TECHNICAL MEMORANDUM

260

Development of a simplified Kalman filter

M. Fisher

Research Department

September 1998 (SAC Paper)

European Centre for Medium-Range Weather Forecasts
Europäisches Zentrum für mittelfristige Wettervorhersage
Centre européen pour les prévisions météorologiques à moyen

Development of a simplified Kalman Filter

Mike Fisher

Introduction

A practical methodology for extending the ECMWF 4D-Var analysis system to a simplified Kalman Filter was suggested by Courtier (1993). The first practical demonstration of the system was made by Rabier **et al.** (1997). This paper documents the continuing development of a simplified Kalman Filter at ECMWF. It extends the preliminary results presented by Rabier **et al.** (1997) to include more convincing evidence of a positive impact on forecast scores.

A companion paper (Ehrendorfer, 1998) presents an explicit low-resolution extended Kalman filter. Although only recently developed, this is already providing insight into the behaviour of the simplified Kalman filter. In particular, it has provided estimates of the fractions of the variances of analysis and forecast error which are represented by the low-dimension subspace which is explicitly evolved by the simplified Kalman filter. The explicit low-resolution Kalman filter is expected to become an invaluable tool for diagnosing and testing the simplified Kalman filter.

Four-dimensional variational data assimilation (4D-Var) expresses the problem of optimally combining observational and forecast data to produce an analysis of the state of the atmosphere, in terms of the minimization of a cost function

$$J(\delta\mathbf{x}) = \delta\mathbf{x}^{\mathrm{T}}\mathbf{B}^{-1}\delta\mathbf{x} + \sum_{i=0}^{M}(\mathbf{H}_i\mathbf{x}_i - \mathbf{y}_i)^T\mathbf{R}_i^{-1}(\mathbf{H}_i\mathbf{x}_i - \mathbf{y}_i) \qquad (1)$$

where the sum is over a set of "time windows" of length a few timesteps of the forecast model. For each time-window, $\mathbf{x}_i$ is calculated by a short integration of the model with initial conditions $\mathbf{x}_{i-1}$. Each time window groups together observations, represented by the vector $\mathbf{y}_i$, which are treated as if contemporaneous. The matrices $\mathbf{R}_i$ are covariance matrices of errors in the observation departures $(\mathbf{H}_i\mathbf{x}_i - \mathbf{y}_i)$, and the operators $\mathbf{H}_i$ map model variables to observations.

The first term on the right hand side of equation 1 is the background term, conventionally denoted J_b. Here, $\delta\mathbf{x} = (\mathbf{x}_0 - \mathbf{x}_b)$ represents the departures of the model variables at the start of the analysis from the background $\mathbf{x}_b$. In a cycling analysis system, the background fields are the model fields produced at timestep M of the preceding analysis. $\mathbf{B}$ is an approximation to the covariance matrix of background error.

It is known (see, for example, Daley 1991) that for a perfect, linear model and linear observation operators, both 4D-Var and the Kalman filter (Kalman, 1960) give the same values for the model variables at the end of the 4D-Var assimilation period (step M in equation 1). It is also known (see, for example, Jaszwinski 1970) that the Kalman filter is optimal in the sense of producing the best linear unbiased estimate of the model state at step M. The fundamental difference between the two analysis methods is that the Kalman filter explicitly propagates the covariance matrix for errors in the model variables. 4D-Var does not. Thus, if we wish to produce a new analysis using the model variables at step M of the current analysis as the background, we are forced, in 4D-Var, to use an approximation to the covariance matrix of errors in the background. The approximation is typically based on climatological error statistics, and is subject to a number of simplifications in order that the matrix may be represented in terms of a small number of parameters.

It is not practical, for an NWP model, to propagate explicitly the covariance matrix of errors in the model state. Indeed, the matrix is too large even to be stored in current computer memories. We must therefore restrict our attention to a small subset of the information contained by the matrix. The remainder must be represented more approximately (for example

using the climatological background error covariance matrix of 4D-Var). Any analysis method which results from restricting the explicit covariance evolution of the Kalman filter to the evolution of a subset of the information it contains, will be referred to as a **simplified** Kalman filter.

This paper describes one approach to the construction a simplified Kalman filter. In the next section, the mathematical formulation is outlined. Following this, the implications of the particular choices made in its formulation are discussed, as are the restrictions and approximations imposed by the requirement that the analysis should run within available computational resources.

The ensemble Kalman filter (Evensen 1994, Evensen and van Leeuwen 1996, and Houtekamer and Mitchell 1998) represents an alternative approach to the approximation of the Kalman filter. The similarities and differences between this approach and the simplified Kalman filter are discussed.

Some experimental results are discussed in the penultimate section, which is followed by some tentative conclusions.

Mathematical Formulation

In the preceding section, it was shown that the fundamental difference between 4D-Var and the Kalman filter lies in the specification of the covariance matrix of background error, which in 4D-Var must be approximated, but which in the Kalman filter is generated explicitly by the preceding analysis cycle. Construction of a simplified Kalman filter amounts to improving the approximation of the covariance matrix of background error in 4D-Var, to include dynamical evolution of a subset of covariance information.

The first task in formulating the simplified Kalman filter is to identify the subset which should be subject to explicit evolution. The most general way to do this is to identify a subspace of the phase space of the model. This may be done by choosing a set of directions, $\{\mathbf{s}_k; k = 1 \ldots K\}$ which span the subspace.

Once the subspace has been identified, we may partition the background departure, $\delta\mathbf{x}$, into a component $\delta\mathbf{x}_S$ which represents the projection of $\delta\mathbf{x}$ onto the directions $\mathbf{s}_k$, and a component $\delta\mathbf{x}_{\bar{S}}$ which is orthogonal to $\{\mathbf{s}_k; k = 1 \ldots K\}$.

In a multivariate analysis system, the partitioning of $\delta\mathbf{x}$ must be done with care. It is clearly nonsensical to use the Euclidean inner product to define projection and orthogonality for vectors whose components include, for example, vorticities and temperatures. At the least, the vectors must be non-dimensionalised. In 4D-Var, the matrix $\mathbf{B}^{-1}$ defines an obvious inner product:

$$\langle \mathbf{x},\mathbf{y} \rangle_{\mathbf{B}} \equiv \mathbf{x}^{\mathrm{T}} \mathbf{B}^{-1} \mathbf{y} \tag{2}$$

The vectors $\delta\mathbf{x}_S$ and $\delta\mathbf{x}_{\bar{S}}$ are defined as the projection onto the subspace and the orthogonal complement of the background departures with respect to the inner product $\langle .,. \rangle_{\mathbf{B}}$. That is,

$$\begin{aligned} \delta\mathbf{x}_S &= \sum_{k=1}^{K} \alpha_k \mathbf{s}_k \\ \delta\mathbf{x}_{\bar{S}} &= \delta\mathbf{x} - \delta\mathbf{x}_S \end{aligned} \tag{3}$$

where the coefficients α_k are fully determined from the requirement that $\langle \delta\mathbf{x}_{\bar{S}}, \mathbf{s}_k \rangle_{\mathbf{B}} = 0$ for $k = 1 \ldots K$. (For the special case that the vectors $\mathbf{s}_k$ are mutually orthogonal with respect to the inner product $\langle .,. \rangle_{\mathbf{B}}$, we have $\alpha_k = \langle \mathbf{s}_k, \delta\mathbf{x} \rangle_{\mathbf{B}} / \langle \mathbf{s}_k, \mathbf{s}_k \rangle_{\mathbf{B}}$.)

