layers, about 1°C during the first 24 hours at 850 mb. Further experiments would be required to unravel the causes of these biases. It is clear however that they make it difficult to achieve a close fit to both the −12 hours and 0 hours observations.

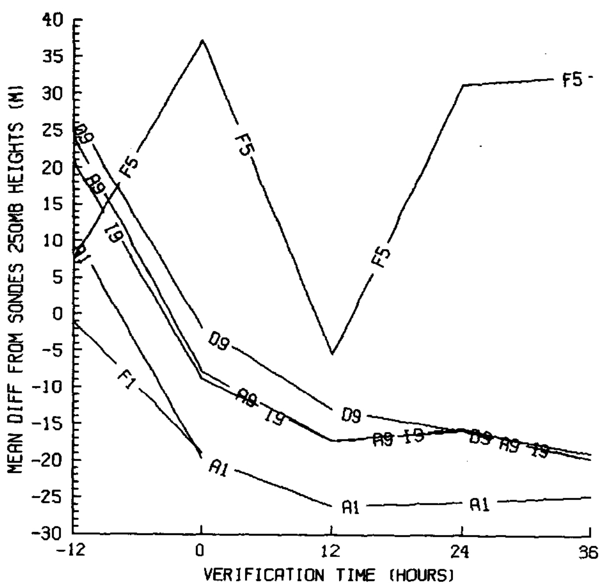


FIG. 5. As in Fig. 4 but for the mean 250 mb height difference of observations minus forecast.

Another failing of NNMI, particularly of adiabatic implementations like that used here, is the underprediction of rainfall in the subsequent short-period forecast. It was hoped that by moving the NNMI to −12 hours, instead of 0 hours as in the current operational system, this spinup problem would be alleviated. However although there were differences between the rainfall forecasts of, for instance, A9 and G1, there was not a clear signal that one was better for the single case studied.

c. Penalty functions, derivatives, and descent algorithm

We discussed in section 6c the approximations made in deriving expressions for the penalty function and its derivatives. Practically, our objective is not to set up an algorithm for finding the exact minimum of the total penalty function, but just to perform a few iterations which decrease it from that given by conventional 3DOI analysis/forecast cycles. The total penalty needs to be used in a practical scheme at least as a check that it is decreasing. The simplest approximation to the Gauss-Newton descent algorithm did not converge; it was necessary to modify it by including a step-length. Some form of total penalty, formed as a weighted sum of the partial penalties shown in Figs. 1, 2 and 3, would probably be a sufficiently good measure of "improvement" in the analysis, for detecting convergence. It would have been possible with such a sum to detect the lack of convergence which eventually forced the halving of the step-length in experiment A iteration 5 and experiment D iteration 3. Note that, in contradiction to the result from Lorenc (1988), the background penalty is important for this. The Lorenc (1988) result that the background penalty always remains small is no longer true in the presence of our other approximations.

Neither experiment A nor experiment D achieved as close a fit to the observations as was achieved in the similar idealized experiments of Lorenc (1988). This was of course to be expected; the earlier experiments were of "identical twin" type, with model generated observations. Thus there were no model errors analogous to those discussed in the last subsection. The exact adjoint used in the idealized experiments enabled changes to be made to the advecting wind in response to tendency information from an advected tracer. The linearized approximation to an adjoint used in the present work was probably not accurate enough to get this effect, which needs a better adjoint for the dynamical part of the model. It would need a large effort to code the accurate adjoint of the full forecast model, including its physical parameterizations (without which its forecasts are significantly degraded). Such an effort is probably premature; research on using observations such as cloud amounts and deduced diabatic heating rates in three-dimensional analysis schemes is still in its infancy. So even if the forecast adjoint is improved, any improvement to the descent algorithm used in this work should take into account that the calculated gradients of the total penalty function are only approximate.

We discussed in section 7d an "unmeteorological" feature, apparently caused by inconsistent approximations in our adjoints of the forecast and the error covariances. Our approximation to the latter is exactly analogous to that used in the operational NMC scheme for estimating background error covariances, something which should be done using a Kalman-Bucy filter. The crude smoothing which replaced this in our adjoint scheme was completely untuned; probably considerable improvements are possible even to this. If the operational scheme were to be improved, for instance by implementing a simple advection of variances by the mean flow, then presumably its adjoint could be used in this scheme.

