347

Developments in 4D-Var and Kalman Filtering

Mike Fisher and Erik Andersson

Research Department

September 2001

For additional copies please contact

The Library
ECMWF
Shinfield Park
Reading, Berks RG2 9AX
library@ecmwf.int

Series: ECMWF Technical Memoranda

A full list of ECMWF Publications can be found on our web site under:
http://www.ecmwf.int/pressroom/publications.html

Developments in 4D-Var and Kalman Filtering

by Mike Fisher and Erik Andersson

Abstract

We discuss the status and the performance of the reduced rank Kalman filter (RRKF) as implemented within the framework of ECMWF's 4D-Var data assimilation system, as well as other new developments related to the specification and cycling of errors in 4D-Var. The presumption that the RRKF, through its incorporation of singular vector structures in the analysis, would lead to substantial forecast improvement has yet to be demonstrated. Extensive experimentation has taken place testing a variety of RRKF configurations - some at highest resolution affordable. Results show substantial forecast impact on a case to case basis in both the positive and negative directions, with near-neutral results on average over large samples. Careful definition of the unstable sub-space resolved by the RRKF and a better characterisation of the analysis error covariance have been identified as the key issues.

Other important developments of the 4D-Var system will enable further increase in analysis resolution, and prepare the ground for the use of future high density (and high frequency) satellite observations. A method for Hessian pre-conditioning is described. Recent re-evaluations of the background error covariance matrix are discussed and modifications to the background error formulation to allow a level of regional variation and flow dependence in the statistics are presented. We show that the definition of the static background error covariance matrix is crucially important for the performance of 4D-Var and also to the RRKF as it will influence its ability to accurately describe the location and structure of growing errors in the assimilation.

1. Introduction

Research on the predictability of synoptic-scale weather systems has identified the structures that amplify most rapidly during the early stages of a forecast. Errors in initial conditions that project substantially onto rapidly amplifying structures will quickly develop into forecast errors. Within the process of data assimilation it is particularly important to control the rapidly growing components of error as short-range forecasts are relied upon for the accurate propagation of information from one analysis time to the next. Methods have been developed that allow the computation of the fastest growing (or most unstable) modes of an atmospheric state, given a suitable definition of forecast error and forecast error growth over a pre-defined time interval. Singular vector techniques (Molteni and Palmer 1993; Buizza and Palmer 1995), adjoint sensitivity techniques (Rabier *et al.* 1996; Klinker *et al.* 1998) and ensemble techniques have been widely applied in the areas of ensemble prediction (Molteni *et al*. 1996) and observation targeting (Palmer *et al.* 1998) for example. In this paper we use these three techniques with the aim to improve the specification of background error covariances and to develop flow-dependent cycling of errors within the framework of ECMWF's operational 4D-Var system (Rabier *et al.* 2000; Mahfouf and Rabier 2000). In particular we will use singular vectors (SVs) to define a subspace of relatively small dimension for flow dependent propagation of errors with the reduced rank Kalman filter (RRKF) (Fisher 1998a).

The term "key analysis errors" was introduced by Klinker *et al.* (1998) to describe estimates, obtained through an iterative procedure based on the adjoint sensitivity technique, of the part of the analysis error that is largely responsible for the short-range forecast errors. When calculated with respect to a common norm (e.g. total energy) it is apparent that key analysis errors and SVs share many important characteristics: their amplitudes are maximum in the lower troposphere (around 750 hPa), they tilt backwards with height and they tend to be localized in the most baroclinically unstable regions. Gelaro *et al.* (1998) showed that the "key analysis errors" projects strongly on SVs, to the extent that a linear combination of the leading 30 SVs describes a large fraction of their variance.

The paradigm that a relatively small number of vectors can express a significant fraction of the most important part of analysis error has determined our approach to the development of an approximate, "reduced-rank" Kalman Filter (RRKF) (Fisher 1998a). The methodology was first outlined by Courtier (1993) and its subsequent development and testing has been ongoing for several years. Despite some early encouraging results based on small samples (Fisher 1998a; Rabier *et al.* 1997), the system has so far failed to live up to its initial promise. In part, this may be due to unrealistically optimistic expectations encouraged by the remarkable success of the "key analysis errors" in correcting medium-range forecast failures (Klinker *et al.* 1998; Rabier *et al.* 1996).

Barkmeijer *et al.* (1998; 1999) and others have demonstrated that the structure of SVs at initial time depends crucially on the norms used to measure error growth. The optimal choice in the context of data assimilation is to use as initial norm our best estimate of the analysis error covariance, $\mathbf{P}^a$. With such a choice the computed SVs will evolve, when propagated in time by the forecast model, to optimally span the evolution of the short-range forecast error covariance matrix. The twelve-hour evolved SVs (in the case of 12-hourly cycling) will then provide the optimal small-dimension basis (or subspace) for the construction of a flow-dependent prediction error covariance matrix, $\mathbf{P}^f$, to be used in the following 4D-Var cycle. It is presently not clear how good an estimate of $\mathbf{P}^a$ is required in order for the RRKF to produce flow-dependent SV-based estimates of $\mathbf{P}^f$ that perform significantly better than a static background error covariance ($\mathbf{B}$) in the context of an operational data assimilation system.

A second reason that a large positive impact was expected from the RRKF is that the superior performance of 4D-Var compared with 3D-Var was attributed to the dynamical evolution of the covariance matrix in the former system, compared with the static covariance matrix of the latter (Thépaut *et al.* 1993; 1996; Rabier *et al.* 2000). It has been argued that, since the covariance matrix in the RRKF is even more flow-dependent, we should expect it to give a correspondingly larger improvement in performance. Two results are presented in this paper which call this interpretation into doubt. The first result is a simple counter example to the hypothesis that covariance evolution necessarily explains the differences between 4D-Var and 3D-Var. The counter example (in Appendix D) is an extremely simple system for which covariance evolution does not occur, but for which 4D-Var is nevertheless demonstrably superior to 3D-Var. The second result is a 4D-Var analysis experiment for which the initial time of the 4D-Var assimilation window is displaced back in time by several hours. This is equivalent to replacing the static background error covariance matrix of the conventional 4D-Var analysis with a covariance matrix that has been dynamically evolved for several hours. The impact of this substitution on forecast skill is shown to be entirely neutral.