The background term of the 4D-Var cost function (that is, the first term on the right hand side of equation 1) may be split into three contributions

$$J_b = \delta \mathbf{x}_S^T \mathbf{B}^{-1} \delta \mathbf{x}_S + 2\delta \mathbf{x}_{\bar{S}}^T \mathbf{B}^{-1} \delta \mathbf{x}_S + \delta \mathbf{x}_{\bar{S}}^T \mathbf{B}^{-1} \delta \mathbf{x}_{\bar{S}} \quad (4)$$

The first term involves only the components of $\delta \mathbf{x}$ which lie in the subspace identified by the vectors $\{\mathbf{s}_k; k = 1 \ldots K\}$. The last term in equation 4 involves only components which are orthogonal to the subspace. The remaining term involves cross-covariances between the identified subspace and its orthogonal complement. It is simply twice the inner product of $\delta \mathbf{x}_S$ and $\delta \mathbf{x}_{\bar{S}}$, and is identically zero, since these vectors are orthogonal with respect to the inner product $\langle .,. \rangle_{\mathbf{B}}$.

In the simplified Kalman filter, the last term of equation 4 is evaluated in the same way as in 4D-Var. The first two terms contain components in the identified directions. The simplified Kalman filter attempts to optimize the analysis in these directions by replacing the approximate covariance matrix of background error $\mathbf{B}$ in the first two terms of equation 4 by a matrix which is close to the true covariance matrix of background errors, $\mathbf{P}_{\mathbf{x}}^f$.

The background term of the cost function in the simplified Kalman filter is

$$J_b = \delta \mathbf{x}_S^T (\mathbf{P}_{\mathbf{x}}^f)^{-1} \delta \mathbf{x}_S + 2\alpha \delta \mathbf{x}_{\bar{S}}^T (\mathbf{P}_{\mathbf{x}}^f)^{-1} \delta \mathbf{x}_S + \delta \mathbf{x}_{\bar{S}}^T \mathbf{B}^{-1} \delta \mathbf{x}_{\bar{S}} \quad (5)$$

The factor α in the second term of equation 5 is required to ensure that the background cost function remains positive definite. A typical value of α is 0.5. Note that in 4D-Var, the second term in equation 5 is identically zero. So, the introduction of the factor α represents a reduction of the cross-correlation term towards the value it takes in 4D-Var.

To proceed further, let us introduce a change of variable matrix, $\mathbf{L}$, with the property $\mathbf{L}\mathbf{L}^T = \mathbf{B}$. The background cost function may then be written as

$$J_b = \chi_S^T (\mathbf{P}_{\chi}^f)^{-1} \chi_S + 2\alpha \chi_{\bar{S}}^T (\mathbf{P}_{\chi}^f)^{-1} \chi_S + \chi_{\bar{S}}^T \chi_{\bar{S}} \quad (6)$$

where $\chi_S = \mathbf{L}^{-1} \mathbf{x}_S$, $\chi_{\bar{S}} = \mathbf{L}^{-1} \mathbf{x}_{\bar{S}}$ and where $\mathbf{P}_{\chi}^f = \mathbf{L}^{-1} \mathbf{P}_{\mathbf{x}}^f \mathbf{L}^{-T}$ is the true covariance matrix of background error transformed according to the change of variable.

The change of variable is not uniquely defined by the requirement that $\mathbf{L}\mathbf{L}^T = \mathbf{B}$. In particular, we may replace the matrix $\mathbf{L}$ by $\mathbf{L}\mathbf{X}$ where $\mathbf{X}$ is any orthogonal matrix (that is, $\mathbf{X}$ is a non-singular matrix for which $\mathbf{X}\mathbf{X}^T = \mathbf{I}$.) The simplified Kalman filter uses this non-uniqueness to define a change of variable which sets to zero all except the first K elements of the vectors $\{\mathbf{L}^{-1}\mathbf{s}_k; k = 1 \ldots K\}$. Since χ_S is the (Euclidean) projection of the background departures onto the space spanned by the vectors $\mathbf{L}^{-1}\mathbf{s}_k$, it follows that only the first K elements of χ_S are non-zero. Consequently, to evaluate the background cost function (equation 6) it is sufficient to know the first K columns of $(\mathbf{P}_{\chi}^f)^{-1}$.

The details of the construction of the orthogonal matrix $\mathbf{X}$ were discussed by Rabier *et al.* (1997). Clearly, $\mathbf{X}$ cannot be stored as a full matrix. Rather, it is constructed as a sequence of K Householder transformations, each of which is fully specified by the elements of a single vector of the same dimension as the control vector.

Evaluation of the background cost function (equation 6) requires knowledge of the first K columns of $(\mathbf{P}_{\chi}^f)^{-1}$. Suppose now that, for each of the vectors $\mathbf{s}_k$, we have a vector $\mathbf{z}_k$ which represents the action of the inverse of the true covariance matrix of background error on $\mathbf{s}_k$. That is,

$$\mathbf{z}_k = (\mathbf{P}_{\mathbf{x}}^f)^{-1} \mathbf{s}_k \quad (7)$$

Replacing $\mathbf{P}_{\mathbf{x}}^f$ in equation 7 by $\mathbf{L}\mathbf{P}_{\chi}^f\mathbf{L}^T$, multiplying by $\mathbf{L}^T$ and rearranging gives

$$\mathbf{L}^T \mathbf{z}_k = (\mathbf{P}_{\chi}^f)^{-1} (\mathbf{L}^{-1} \mathbf{s}_k) \quad (8)$$

Now, let $\mathbf{Z}$ be the rectangular matrix whose columns are the vectors $\mathbf{L}^T\mathbf{z}_k$, and let $\mathbf{S}$ be the $K \times K$ matrix whose columns are the first K elements of the vectors $\mathbf{L}^{-1}\mathbf{s}_k$. (Remember that the change of variable matrix, $\mathbf{L}$, has been constructed so that only the first K elements of the vectors $\mathbf{L}^{-1}\mathbf{s}_k$ are non-zero.) Also, let us partition $(\mathbf{P}_\chi^f)^{-1}$:

$$(\mathbf{P}_\chi^f)^{-1} = \begin{pmatrix} \mathbf{E} & \mathbf{F}^T \\ \mathbf{F} & \mathbf{G} \end{pmatrix} \tag{9}$$

where $\mathbf{E}$ is a square matrix of dimension $K \times K$.

With these definitions we may write equation 8 as

$$\mathbf{Z} = \begin{pmatrix} \mathbf{E} & \mathbf{F}^T \\ \mathbf{F} & \mathbf{G} \end{pmatrix} \begin{pmatrix} \mathbf{S} \\ \mathbf{0} \end{pmatrix} \tag{10}$$

Equivalently,

$$\mathbf{Z} = \begin{pmatrix} \mathbf{E} \\ \mathbf{F} \end{pmatrix} \mathbf{S} \tag{11}$$

The matrix $\mathbf{S}$ is non-singular, provided that the change of variable is also non-singular and that the vectors $\{\mathbf{s}_k; k = 1 \ldots K\}$ are linearly independent. It is also of low dimension ($K \times K$). Hence, we may multiply equation 11 to the right by $\mathbf{S}^{-1}$ to get an equation for the elements of the matrices $\mathbf{E}$ and $\mathbf{F}$, i.e. for the first K columns of the inverse of the true background error covariance matrix in control-space.

To summarise, the background error term of the simplified Kalman filter consists of the following steps. First, a subspace is identified. Next, the background departures are split into a part which projects (in the sense of the $\langle .,. \rangle_{\mathbf{B}}$ inner product) onto the subspace, and a remainder which is orthogonal to the subspace. This allows the background cost to be split into three contributions: one due to covariances within the subspace; another due to covariances in the space orthogonal to the subspace; and a third resulting from cross-covariance between the subspace and its orthogonal complement.

The approximate inverse covariance matrix of background error used in 4D-Var is retained for the contribution due to covariances in the space orthogonal to the subspace, but is replaced by a matrix which is closer to the inverse of the true covariance matrix of background errors, $\mathbf{P}_\mathbf{x}^f$, for the other contributions to the background cost.

The background term is entirely specified by two sets of vectors, $\{\mathbf{s}_k; k = 1 \ldots K\}$ and $\{\mathbf{z}_k; k = 1 \ldots K\}$. The former define the subspace whereas the latter define the action of the inverse of the true covariance matrix of background error on the subspace. It is worth noting that this allows considerable flexibility both in the choice of subspace, and in the degree of approximation used to specify $\mathbf{P}_\mathbf{x}^f$.