8. Conclusions

We have shown that a four-dimensional analysis of the full operational observational database can be made, by iterating modifications to the operational analysis and forecast codes and an approximate adjoint model. Computer resources required are only an order of magnitude greater than those for the operational scheme. This is much less than the theoretical requirements of some other proposed algorithms. It means that the technique is practicable now for research experiments, such as producing a dynamically consistent four-dimensional analysis from a special set of observations. It should become operationally practicable by the next generation of computers (as long as the requirements of the forecast model do not grow to match the available computer!).

The derivation of the scheme emphasizes that it can be regarded as an extension of current three-dimensional analysis methods. It should be possible to carry over the results of past and continuing efforts to develop these, by using the three-dimensional analysis code as part of the four-dimensional scheme. Experiments with the scheme highlighted deficiencies in the current operational scheme, in the observational error variances assumed, in the "balance" given by the nonlinear normal-mode initialization, and in the systematic errors of the forecast model.

The preliminary experiments described in this paper indicate that further work is necessary on improving several aspects of the scheme, particularly the descent algorithm and the adjoint forecast of covariances. Forecasts from the analyses did not verify quite as well as those from the operational scheme, for the one case studied.

The scheme has the potential to use a more complete time-coverage of observations. It would be interesting to test this by analyzing data from special observational efforts such as the GALE experiment.

Acknowledgments. This work would not have been possible without the enthusiastic support of NMC Development Division staff in providing and modifying their codes. Thanks are particularly due to Drs. Geoff DiMego, Jim Hoke, Dave Parrish, and Jim Tuccillo. The work was performed while the author was visiting the National Meteorological Center as an exchange scientist from the UK Meteorological Office.

APPENDIX

Penalty Function and Derivatives

The "best" analysis can be obtained by an iterative search for the minimum of

$$L(\mathbf{v}_i) = \mathbf{dy}_i^*(\mathbf{O} + \mathbf{F})^{-1}\mathbf{dy}_i/2 + \mathbf{v}_i^*\mathbf{B}\mathbf{v}_i/2$$
$$= L_f(\mathbf{v}_i) + L_b(\mathbf{v}_i) \quad \text{(A1)}$$

where L_f and L_b are notations for the individual components of L.

To manipulate the four-dimensional distribution of observations using the three-dimensional analysis program 3DOI, we partition them into N time-slices, indicated by subscript m:

$$\mathbf{y}^* = \{\mathbf{y}_1^*, \mathbf{y}_2^*, \cdots \mathbf{y}_m^*, \cdots \mathbf{y}_N^*\}. \quad \text{(A2)}$$

The four-dimensional field $\mathbf{x}$, defined by a forecast G_n from the initial conditions $\mathbf{w}$, is similarly partitioned:

$$\mathbf{x}_i^* = \{\mathbf{x}_{i1}^*, \mathbf{x}_{i2}^*, \cdots \mathbf{x}_{im}^*, \cdots \mathbf{x}_{iN}^*\} \quad \text{(A3)}$$

$$\mathbf{x}_{im} = G_{nm}(\mathbf{w}_i). \quad \text{(A4)}$$

For simplicity we do not interpolate in time, but assume that all observations in time-slice m are valid at

T_m. Hence K_n becomes a space only interpolation at each T_m:

$$\mathbf{y}_{im} = K_{nm}(\mathbf{x}_{im}). \tag{A5}$$

If there is no correlation between observational and representativeness errors in different time-slices, then **O** and **F** can be partitioned into submatrices which can be inverted separately, and the observational penalty L_f can be partitioned into time-slices:

$$L_{fm}(\mathbf{v}_i) = \mathbf{dy}^*_{im}(\mathbf{O}_m + \mathbf{F}_m)^{-1}\mathbf{dy}_{im}/2. \tag{A6}$$

The partitioned total penalty is

$$L(\mathbf{v}_i) = \sum_{m=1}^{N} L_{fm}(\mathbf{v}_i) + L_b(\mathbf{v}_i). \tag{A7}$$