Despite the results mentioned above, we remain optimistic that a well-formulated approximate Kalman filter should produce a significant improvement to the accuracy of the analyses and the skill of the forecasts. However, it is clear that such an improvement will not be achieved without first improving our understanding of the problems involved. Current emphasis has therefore shifted away from operational implementation at a specific future date towards a more open-ended development strategy for a SV-based 4D-Var RRKF. The current paper provides results and discussions on what we have identified as the three main issues:

1. **The definition of the initial norm**. We investigate four different approximations for $\mathbf{P}^a$ as initial norm in the SV computations: total energy; the background error covariance matrix $\mathbf{B}$; a static approximation $\mathbf{A}$ obtained from an ensemble of assimilations; and the 4D-Var Hessian.

2. **The definition of the resolved subspace.** In a full Kalman Filter error covariances grow and evolve according to the model dynamics during the forecast step and they are reduced according to the Kalman gain-matrix during the analysis step. We have investigated to what extent this process takes place within the sub-space resolved by the RRKF. We shall see that this relates to the structural evolution of the SVs within the 12-hour time interval between analyses.
3. **The formulation of the background term.** The background term is crucially important for analysis performance (Andersson *et al.* 1998), so also in 4D-Var and RRKF. In RRKF the $\mathbf{B}$ matrix (the approximation to $\mathbf{P}^f$) used in 4D-Var will influence the analysis Hessian which in turn will influence the SVs (in the case that the Hessian is used as the SV initial norm). An ensemble-based approach has recently been adopted for the computation of $\mathbf{B}$. Regional inhomogeneity has been incorporated through a wavelet formulation. Vertical co-ordinate transformations are being developed which will make isentropes and/or the boundary layer height co-ordinate surfaces in the background term.

In the coming years further enhancements of the 4D-Var system are planned in preparation for cloud and rain assimilation and for the arrival of a large variety of satellite data, although this is not within the focus of the current paper. Some of the progress reported on here may nevertheless have a profound impact also on these other developments and pave the way to higher analysis resolution and the use of higher density satellite data.

The main part of this paper is devoted to new results and discussions, whereas mathematical details have been collected in a set of appendices towards the end of the paper. The outline of the paper is as follows: In Section 2 we summarize the current status with respect to the cycling of errors in the operational 4D-Var system, followed by a brief outline of the RRKF and the configuration in which it is usually run. In Section 3 we present result from extensive data assimilation and forecast experimentation testing several variations of the RRKF scheme. A critical re-examination of the importance of covariance propagation for the performance of 4D-Var is presented in Section 4, where our findings cast doubt on the generally accepted view that it is a dominant effect. New developments in the background term formulation are presented in Section 5 and their significance for the future progress of 4D-Var is discussed. Conclusions follow in Section 6. The appendices provide details on: An ensemble-based Kalman filter for the propagation of variances; Hessian-eigenvector preconditioning; The wavelet J_b-formulation; and “There is more to 4D-Var than covariance evolution!”.

2. Cycling of error covariances

For a linear system the Kalman Filter provides the formalism for optimal cycling of error covariances. The Kalman filter evolution of covariances may be divided into an analysis step (at time t) and a forecast step (from time t to $t+1$):

$$\mathbf{P}_t^a = (\mathbf{I} - \mathbf{K}\mathbf{H})\mathbf{P}_t^f \tag{1}$$

$$\mathbf{P}_{(t+1)}^f = \mathbf{M}\mathbf{A}_t\mathbf{M}^T + \mathbf{Q}_t \tag{2}$$

where $\mathbf{M}$ represents an integration of the linear forecast model over the interval $[t,t+1]$, $\mathbf{K} = \mathbf{P}_t^f\mathbf{H}^T(\mathbf{H}\mathbf{P}_t^f\mathbf{H}^t + \mathbf{R})^{-1}$ is the Kalman gain, $\mathbf{H}$ is the observation operator, $\mathbf{R}$ is the observation error covariance and $\mathbf{Q}$ is the model error covariance (following the notation of Ide *et al.* 1995).

2.1 4D-Var cycling

It is well known that the 4D-Var analysis at the end of the assimilation period is equivalent to a Kalman filter analysis over the same interval, given identical inputs - specifically that $\mathbf{P}_{(t=0)}^f = \mathbf{B}$ (the static background

error covariance) and $\mathbf{Q} = 0$ (the perfect model assumption). The 4D-Var algorithm itself does not provide an estimate of $\mathbf{P}^a_t$ and the prediction error covariance $\mathbf{P}^f_{(t+1)}$, required by the Kalman filter algorithm as input to the next analysis cycle, cannot be computed.

In the most straight forward practical implementations of 4D-Var $\mathbf{P}^f_{t=0}$ is replaced by the static $\mathbf{B}$ at every analysis cycle. In such implementations no cycling of covariance information takes place. The dynamical component of prediction error and the effects of variations in data distribution are both neglected. However, in ECMWF's operational 4D-Var data density and the seasonal variations in prediction error are taken approximately into account, using methods proposed by Fisher and Courtier (1995) and Fisher (1996). The analysis error covariance $\mathbf{P}^a_t$ is estimated using the combined Lanczos/conjugate-gradient method which finds approximately the leading eigenvectors of the 4D-Var Hessian and the associated eigenvalues. The leading Hessian eigenvectors describe the directions in control-vector space in which the information from the observations is most important. The simple error-growth model of Savijärvi (1995) is used to propagate the error variances to the time of the next cycle. This model represents exponential error-growth of small errors and the asymptotic behaviour of large errors towards a climatological variance (Fisher 1996). It lacks the dynamical i.e. flow-dependent effects on error growth.

A further refinement which allows flow-dependent cycling of prediction-error variances has been developed by Andersson and Fisher (1999). The method can be described as an ensemble-based Kalman filter, with the members randomly drawn from a population with covariance matrix $\mathbf{B}$. The ensemble is evolved to time $t+1$ by applying the tangent linear model $\mathbf{M}$ to each member of the ensemble. A brief description is given in Appendix A, for completeness. The method is affordable and could be implemented operationally as a future complement to the RRKF, or on its own. It is so far being used as a diagnostic tool to calculate the evolution of the effective background error variances within the 4D-Var assimilation period and also (as will be demonstrated later in this paper) to diagnose to what extent the effective background error variances have been modified by the RRKF. The RRKF was developed in an endeavour to cycle not only variances but also the dominant covariance structures, as explained in the following section.