It remains to demonstrate how the vectors $\mathbf{s}_k$ and $\mathbf{z}_k$ may be generated at reasonable computational cost. The method which has been adopted makes use of the Hessian singular vector calculation developed by Barkmeijer *et al.* (1998). This calculation produces the leading eigenvectors of the following generalized eigenvector problem

$$\mathbf{M}_{0 \to T}^T \mathbf{W} \mathbf{M}_{0 \to T} \mathbf{s}(0) = \lambda (\mathbf{P}_\mathbf{x}^a)^{-1} \mathbf{s}(0) \tag{12}$$

Here, $\mathbf{M}_{0 \to T}$ represents the resolvent of the tangent linear model integrated for the period $0 \le t \le T$. The matrix $\mathbf{W}$ is positive definite and defines a norm by which the size of perturbations at time T may be measured. $\mathbf{P}_\mathbf{x}^a$ is the covariance

matrix of analysis error, and defines a norm by which the size of perturbations at time $t = 0$ may be measured. If analysis errors are assumed to have a Gaussian distribution, the norm at time $t = 0$ is proportional to the likelihood (i.e. the log of the probability) of the perturbation according to the analysis error distribution. In this case, the leading eigenvectors of equation 12 are the vectors which, for a given initial likelihood, attain maximum size (as measured by the **W**-norm) at time T.

In practice, the covariance matrix of analysis error is not available. However, as shown by Rabier and Courtier (1992) its inverse is approximated by the Hessian matrix of the cost function (the approximation being consistent with approximations made in specifying the observation operators and the covariance matrices of background and observation error). The Hessian singular vector calculation replaces $(\mathbf{P}_{\mathbf{x}}^{a})^{-1}$ in equation 12 by the Hessian matrix of a 3D-Var analysis.

Hessian singular vectors would seem to be good candidates to define a subspace for the simplified Kalman filter, since they represent directions in which errors are likely to grow rapidly during the initial stages of a forecast. They have the additional advantage that the vectors $\{\mathbf{z}_k; k = 1 \ldots K\}$, which are needed to specify the background cost function of the simplified Kalman filter, may be generated at no additional cost during the Hessian singular vector calculation. To demonstrate this, let us split the resolvent of the tangent linear model in equation 12 into a product of resolvents:

$$\mathbf{M}_{0 \to T} = \mathbf{M}_{t \to T} \mathbf{M}_{0 \to t} \tag{13}$$

where t is some intermediate time with $0 \le t \le T$.

Equation 12 may then be written as

$$\mathbf{M}_{0 \to t}^{T} \mathbf{M}_{t \to T}^{T} \mathbf{W} \mathbf{M}_{t \to T} \mathbf{M}_{0 \to t} \mathbf{s}(0) = \lambda (\mathbf{P}_{\mathbf{x}}^{a})^{-1} \mathbf{s}(0) \tag{14}$$

Assuming invertible tangent linear dynamics[1] and multiplying to the left by $\mathbf{M}_{0 \to t}^{-T}$ gives, after a little rearrangement,

$$\mathbf{M}_{t \to T}^{T} \mathbf{W} \mathbf{M}_{t \to T} \mathbf{s}(t) = \lambda \left(\mathbf{M}_{0 \to t} \mathbf{P}_{\mathbf{x}}^{a} \mathbf{M}_{0 \to t}^{T} \right)^{-1} \mathbf{s}(t) \tag{15}$$

where $\mathbf{s}(t) = \mathbf{M}_{0 \to t} \mathbf{s}(0)$ is the vector $\mathbf{s}(0)$, evolved using the tangent linear dynamics to time t.

Now, $\mathbf{P}_{\mathbf{x}}^{a}$ is the covariance matrix of analysis error. Hence, if model errors are assumed to be negligible for the interval $[0,t]$, then the matrix $\mathbf{M}_{0 \to t} \mathbf{P}_{\mathbf{x}}^{a} \mathbf{M}_{0 \to t}^{T}$ which appears on the right hand side of equation 15 is the covariance matrix of forecast error at time t for a forecast whose initial conditions are provided by the analysis at time 0.

If we define the vector $\mathbf{z}(t)$ to be $1/\lambda$ times the left hand side of Equation 15, we may write

$$\mathbf{z}(t) = (\mathbf{P}_{\mathbf{x}}^{f})^{-1} \mathbf{s}(t) \tag{16}$$

The vectors $\mathbf{s}(t)$ and $\mathbf{z}(t)$ are in precisely the relationship required of the pairs of vectors $\{\mathbf{s}_k; k = 1 \ldots K\}$ and $\{\mathbf{z}_k; k = 1 \ldots K\}$ (equation 7). We may therefore identify $\{\mathbf{s}_k; k = 1 \ldots K\}$ and $\{\mathbf{z}_k; k = 1 \ldots K\}$ as the vectors $\mathbf{s}(t)$ and $\mathbf{z}(t)$ corresponding to the K leading eigenvalues of equation 12. For each converged eigenvalue, the vectors $\mathbf{s}(t)$ and $\lambda \mathbf{z}(t)$ are generated during the course of the Hessian singular vector calculation. They are therefore available at no extra cost.

1. The assumption of invertible dynamics is not fundamental to the derivation of equation 16, but has been made to simplify the notation. A derivation involving, for example, the Moore-Penrose generalized inverse of $\mathbf{M}_{0 \to t}^{T}$ is possible.

Discussion

The simplified Kalman filter produces an approximation to the inverse of the covariance matrix of background error. That is, it attempts to improve the weights given to background departures in the background cost function of 4D-Var. An alternative approach is to improve the representation of the background covariance matrix itself. It is not clear which is the better approach. Under the assumption of a Gaussian error distribution, the background cost is a measure of the likelihood (i.e. the logarithm of the probability) of a given departure from the background. On the other hand, it is the background covariance matrix, not its inverse, which defines the structures generated by the analysis. The covariance matrix is arguably a more fundamental quantity than its inverse.

Ultimately, it is the Kalman gain matrix, $\mathbf{K} = \mathbf{P}^f\mathbf{H}^T(\mathbf{R}+\mathbf{H}\mathbf{P}^f\mathbf{H}^T)^{-1}$, which determines the analysis increments. The ensemble Kalman filters described by Evensen (1994) and by Houtekamer and Mitchell (1998) produce estimates of the matrices $\mathbf{P}^f$ and $\mathbf{H}\mathbf{P}^f\mathbf{H}^T$ based on the ensemble members. The matrix $(\mathbf{R}+\mathbf{H}\mathbf{P}^f\mathbf{H}^T)$, is then explicitly inverted (or factorized) to form the Kalman gain matrix. This is easily performed using standard techniques for the few hundred observations used by both Evensen and by Houtekamer and Mitchell. It is not practical for the 10^5 or more observations assimilated during a typical operational NWP analysis.

Equation 9 represented the inverse of the true covariance matrix of background error in the space of the analysis control variable as a partitioned matrix. The simplified Kalman filter tries to retain the sub-matrices $\mathbf{E}$ and $\mathbf{F}$, but replaces the matrix $\mathbf{G}$ by the corresponding static approximation, which for the control vector is simply the identity matrix. The inverse of the background error covariance matrix for the control vector of the simplified Kalman filter is

$$\begin{pmatrix} \mathbf{E} & \mathbf{F}^T \\ \mathbf{F} & \mathbf{I} \end{pmatrix} \tag{17}$$

The inverse of this matrix may be written as

$$\begin{pmatrix} (\mathbf{E}-\mathbf{F}^T\mathbf{F})^{-1} & -(\mathbf{E}-\mathbf{F}^T\mathbf{F})^{-1}\mathbf{F}^T \\ -\mathbf{F}(\mathbf{E}-\mathbf{F}^T\mathbf{F})^{-1} & \mathbf{I}+\mathbf{F}(\mathbf{E}-\mathbf{F}^T\mathbf{F})^{-1}\mathbf{F}^T \end{pmatrix} \tag{18}$$

None of the sub-matrices in equation 18 is identical with the corresponding sub-matrix of the true covariance matrix of background error, except in the special case $\mathbf{F} \equiv \mathbf{0}$. This occurs when the subspace defined by the vectors $\{\mathbf{s}_k; k = 1 \ldots K\}$ coincides with a subspace spanned by K eigenvectors of the true covariance matrix of background error in the space of the control variable. This corresponds to the original outline of the simplified Kalman filter presented by Courtier (1993). The current simplified Kalman filter includes this as a special case. It corresponds to the choice $T = t$ and $\mathbf{W} = \mathbf{B}^{-1}$ in the Hessian singular vector calculation. At this stage, it is unclear whether the advantage of correctly specifying both the background error covariance matrix and its inverse in the selected subspace outweighs the potential advantage of optimising the analysis in a subspace in which errors will grow rapidly during the initial stages of the subsequent forecast.