A similar partitioning can be done for the vector of partial first derivatives of the penalty function:

$$L'(\mathbf{v}_i) = \sum_{m=1}^{N} L'_{fm}(\mathbf{v}_i) + L'_b(\mathbf{v}_i) \tag{A8}$$

and for the matrix of partial second derivatives:

$$L''(\mathbf{v}_i) = \sum_{m=1}^{N} L''_{fm}(\mathbf{v}_i) + L''_b(\mathbf{v}_i). \tag{A9}$$

We assume locally valid linearizations **K** and **G** exist:

$$K_{nm}(\mathbf{x}_{im} + \mathbf{dx}_m) = K_{nm}(\mathbf{x}_{im}) + \mathbf{K}_m\mathbf{dx}_m \tag{A10}$$

$$G_{nm}(\mathbf{w}_i + \mathbf{dw}) = G_{nm}(\mathbf{w}_i) + \mathbf{G}_m\mathbf{dw}. \tag{A11}$$

Then we get

$$L'_{fm}(\mathbf{v}_i) = \mathbf{B}\mathbf{G}^*_m\mathbf{K}^*_m(\mathbf{O}_m + \mathbf{F}_m)^{-1}\mathbf{dy}_{im}. \tag{A12}$$

Note that $\mathbf{G}_m$ and $\mathbf{K}_m$ are both in general functions of $\mathbf{v}_i$. Neglecting this dependence in comparison with that of $\mathbf{dy}_{im}$, gives

$$L''_{fm}(\mathbf{v}_i) = \mathbf{B}\mathbf{G}^*_m\mathbf{K}^*_m(\mathbf{O}_m + \mathbf{F}_m)^{-1}\mathbf{K}_m\mathbf{G}_m\mathbf{B}. \tag{A13}$$

The partial derivatives of L_b are

$$L'_b(\mathbf{v}_i) = \mathbf{B}\mathbf{v}_i$$
$$= \mathbf{w}_i - \mathbf{w}_b \tag{A14}$$

$$L''_b(\mathbf{v}_i) = \mathbf{B}. \tag{A15}$$

REFERENCES

Bell, R. S., 1986: The Meteorological Office fine-mesh data assimilation scheme. *Meteor. Mag.*, **115**, 161–177.

Courtier, P., and O. Talagrand, 1987: Variational assimilation of meteorological observations with the adjoint vorticity equations. Part II: Numerical results. *Quart. J. Roy. Meteor. Soc.*, **113**, 1329–1368.

DiMego, G. J., 1987: The National Meteorological Center regional analysis system. *Mon. Wea. Rev.*, in press.

Ghil, M., S. E. Cohn, J. Tavantzis, K. Bube and E. Isaacson, 1981: Applications of estimation theory to numerical weather prediction. *Dynamical meteorology: Data assimilation methods.* L. Bengtsson, M. Ghil, and E. Kallen, Eds., Springer-Verlag, 139–224.

Gill, P. E., W. Murray and M. H. Wright, 1982: Practical optimization. Academic Press. 83–154.

Hollingsworth, A., A. C. Lorenc, M. S. Tracton, K. Arpe, G. Cats, S. Uppala and P. Kallberg, 1985: The response of numerical weather prediction systems to FGGE level IIb data. Part I: Analyses. *Quart. J. Roy. Meteor. Soc.*, **111**, 1–66.

Hoffmann, R. N., 1986: A four dimensional analysis exactly satisfying equations of motion. *Mon. Wea. Rev.*, **114**, 388–397.

Lewis, J. M., and S. C. Bloom, 1978: Incorporation of time continuity into subsynoptic analysis by using dynamical constraints. *Tellus*, **30**, 496–516.

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

Lorenc, A. C., 1986: Analysis methods for numerical weather prediction. *Quart. J. Roy. Meteor. Soc.*, **112**, 1177–1194.

——, 1988: Optimal nonlinear objective analysis. *Quart. J. Roy. Meteor. Soc.*, **114**, 205–240.

Navon, I. M., and D. M. Legler, 1987: Conjugate-gradient methods for large scale minimization in meteorology. *Mon. Wea. Rev.*, **115**, 1479–1502.

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

Sela, J. G., 1980: Spectral modeling at the National Meteorological Center. *Mon. Wea. Rev.*, **108**, 1279–1292.

Williamson, D., and R. Daley, 1983: A unified analysis-initialization technique. *Mon. Wea. Rev.*, **111**, 1517–1536.