2.2 A brief description of the RRKF

The ECMWF reduced rank Kalman filter (RRKF) is described by Fisher (1998) and by Rabier *et al*. (1997). We refer the reader to these papers for details of the algorithms used. Here we give a brief outline of the main features of the RRKF, and the configuration in which it is usually run.

From the point of view of the analysis, the RRKF consists of a modification to the background cost function of 4D-Var. In 4D-Var, the background cost function may be written as:

$$J_b = \frac{1}{2}(\mathbf{x}-\mathbf{x}_b)^T \mathbf{L}^{-T}\mathbf{L}^{-1}(\mathbf{x}-\mathbf{x}_b) = \frac{1}{2}\chi^T\chi \qquad (3)$$

where $\mathbf{L}^{-1}$ is the matrix representing the inverse change-of-variable which transforms model variables to the control variable χ of the minimization: $\chi = \mathbf{L}^{-1}(\mathbf{x}-\mathbf{x}_b)$. The background error covariance matrix used by 4D-Var is defined implicitly by the change-of-variable such that $\mathbf{B} = \mathbf{L}\mathbf{L}^T$ (and $\mathbf{B}^{-1} = \mathbf{L}^{-T}\mathbf{L}^{-1}$).

In RRKF the background cost function is modified to:

$$J_b = \frac{1}{2}\chi^T \mathbf{X}^T \begin{pmatrix} \mathbf{E} & \mathbf{F}^T \\ \mathbf{F} & \mathbf{I} \end{pmatrix} \mathbf{X}\chi \qquad (4)$$

Here, $\mathbf{X}$ is an orthogonal matrix (i.e. $\mathbf{X}^T\mathbf{X} = \mathbf{I}$) which represents a coordinate rotation such that the top left $K \times K$ submatrix of the innermost matrix corresponds to some chosen K-dimensional subspace. The matrix $\mathbf{E}$ has dimension $K \times K$, and determines the cost associated with background departures in the chosen subspace. The matrix $\mathbf{F}$ defines the cross-correlation between background departures in the subspace and those which are orthogonal to the subspace (with respect to the implied inner product). Note that if $\mathbf{E} = \mathbf{I}$ and $\mathbf{F} = \mathbf{0}$, then the background cost function of the RRKF is identical to that of 4D-Var.

The orthogonal matrix $\mathbf{X}$ is defined in practice by specifying a set of K vectors $\mathbf{s}_1, \mathbf{s}_2, \ldots, \mathbf{s}_K$ which span the required subspace. From these vectors, it is straightforward (and numerically stable) to construct $\mathbf{X}$ as a sequence of Householder transformations. The matrices $\mathbf{E}$ and $\mathbf{F}$ are then defined by specifying a second set of vectors $\mathbf{z}_1, \mathbf{z}_2, \ldots, \mathbf{z}_K$ such that $\mathbf{z}_i = (\mathbf{P}^f)^{-1}\mathbf{s}_i$, where $\mathbf{P}^f$ is the required flow-dependent covariance matrix of prediction error.

Usually, the vectors $\mathbf{s}_1, \mathbf{s}_2, \ldots, \mathbf{s}_K$ are chosen to be partially-evolved Hessian singular vectors. The reason for this choice is that if the vectors are evolved for a period equal to the cycling period of the analysis (typically 12 hours), then the vectors $\mathbf{z}_1, \mathbf{z}_2, \ldots, \mathbf{z}_K$ may be determined easily and cheaply during the singular vector calculation. (Note that this partial-evolution period is different from the optimization period, which is typically 48 hours.)

The Hessian singular vector calculation makes the assumption that the inverse of the Hessian of the analysis cost function (which is the *theoretical* analysis error covariance) provides an accurate characterisation of the *actual* analysis error covariance. This is an important assumption and is subject to the validity of major approximations within 4D-Var. Furthermore, the Hessian singular vector subspace is propagated using the tangent linear dynamics to give $\mathbf{P}^f$ (of rank K), without taking model error into account. It is subject to these approximations that the vectors $\mathbf{s}_1, \mathbf{s}_2, \ldots, \mathbf{s}_K$ and $\mathbf{z}_1, \mathbf{z}_2, \ldots, \mathbf{z}_K$ produced by the singular vector calculation satisfy $\mathbf{z}_i = (\mathbf{P}^f)^{-1}\mathbf{s}_i$. It is the validity of these approximations which effectively determines the ability of the RRKF to describe the likely evolution of the fastest growing components of forecast error.

2.3 In summary

In standard 4D-Var there is effectively no cycling of errors as the prediction error covariance matrix $\mathbf{P}^f$ at each analysis time is replaced by the static $\mathbf{B}$. In the theoretical Kalman filter, on the other hand, cycling is optimal, but a realistic model-error source term $\mathbf{Q}$ may be required to keep error variances at a realistic level. The RRKF incorporates some features of the Kalman filter into 4D-Var. Within the K-dimensional resolved subspace, the RRKF mimics the behaviour of the full Kalman filter provided that there is a substantial overlap between the subspaces of any two adjacent cycles. The $K \times K$-dimensional covariances will then essentially evolve according to Eq. (1) and Eq. (2). If there is little or no overlap then the resolved part of $\mathbf{P}^f$ will not evolve effectively with time. Covariances are then said to “leak” from the resolved subspace, requiring covariances to be replenished from the static $\mathbf{B}$ at each cycle. Outside the resolved subspace the RRKF can be expected to perform similarly to the standard 4D-Var.

3. RRKF Experimentation

Early results from experimentation with the reduced-rank Kalman filter were reported by Fisher (1998a) and by Rabier *et al.* (1997). These results were encouraging. However, they were also based on small samples. More recently, longer experiments have been possible - generally with less encouraging results.

3.1 Main forecast Results

Fig. 1 shows anomaly correlation of forecast error for 500hPa geopotential averaged over the Northern Hemisphere for a total of 131 days of RRKF data assimilation and forecast experimentation. The forecast scores are averaged over 5 periods: 14 January 1998 to 15 February 1998; 16 December 1998 to 16 January 1999; 1 August 1999 to 5 September 1999; 15 October 1999 to 5 November 1999; and 21 December 1999 to 28 December 1999.