An important practical reason for constructing an approximation to the inverse covariance matrix rather than to the covariance matrix itself is that it allows the approximation to be built from the pairs of vectors, $\mathbf{z}_k$ and $\mathbf{s}_k$ produced by the Hessian singular vector calculation. A very similar formulation to that presented above would allow the construction of an approximation to the covariance matrix from pairs of vectors $\mathbf{y}_k$ and $\mathbf{s}_k$ satisfying $\mathbf{y}(t) = \mathbf{P}_{\mathbf{x}}{}^f\mathbf{s}(t)$.

In addition to the explicit approximation which is the basis of the simplified Kalman filter, it is necessary in practice to make further simplifications. The main simplification is the use of a 3D-Var Hessian in the Hessian singular vector calculation. The background error covariance matrix used in the singular vector calculation is the same as is used in the simplified Kalman filter analysis. However, since the analysis is performed using 4D-Var, whereas the Hessian singular

vector calculation uses the 3D-Var Hessian, the vectors $\mathbf{z}_k$ calculated by the Hessian singular vector calculation must necessarily be only approximations to the action of the true inverse forecast error covariance matrix on vectors $\mathbf{s}_k$.

Several further approximations are necessary to reduce the computational cost of the Hessian singular vector calculation to a reasonable level given current computing resources. The resolution is reduced to T42 (the main analysis is currently performed at T63). Scatterometer and SSMI observations are excluded, as are observations of all types for latitudes south of 20°S. TOVS observations are thinned to approximately half the horizontal resolution at which they are used in the operational analysis.

The formulation assumes invertibility of the change of variable matrix, $\mathbf{L}$. In particular, calculation of the matrices $\mathbf{E}$ and $\mathbf{F}$ requires that the vectors $\mathbf{z}_k$ be transformed to control space using $\mathbf{L}^T$, whereas the vectors $\mathbf{s}_k$ are transformed to control space using $\mathbf{L}^{-1}$. However, the normalization of background departures in grid space by the standard deviations of background error, which forms part of the change of variable, is not invertible. This is a consequence of the larger number of degrees of freedom in grid space than in spectral space. A consequence of the non-invertibility of the change of variable is that the matrix $\mathbf{E}$ becomes significantly asymmetric. Of course, this asymmetry is easily corrected by replacing $\mathbf{E}$ by $(\mathbf{E}+\mathbf{E}^T)/2$, but it is clear that the non-invertibility of the change of variable remains a major source of numerical inaccuracy in the construction of the matrices $\mathbf{E}$ and $\mathbf{F}$.

The ensemble Kalman filter (Evensen, 1994) was briefly discussed above. There are clearly some similarities between the ensemble Kalman filter and the simplified Kalman filter presented in this paper. Both approximate the covariance evolution for a small subspace. The main difference between the two systems is the way in which this approximation is generated.

In the ensemble Kalman filter, the forecasts which define the subspace are assumed to form a sample drawn at random from a population of forecasts with representative forecast error. Consequently, the covariances between members of the sample approximate the true covariances of forecast error. As the sample size increases, the approximation converges towards the true covariance matrix of forecast error at a rate proportional to the square-root of the number of forecasts which form the sample.

The sample size is inevitably several orders of magnitude smaller than the size of the covariance matrix. This has two important consequences. First, there are directions in phase space which are orthogonal to all the vectors which define the covariance matrix. This is referred to as the "rank problem" by Houtekamer and Mitchell (1998). Their solution is to impose a cut-off radius for the distance between observations and analysis points. If the cut-off radius is sufficiently small, then the number of observations contributing to the analysis at a given analysis point is smaller than the number of forecasts in the sample. Houtekamer and Mitchell (*op. cit.*) show that in this case, the representer matrix $\mathbf{HP}^f\mathbf{H}^T$ (i.e. the forecast error covariance matrix expressed in the space of observations) is of full rank, and the analysis problem is mathematically well formulated.

While the use of a cut-off radius provides a technical solution to the rank problem, its use compromises one of the main advantages of the variational approach to data assimilation; namely the global nature of the analysis. The inter-relationship between the cut-off radius, the number of observations and the number of members in the ensemble is particularly undesirable. It is likely that, as the number of independent pieces of information to be assimilated increases, the approach will eventually require either unfeasibly large ensembles, or too-small cut-off radii.

Evensen and van Leeuwen (1996) chose a different method of attacking the rank problem. They employed an eigenvalue decomposition of the sum of the representer matrix and the observation error covariance matrix. They showed that numerical problems caused by ill-conditioning and the rank problem could be avoided by setting small eigenvalues to zero. An undesirable consequence of their approach is that analysis increments are restricted to the subspace spanned by the ensemble members.

The second consequence of estimating a large covariance matrix using a small sample is that sampling noise produces spurious correlations at large distances. Both the cut-off approach of Houtekamer and Mitchell (1998) and the massaging of eigenvalues employed by Evensen and van Leeuwen (1996) are effective in removing this noise.

The computational cost of the analysis is an important consideration for both the simplified Kalman filter and the ensemble Kalman filter. For the simplified Kalman filter, the additional cost in the analysis due to the increased

complexity of the background cost function is negligible. The cost of calculating the vectors $\{\mathbf{s}_k; k = 1 \ldots K\}$ and $\{\mathbf{z}_k; k = 1 \ldots K\}$, on the other hand, is large. The Hessian singular vector calculation has a nested loop structure. An outer loop contains one tangent linear and one adjoint integration of the model for the period $[0, T]$. The number of iterations of the outer loop determines the number of singular vectors which are calculated. In a typical configuration, around 60 iterations are required to determine 15 singular vectors. (Note, however, that the discrepancy between the number of iterations and the number of singular vectors reduces if more iterations are performed.)

For each iteration of the outer loop there is an inner loop of, typically, 30 iterations. Each iteration of the inner loop requires the calculation of one gradient of the 3D-Var cost function. The number of iterations performed in the inner loop is roughly half the number performed during a 3D-Var analysis. Thus, the computational cost of the inner loop is roughly half the cost of the minimization component of a 3D-Var analysis at T42 resolution. The cost of the inner iterations constitutes around 60% of the total cost of the Hessian singular vector calculation.

The computational cost of the ensemble Kalman filter depends critically on the method chosen to generate the initial perturbations for the ensemble. It is vital to the success of the method that the perturbations are chosen in such a way that the differences between the forecasts evolve into a representative sample of forecast error. The approach of Houtekamer and Mitchell (1998) is to perform an ensemble of independent analyses. That is, a separate analysis is performed for each forecast. The analyses differ in having perturbed first guess fields and observations. If these analyses were performed using 3D-Var at T42 resolution, it is likely that the computational cost of the ensemble Kalman filter would be roughly similar to that of the Hessian singular vector calculation.

Houtekamer and Mitchell (1998) describe the ensemble Kalman filter algorithm as "embarrassingly parallel". This makes it well suited to modern computers. The simplified Kalman filter is less obviously parallel. The main component of the algorithm is a version of the Lanczos algorithm, and is an essentially serial calculation. However, the main computational burden of the algorithm is divided between integrations of the tangent linear and adjoint models, and calculations of analysis gradients. These calculations are already well parallelized for moderate numbers of parallel processors. There are opportunities to increase this parallelism through the use of targeted final-time inner products, which would allow the parallel execution of several singular vector calculations, each targeted to a different region; and through the use of a block version of the Lanczos algorithm, which would allow more than one gradient calculation or model integration to be performed simultaneously. Modifications to the way in which the observation cost function is evaluated may also allow effective use of larger numbers of processors.