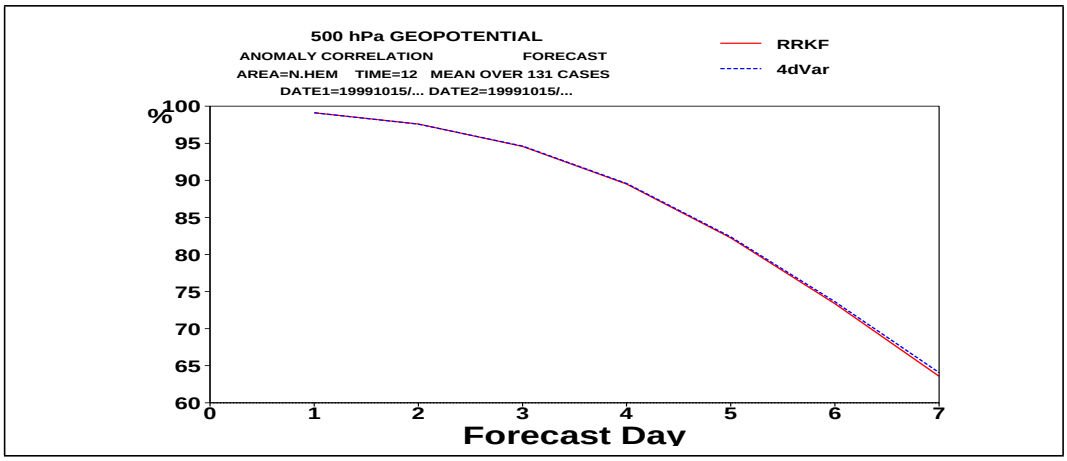

Fig 1: Northern Hemisphere forecast scores for the RRKF (red) and the 4D-Var control experiments (blue), averaged over 131 cases in 5 separate periods.

The control experiments were all 12h 4D-Var. The RRKF subspaces and covariance matrices were defined using Hessian singular vectors with a 48h optimization time, an energy inner product at final time, and a 4D-Var Hessian inner product at initial time. Singular vectors were calculated at T42 resolution and were targeted at final time to the Northern Hemisphere. At each analysis cycle 25 vectors were used to define the subspace.

Three of the experiments included in the sample were affected by an error in the specification of the Hessian used in the singular vector calculation. The effect of this error was to remove from the Hessian calculation all observations from the second half of the 12 hour window. In effect, the Hessian was that of 6h 4D-Var, rather than 12h 4D-Var. No systematic impact of this error on forecast scores could be detected. Fig. 2 shows the mean Northern Hemisphere 500hPa anomaly correlation averaged over the experiments (55 cases) which were not affected by the error.

The mean forecast score for the Northern Hemisphere for the RRKF experiments is nearly identical to that of the 4D-Var controls. Mean forecast scores for the RRKF are, however, marginally positive over the north Pacific (figure 3a). There is also small improvement at short range over North America (not shown), but this is not maintained into the longer range. Forecast scores for Europe (figure 3b) are less skilful for the RRKF than for the control 4D-Var experiments.

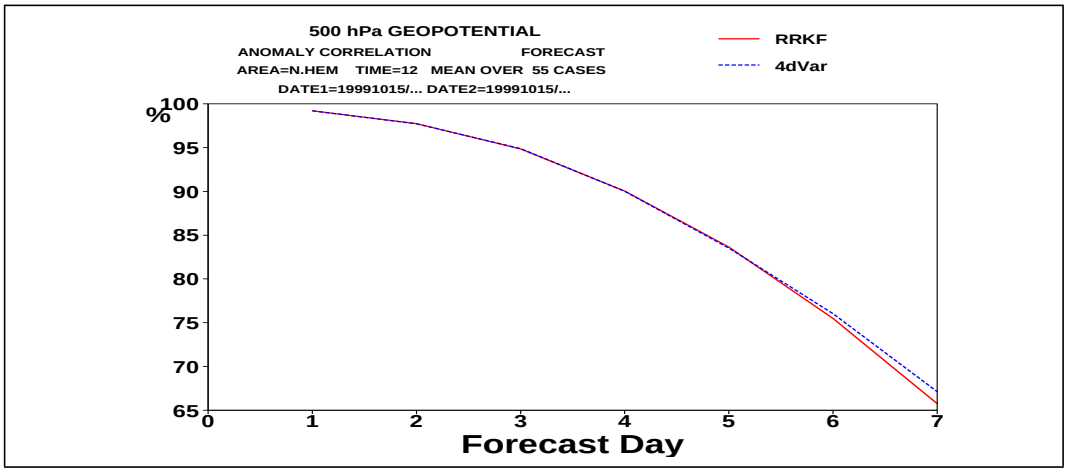

Fig 2: Northern Hemisphere forecast scores for the RRKF (red) and control (blue), averaged over 55 cases unaffected by an error in the specification of the Hessian inner product during the singular vector calculation.

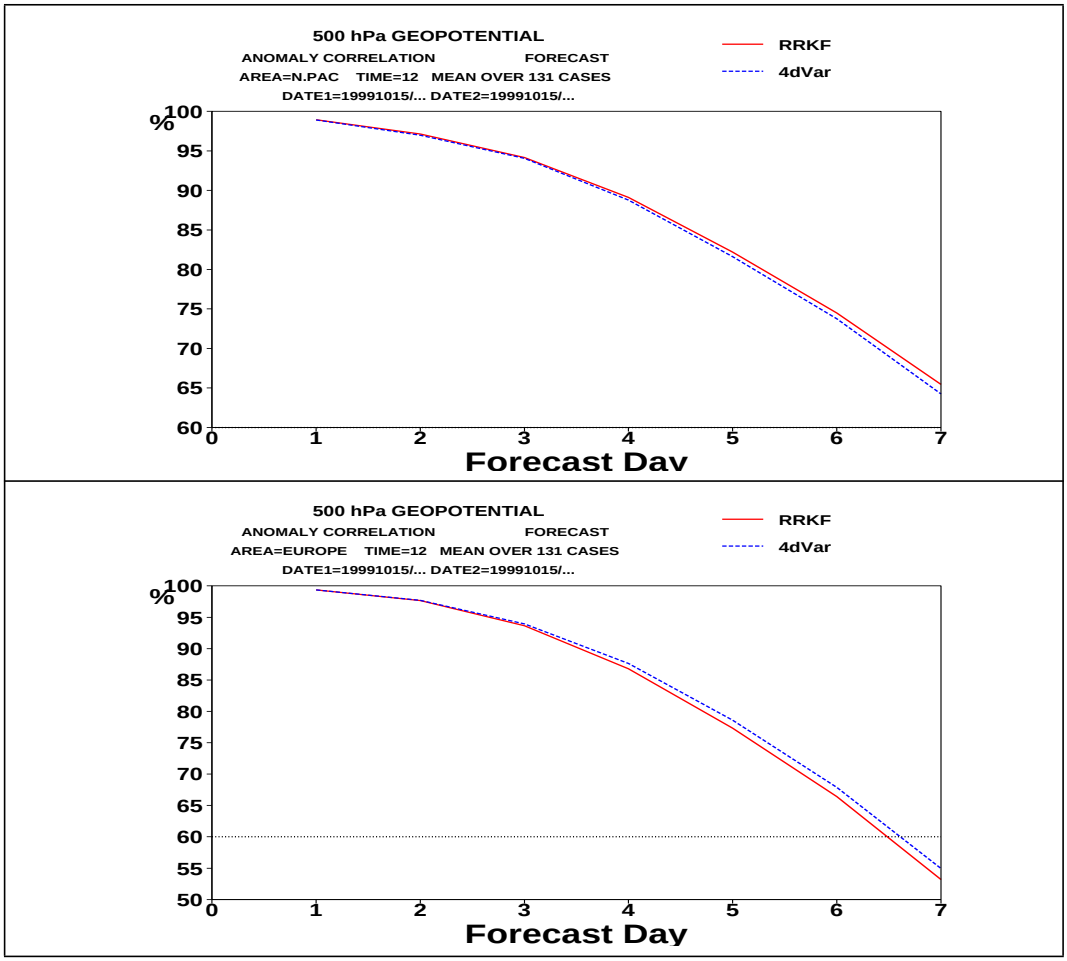

Fig 3: Forecast scores for the RRKF (red) and control (blue) for the North Pacific (top) and for Europe (lower panel), averaged over 131 cases.

Experiments were conducted for the 22-day period 15 October 1999 to 5 November 1999 to assess the impact of the dimension of the subspace used (i.e. the number of Hessian singular vectors which were calculated) and the optimization time for the singular vector calculation. This period was chosen because, with a 25-

dimensional subspace and 48 hour optimization time the RRKF appeared to have a slightly positive impact. The results are summarized in figures 4 and 5 for subspaces of dimension 5,10, 25 and 50 and for optimization

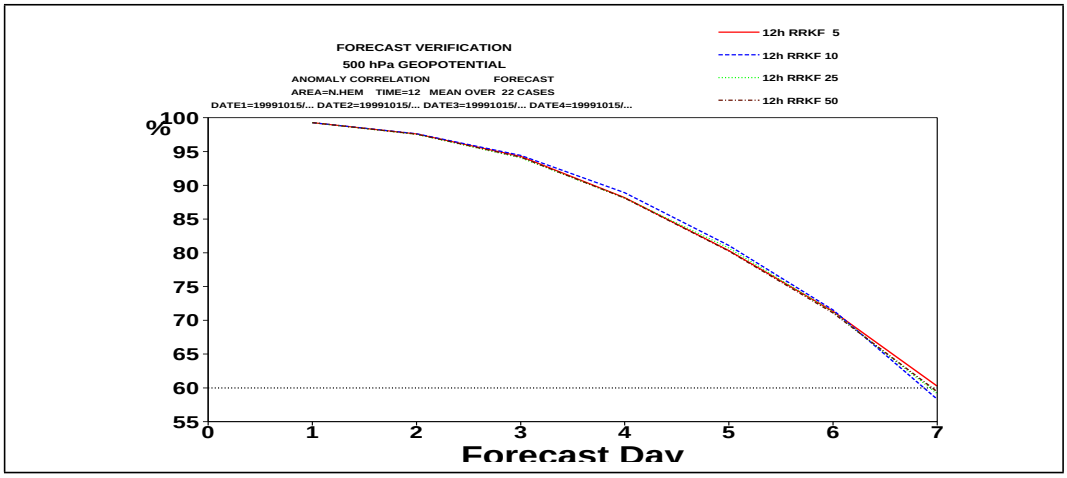

Fig 4: Northern Hemisphere forecast scores for the RRKF for different subspace dimensions (see legend). The singular vector optimization time for all cases is 48 hours.

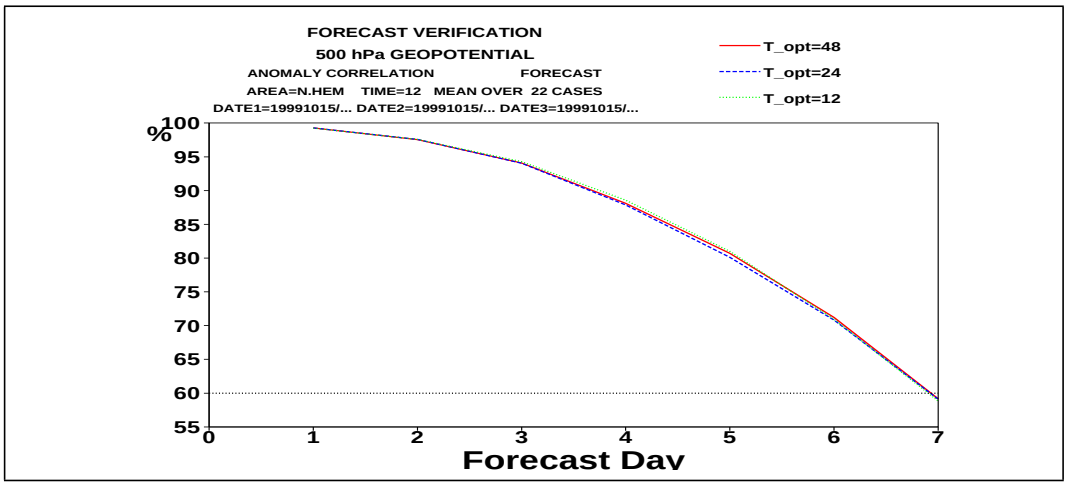

Fig 5: Northern Hemisphere forecast scores for the RRKF with different optimization periods (see legend) used in the Hessian singular vector calculation. The subspace dimension is 25 for all cases.

times of 12, 24 and 48 hours. There is no clear indication that changing either the dimension of the subspaces or the optimization time for the singular vector calculation has a significant impact on the forecast scores.

The effect of the analysis cycling period (i.e. the length of the data assimilation window) was also evaluated. Fig. 6 shows that the RRKF does not improve upon 6h 4D-Var. (The control in this case is the ECMWF operational system at the time.)