Experimental Results

Five experiments are discussed in this section. They are summarized in the following table. The column marked "T_{opt}" indicates the optimization time for the Hessian singular vector calculation. The column marked "Inner Product" describes the inner product at the final time of the singular vector calculation (i.e. the matrix $\mathbf{W}$ in equation 12).

Expt.	Dates	T_{opt}	Inner Product
A	16-30 April 1998	6 hours	$\langle .,. \rangle_{\mathbf{B}}$
B	16-22 April 1998	6 hours	$\langle .,. \rangle_{\mathbf{B}}$
C	28 Nov-15 Dec 1997	48 hours	Energy (targeted)
D	28 Nov-15 Dec 1997	48 hours	Energy (not targeted)
E	16-22 October 1997	48 hours	Energy (not targeted)

For all experiments, the initial-time inner product of the singular vector calculation was assumed to approximate the inverse of the covariance matrix of analysis error at the beginning of the 4d-Var assimilation window, which was 6 hours long. Consequently, the vectors $\{\mathbf{s}_k; k = 1 \ldots K\}$ which define the subspace for the simplified Kalman filter were produced by evolving the singular vectors for 6 hours.

The optimization time for experiments **A** and **B** was 6 hours. The vectors $\{\mathbf{s}_k; k = 1 \ldots K\}$ were therefore the singular vectors at the optimization time. To the extent that the initial-time inner product accurately represents the inverse of the

analysis error covariance matrix, and that model error may be neglected, the singular vectors at the optimization time are the eigenvectors of the forecast error covariance matrix with respect to the final-time inner product. For experiments **A** and **B**, the final-time inner product corresponds approximately to the inner product defined in equation 2. As a consequence, the subspace defined by the vectors $\{s_k; k = 1 \ldots K\}$ coincides approximately with the subspace spanned by the leading eigenvectors of the covariance matrix of background error in the space of the control variable, as suggested by Courtier (1993). However, the vectors are not exactly eigenvectors. This is because the final-time inner product uses a local projection operator (Buizza and Palmer, 1995) to disregard contributions to the inner product from perturbations south of 30°N. Without this restriction operator, many of the leading singular vectors would lie in the southern hemisphere, due to the relative sparsity of observations.

As discussed in the preceding section, one consequence of choosing a subspace spanned by eigenvectors of the covariance matrix of background error in the space of the control variable is that the elements of the cross-covariance matrix **F** are zero. The subspace for experiments **A** and **B** does not coincide exactly with that spanned by eigenvectors of the background error, so that some small cross-covariance should be present. However, it is suspected that the cross-covariance matrix may be rather inaccurately reconstructed in the simplified Kalman filter. So, to avoid possible numerical problems, the cross covariances were explicitly zeroed in experiments **A** and **B** by setting the coefficient α in equation 5 to zero.

Experiment **B** differed from **A** only in the inner product used at the initial time of the Hessian singular vector calculation. In experiment **A**, the initial inner product was the 3D-Var Hessian, whereas in **B** the Hessian was approximated by the inverse of a covariance matrix constructed by combining the background error correlation matrix with an estimate of the variances of analysis error. (The latter is produced routinely during the analysis as part of the background error variance calculation. For details, see Fisher and Courtier, 1995.) The approximation to the Hessian used for experiment **B** has the practical advantage of greatly reducing the computational cost of the singular vector calculation. It allows the calculation to be expressed as an ordinary eigenvalue problem, which may be solved using a standard Lanczos algorithm. In particular, the inner iterations, which normally constitute 60% of the computational cost of the calculation, are unnecessary. An additional benefit is that no observation processing is required.

The combination of an initial-time inner product defined by an approximation to the analysis error covariance matrix, and a final time inner product defined by an approximation to the covariance matrix of background error favours singular vectors which originate in areas of large analysis error variance and propagate to areas of small background error variance. Most of the singular vectors for experiments **A** and **B** originated on the west coasts of Europe and North America. Both experiments performed 60 outer iterations with 30 inner iterations for the singular vector calculations, resulting in typically 15 converged singular vectors.

Scores for experiments **A** and **B** are shown in Figure 1. Experiment **A** shows a clear positive impact during the later stages of the forecast. The experiment was continued for a further 8 days, during which time the positive impact demonstrated during the first 7 days of the experiment was maintained. (The magnitude of the impact on forecast scores averaged over all 15 days of the experiment was similar to that shown in figure 1.)

Experiment **B** did not improve forecast scores. The lack of a positive impact suggests that the approximations made in specifying the initial-time inner product of the singular vector calculation have a significant effect on the performance of the simplified Kalman filter. In particular, it appears necessary to account for the effect of observations on the correlations as well as the variances of analysis error.

Experiments **C**, **D** and **E** all used a final-time inner product which measured perturbation energy (see Buizza and Palmer, 1995, for details). A local projection operator was used to restrict the inner product to measure energy north of 30°N. In the case of experiment **C**, this area was further restricted to 20°W-20°E and 30°N-70°N. The initial-time inner product was defined by a 3D-Var Hessian using the reduced observation set described in an earlier section. Horizontal variation of the standard deviations of background error was suppressed for these experiments in the singular vector calculation and in the construction of the matrices E and F of the background cost function. This was done to ensure that the change of variable was strictly invertible.

Figure 2 shows anomaly correlation and RMS error scores for 500hPa geopotential over the northern hemisphere for experiment **E**. The simplified Kalman filter has a clear positive impact in this case. The singular vector calculation was run for 60 outer iterations and 20 inner iterations, producing approximately 10 singular vectors per analysis cycle.

A second experiment (**D**) with the same choice of optimization time and inner product as experiment **E**, but for a different period, had a neutral impact. (The configurations of the simplified Kalman filter were unfortunately not identical for experiments **D** and **E** as the number of inner iterations of the singular vector calculation was increased to 35 for the former experiment, resulting in typically 14 rather than 10 converged singular vectors.) The reason for the variation in the impact on forecast scores between these experiments is not currently understood. However, when the inner product at final time was restricted to the area 20°W-20°E and 30°N-70°N (experiment **C**), a small positive impact was again observed. Figure 3 shows scores for experiments **C** and **D**. One beneficial effect of targeting is that the number of singular vectors which are converged after 60 outer iterations increased to about 20 in experiment **C**.

Figure 4 summarizes the impact of the three successful simplified Kalman filter experiments, **E**, **C** and **A**. The plots show the differences in scores between individual forecasts and their control forecasts as crosses. Differences in anomaly correlation and RMS error for geopotential at 1000hPa, 500hPa and 200hPa for the northern hemisphere are plotted. The differences have been normalized by the standard deviation of the sample to allow score-differences for forecast ranges between one and ten days to be shown on the same plot. The solid curve on each plot shows the mean difference in score. This curve has also been normalised by an estimate of the standard deviation of mean score. In estimating this standard deviation, an attempt has been made to account for temporal correlation of score differences by fitting the score differences to a first order auto-correlation model:

$$d_i = \phi d_{i-1} + \varepsilon_i \tag{19}$$

where d_i is the i^{th} score difference, and where ε_i is an independent random variable. It is easily shown that, if N is the total number of score differences and σ^2 is the variance of the individual score differences, the variance of the mean score for this model is given by

$$\left(\frac{1}{N^2} \sum_{i=1}^{N} \sum_{j=1}^{N} \phi^{|i-j|} \right) \sigma^2 \tag{20}$$

In practice, it was found that the autocorrelation coefficient ϕ was little different from zero for the experiments presented here. Consequently, the standard deviations used to normalize the mean score differences shown in figure 4 were in most cases close to $\sigma/\sqrt{N}$.

The mean score differences typically show a negative impact during the first few days of the forecasts. However, the difference is in most cases only one standard deviation away from zero, and is not statistically significant. The positive impact at day 6 and 7, on the other hand is between 2 and 3 standard deviations away from zero, and is therefore significant with a null-hypothesis probability of between 5% and 1%. The individual score differences, shown by the crosses in figure 4 show that the positive impact results from a modest general improvement in the scores for several forecasts, rather than from a few well-forecast outliers.

Conclusions

The simplified Kalman filter presented in this paper represents one approach to reducing the computational expense of the Kalman filter to a level which makes it feasible for use in operational numerical weather prediction.

An important part of the design of the simplified Kalman filter is that it does not attempt to replace the current analysis system. Rather, it tries to extend the already successful ECMWF 4D-Var analysis. This has the practical advantage that the simplified Kalman filter automatically takes advantage of any improvements to the 4D-Var system. In the vast majority of directions in the phase space of the model in which the Hessian singular vectors provide no information about background error covariances, the simplified Kalman filter generates analysis increments which (except for the effects of non-linearity) are identical to those of 4D-Var. This is in marked contrast to current implementations of the ensemble Kalman filter, which are only able to correct the background in directions spanned by the ensemble members.

The results presented in this paper show an encouraging positive impact. This impact is statistically significant. However, it is rather small given the large computational cost of the calculations. The small impact, and its variability when

essentially the same configuration is evaluated for different periods, has made it difficult to demonstrate a clear advantage for any particular choice of final time inner product and optimization time in the singular vector calculation.

In the companion paper (Ehrendorfer, 1998), it is estimated that 100 singular vectors may account for about 50% of the variance of 2-day forecast error and 17% of the variance of analysis error. The proportion of the variance of background error accounted for by 100 singular vectors probably lies between these extremes. The experiments presented in this paper used between 10 and 20 singular vectors. *It is likely that a much larger number of vectors should be used to demonstrate a large impact on forecast scores.* This would require a more expensive singular vector calculation. Changes to the analysis system and the forecast model (for example, the addition of more model levels) are also likely to increase the computational cost of the simplified Kalman filter.

The effect on the performance of the simplified Kalman filter of using a restricted observation set during the singular vector calculation has yet to be evaluated. It is also not known at present whether the use of a 3D-Var Hessian in the singular vector calculation has a detrimental effect on the accuracy with which the inverse forecast error covariance matrix is calculated. For the experiments presented in this paper, it has been assumed that the 3D-Var Hessian represents an approximation to the 4D-Var Hessian at the beginning of the assimilation window. It may be argued that the mid-point of the assimilation window should have been chosen, in which case the covariance information provided by the 3D-Var Hessian should have been evolved for 3 hours rather than 6 hours. The growth rate for forecast error variance in the directions spanned by the leading singular vectors is large. Consequently, the period over which covariance information is evolved has a large impact on the diagnosed variance of background error for the evolved subspace. The ambiguity in the evolution period can best be resolved by using a 4D-Var Hessian in the singular vector calculation. It is planned to test this option. The singular vector calculation will be roughly a factor of 2 more expensive in this configuration.

The large number of approximations involved in the current implementation of the simplified Kalman filter makes it difficult to properly evaluate the system. In order to test the effects of these approximations, it is highly desirable that a "benchmark" experiment should be run, for which as many approximations as possible are removed. This would allow a proper evaluation of the individual effects of the various approximations, and would provide an indication of the full potential of the system for improving forecast scores. The benchmark experiment should calculate a significantly larger number of singular vectors than the experiments presented here. It should use the 4D-Var Hessian with a full set of observations and, if possible, at the same resolution as the analysis. This would require a singular vector calculation which is at least a factor of 10 more computationally expensive than those used in the experiments presented in this paper. To put this into perspective, each singular vector calculation would require roughly the same computer resources as a 12 day integration of the T639/L31 configuration of the forecast model.

Currently, the Hessian singular vector calculation executes for about 2 hours on 8 processors of the Fujitsu VPP700 computer. A significant increase in the computational cost would clearly require that the calculation run efficiently on a much larger number of processors. This is not the case at present for two reasons. First, the low resolution of the calculation results in vectors which are too short to make effective use of the vector arithmetic units of the VPP700. Second, there is a load imbalance in the calculation of the observation term of the analysis cost function. Both problems are being addressed.

There are several practical and theoretical questions to be answered before the optimal configuration of the simplified Kalman filter can be decided. Many of these questions are best answered in the context of the explicit low resolution extended Kalman filter developed by Ehrendorfer (1998). It is expected that the explicit Kalman filter will prove invaluable in the further development of a simplified Kalman filter at ECMWF.

References

Barkmeijer, J., M. van Gijzen and F. Bouttier 1998, Singular Vectors and Estimates of the Analysis Error Covariance Metric. Q. J. Roy. Meteorol. Soc., 124, 1695-1713

Courtier, P. 1993, Introduction to Numerical Weather Prediction Data Assimilation Methods. Proc. ECMWF Seminar on Developments in the Use of Satellite Data in Numerical Weather Prediction. 6-10 September 1993.

Daley, R. 1991, Atmospheric Data Analysis. Cambridge Atmospheric and Space Science Series (Cambridge University Press).

Evensen, G. 1994, Sequential data assimilation with a nonlinear quasi-geostrophic model using Monte-Carlo methods to forecast error statistics. J. Geophys. Res., 99 (C5), 10143-10162.

Evensen, G. and P.J. van Leeuwen, 1996, Assimilation of Geosat Altimeter Data for the Agulhas Current Using the Ensemble Kalman Filter with a Quasigeostrophic Model, Mon. Wea. Review, 124, 85-96.

Fisher, M. and P. Courtier 1995, Estimating the Covariance Matrices of Analysis and Forecast Error in Variational Data Assimilation, ECMWF Research Dept. Tech. Memo 220.

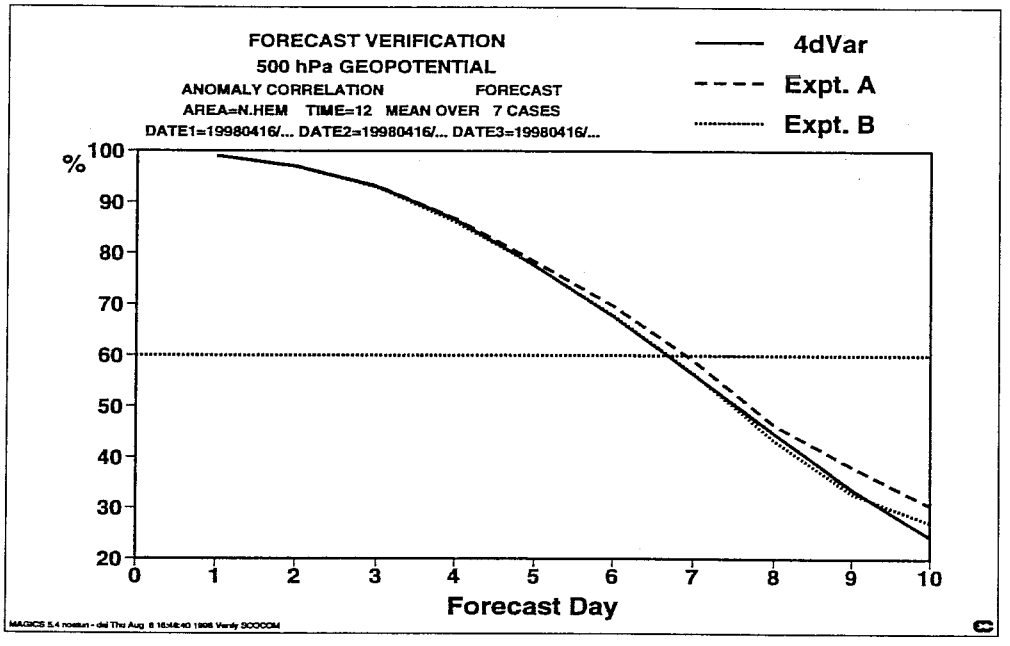
Jaszwinski, A.H. 1970, Stochastic Processes and FIltering Theory. Academic Press, New York.