An experiment was run for which the subspace was defined by 200 Hessian singular vectors, calculated at T63 resolution (rather than the usual T42). Forecast scores for the RRKF are again similar to the control experiment, and are shown in figure 7. The increases in subspace dimension and in the resolution of the singular vectors were made computationally possible by replacing the analysis Hessian in the singular vector calculation by a static approximation to the covariance matrix of analysis error calculated from differences between contemporaneous analyses from an ensemble of data assimilation experiments.

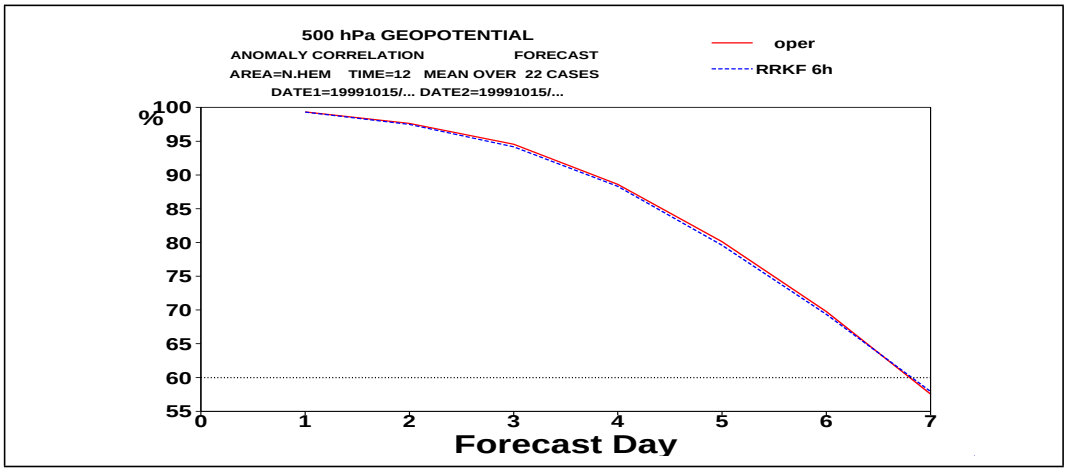

Fig 6: Northern Hemisphere forecast scores for the RRKF (blue) and control (red), both with 6-hourly cycling.

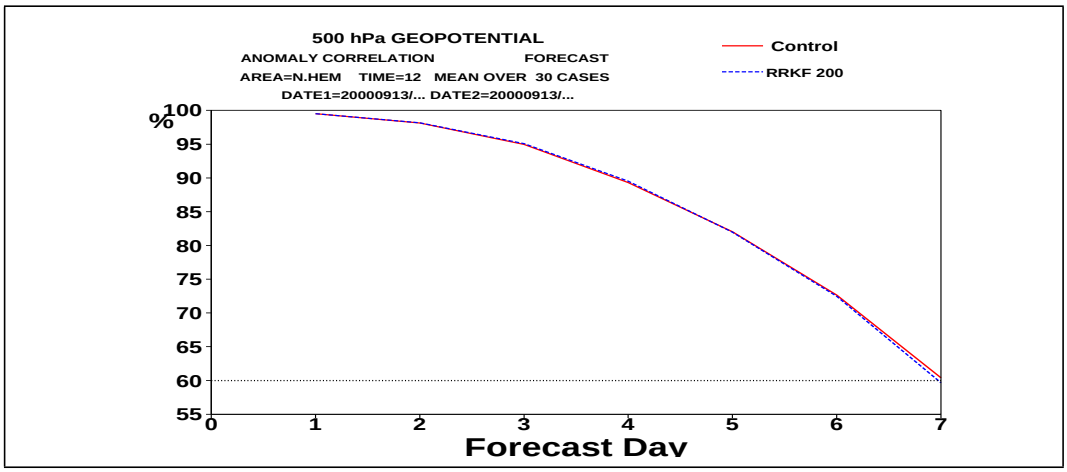

Fig 7: Northern Hemisphere forecast scores for an RRKF experiment using 200 Hessian singular vectors calculated at T63 resolution (blue) and control (red). A static approximation of analysis error covariance was used as singular-vector initial norm (instead of the Hessian).

Fig. 8 shows the rms amplitude of the temperature at model level 39 (approximately 500hPa) averaged over all 200 vectors for 3 UTC on 20 September 2000. Superimposed is the 500hPa geopotential height analysis for the same date. It is clear that the singular vectors tend to have amplitude in the dynamically unstable regions, and that all such regions in the extra-tropical Northern Hemisphere contain at least one singular vector.

Visual inspection of the Hessian singular vectors in the experiments reported on in this section revealed relatively vertical, nearly barotropic structures, in stark contrast to those obtained with an energy initial norm. This was also found to be the case in a study by Barkmeijer *et al.* (1999). However, more recent work has shown that using a more realistic background error covariance matrix produces a more baroclinic singular vector structure. The choice of RRKF subspace and the role of the initial norm will be further discussed in the following two sections.