Kalman, R.E. 1960, A New Approach for Linear Filtering and Prediction Problems. Trans ASME, Ser D, J Basic Eng, 82, 35-45.

Houtekamer, P.L. and H.L. Mitchell, 1998, Data Assimilation Using an Ensemble Kalman Filter Technique, Mon. Wea. Review, 126, 796-811.

Rabier, F., J-F. Mahfouf, M. Fisher, H. Järvinen, A.Simmons, E.Andersson, F. Boutier, P. Courtier, M. Hamrud, J. Haseler, A. Hollingsworth, L. Isaksen, E. Klinker, S. Saarinen, C. Temperton, J-N. Thépaut, P. Undén and D. Vasiljevic, 1997, Recent Experimentation on 4D-Var and First Results from a Simplified Kalman Filter, ECMWF Research Dept. Tech. Memo. 240.

Rabier, F. and P. Courtier, 1992, Four-dimensional Assimilation in the Presence of Baroclinic Instability, Q. J. Roy. Meteorol. Soc., 118, 649-672



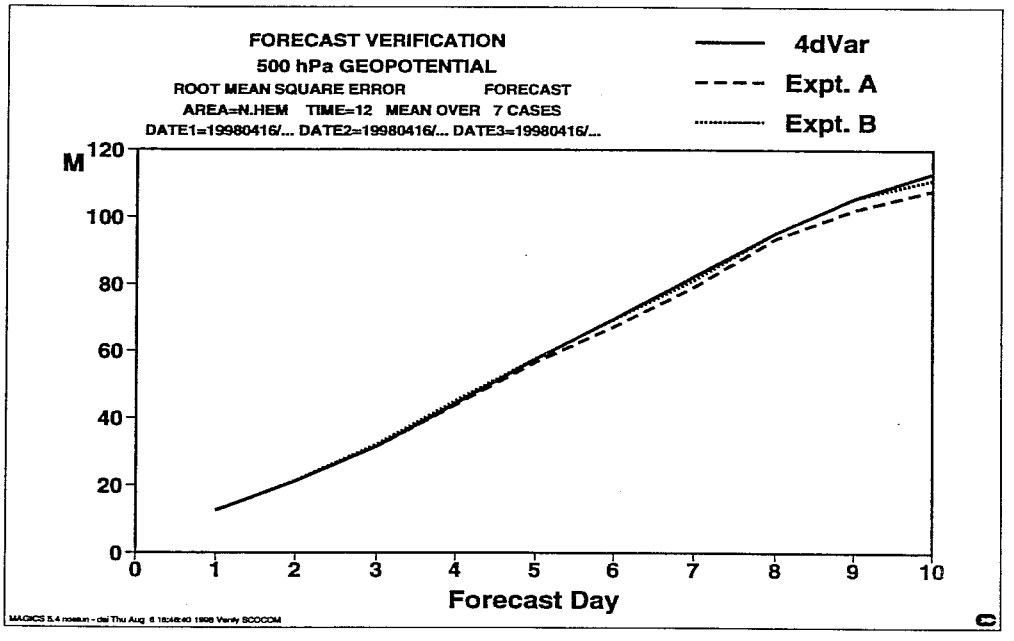


Figure 1: Anomaly correlation and RMS error for 500hPa geopotential for forecasts initialized from experiments **A** and **B**, and for a 4D-Var control

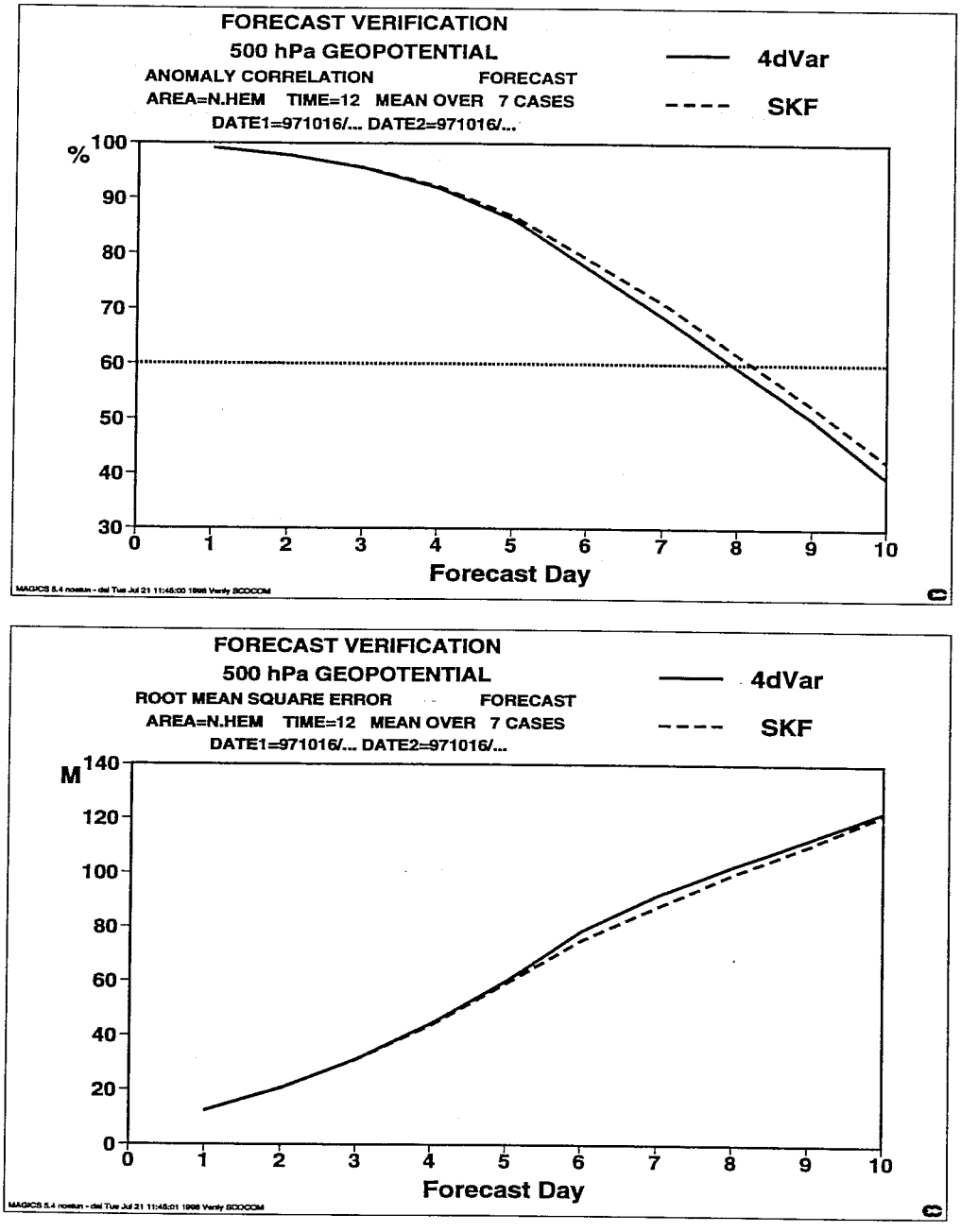
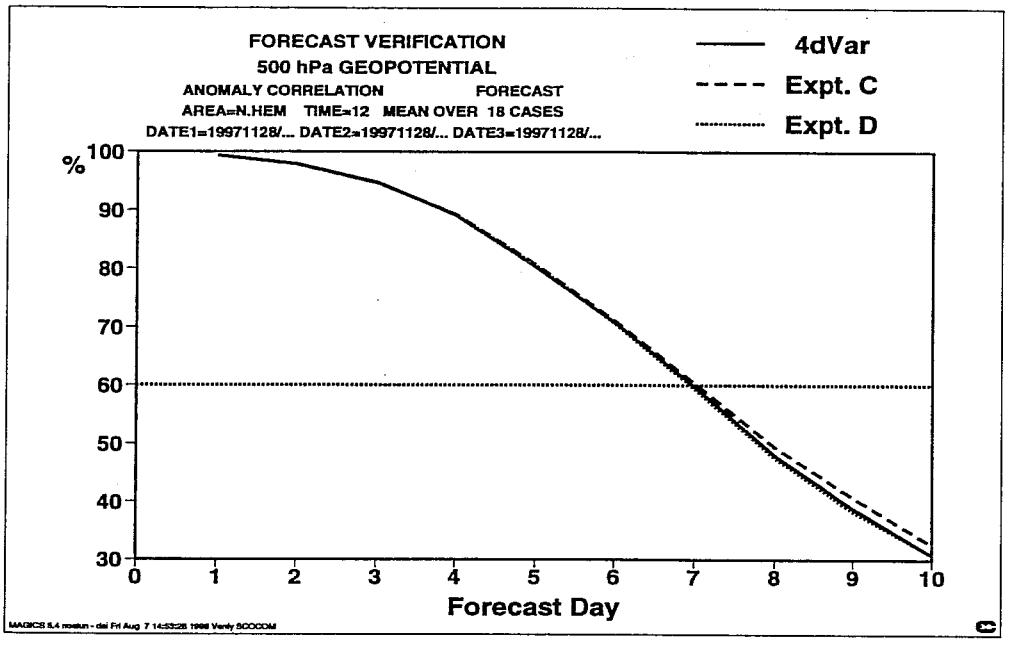