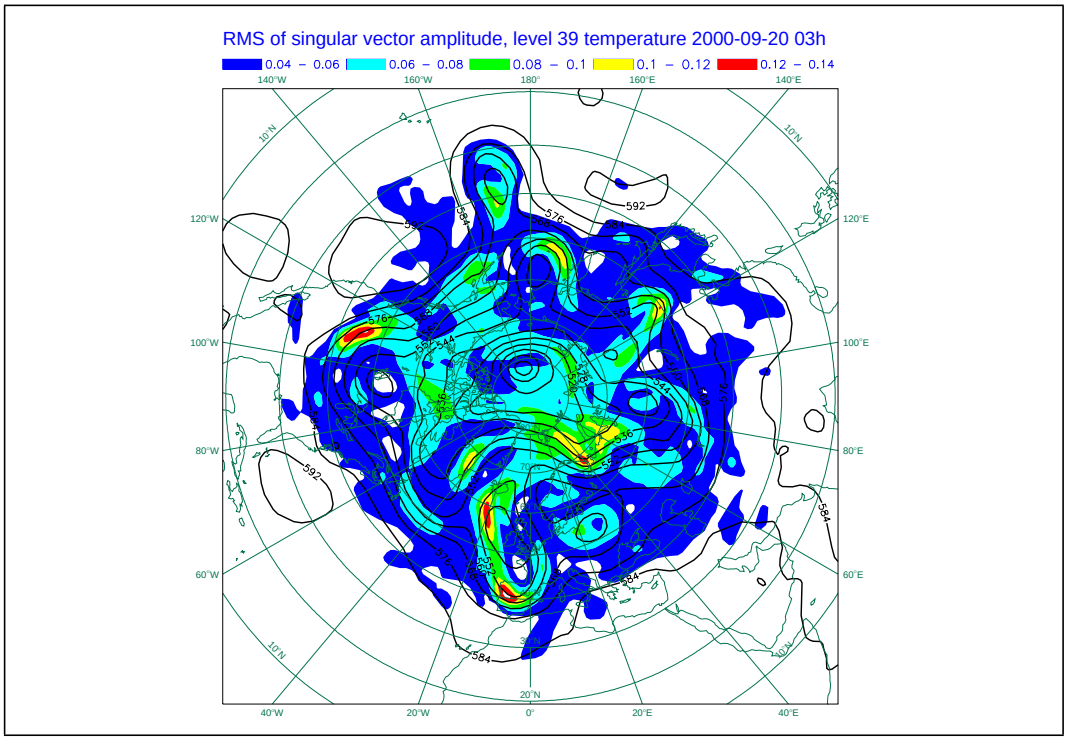

Fig 8: RMS amplitude of level 39 temperature for 200 singular vectors (shaded) using a static approximation of analysis error as initial norm, with the 500hPa geopotential height field (contoured) superimposed.

3.2 Use of Different Subspaces

Following the results reported above, a series of 3D-Var experiments was run to evaluate the possibility of replacing the Hessian singular vector subspace in the RRKF with unstable subspaces defined using other techniques. The RRKF relies on an implicit mechanism to propagate the covariance matrix for the subspace during the Hessian singular vector calculation. This mechanism cannot be used for other subspaces. Since the object of the experiments was to demonstrate sensitivity of the analysis to the choice of unstable subspace, no attempt was made to propagate the covariance matrix. Instead, the subspace covariance matrix was defined by setting $\mathbf{F} = \mathbf{0}$ and $\mathbf{E} = (1/\alpha^2)\mathbf{I}$ in Eq. (4). This corresponds to an inflation of the background error variance in the unstable directions by a factor of α^2. In most cases, a value of $\alpha^2 = 10$ was used.

To help identify the precise effect of the modified background cost function on the analysis, the background field for each cycle of each analysis was replaced by the corresponding background field of the control 3D-Var experiment. Thus the difference between each analysis and the control experiment was due entirely to differences in the background cost function, and not to differences in background fields, quality control decisions, *et cetera*, accumulated from earlier cycles.

Fig. 9 shows mean Northern Hemisphere forecast scores for the control experiment, and an experiment in which the subspace was defined by the leading 25 initial-time “energy” singular vectors (i.e. singular vectors calculated using an energy inner product at both initial and final time and with a 48 hour optimization time). The experiments were run for the period 15-24 October 1999. Also shown is an experiment in which the subspace was defined by the leading 25 48-hour evolved singular vectors. Both experiments show a remarkably small mean impact from modifying the background cost function. Mean forecast scores are nearly

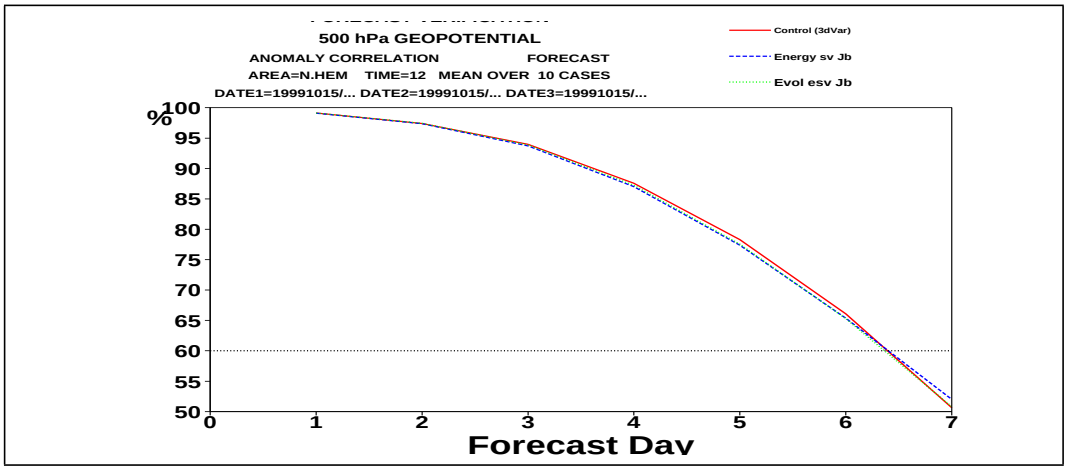

Fig 9: Northern Hemisphere forecast scores showing the effect of inflating background error in spaces defined by initial-time energy singular vectors (blue), evolved singular vectors (green) and the 3D-Var control (red).

as good as those of the control experiment, despite the fact that a significant modification to the background covariance matrix has been made in a highly unstable subspace.

The effect of the modified background cost function on the analysis is shown in figure 10 for the space defined

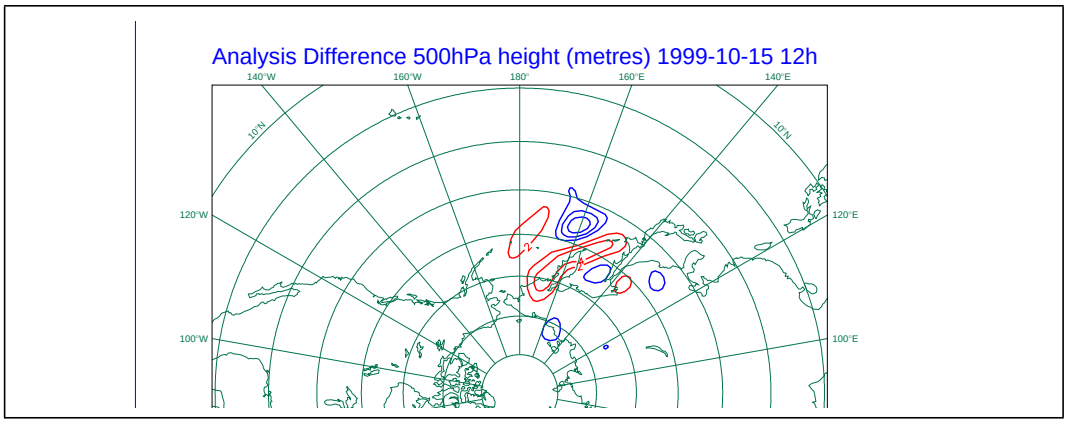

Fig 10: The effect on the 500hPa height analysis (19991015-12 UTC) of inflating background error variance by a factor of 10 in the space spanned by the leading 25 energy singular vectors. The contour interval is 2m. Negative analysis differences are shown in blue and positive in red.

by initial-time energy singular vectors. Geopotential heights at 500hPa are modified by several metres (i.e by large fractions of typical observation and background errors which both are less than 10 m). Fig 11 shows a cross section of the difference in analysed temperature for the same date taken along the line 50N, 135E to 40N, 175E. Clearly, the lack of impact on mean forecast scores is not simply due to a lack of impact on the analysis.