Figure 2: Anomaly correlation and RMS error for 500hPa geopotential for forecasts initialised from two analyses experiments. The dashed curves show scores for forecasts initialised using the simplified Kalman filter (experiment **E**). The solid curves are for 4D-Var.



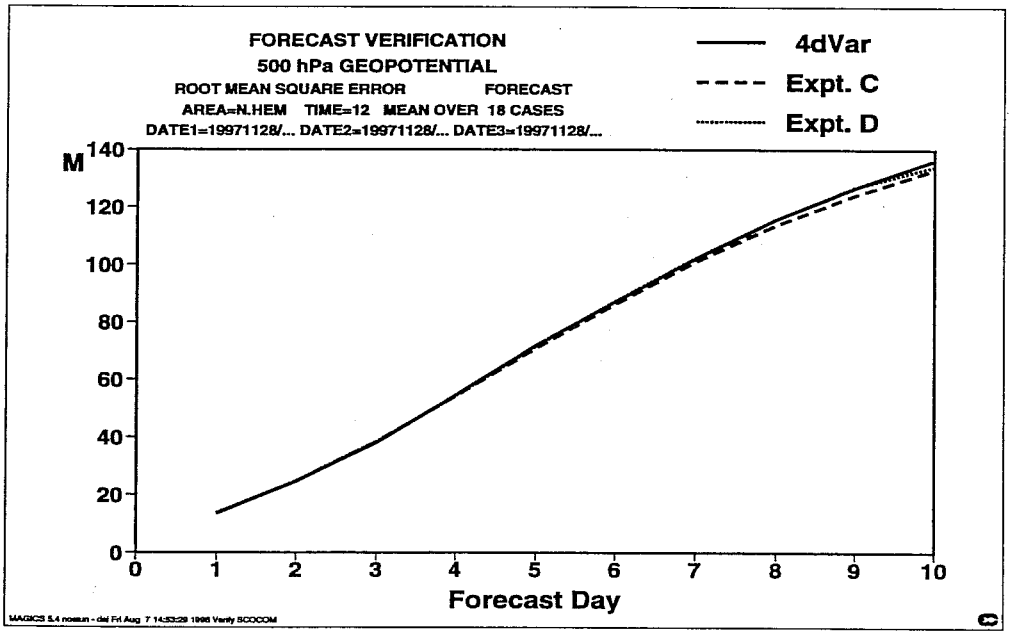


Figure 3: Anomaly correlation and RMS error for 500hPa geopotential for forecasts initialised from experiments **C** and **D**, and for a 4D-Var control.

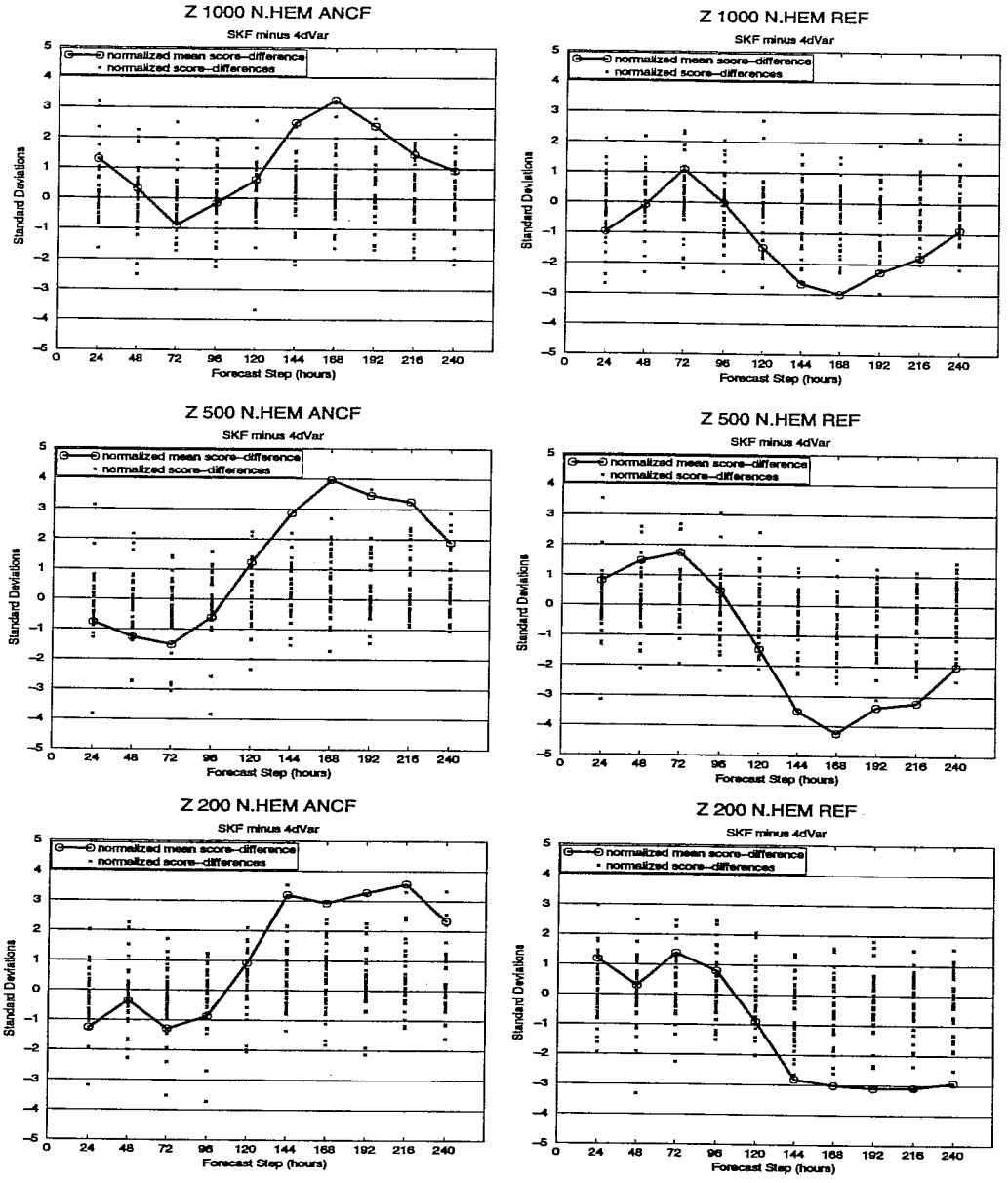


Figure 4: Individual and mean differences in anomaly correlation and RMS height score between forecasts from simplfied Kalman filter and 4D-Var analyses. The scores have been normalized by an estimate of the standard deviation of individual or mean score. In the case of mean scores, the estimated standard deviation includes a correction to account for the effect of autocorrelation.