The forecast score for 1000hPa geopotential height over North America is shown for two consecutive dates in figure 12. The upper panel corresponds to forecasts run from the analyses whose difference is shown in figure 10. There is a positive impact of the analysis difference on the forecast score. By contrast, the lower panel shows a negative impact, for the subsequent date. It seems that the neutral mean forecast score for the period

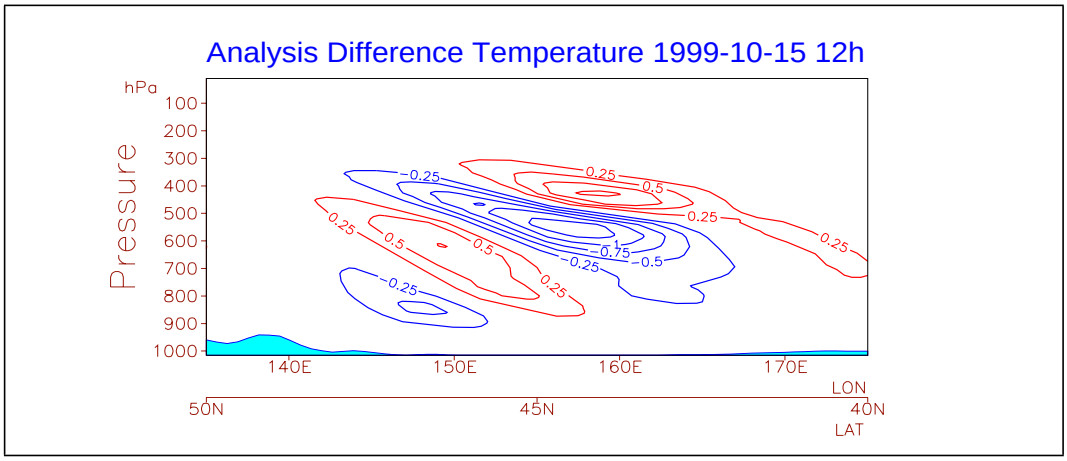

Fig 11: The effect on the temperature analysis of inflating background error variance by a factor of 10 in the space spanned by the leading 25 energy singular vectors, in a cross section through the feature shown in figure 10.

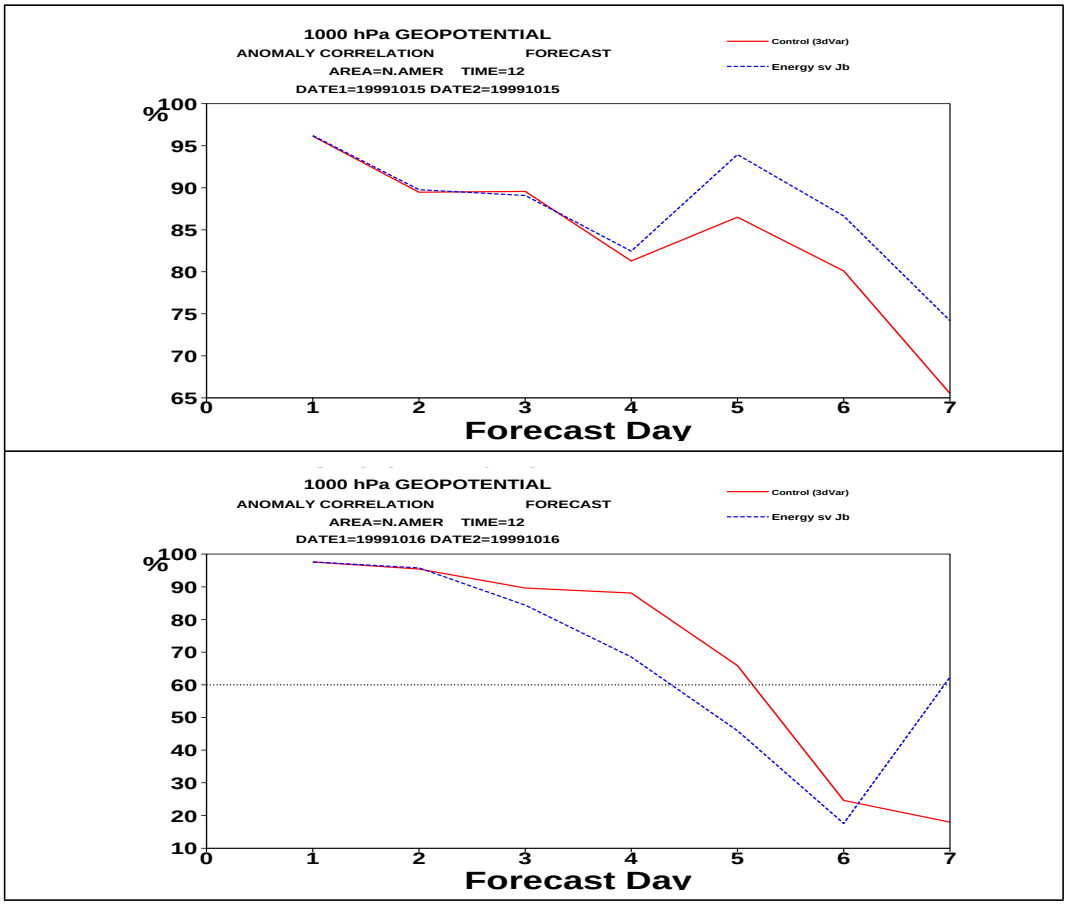

Fig 12: Forecast scores for 1000hPa geopotential height over North America for two consecutive days (19991015-12 UTC top, and 19991016-12 UTC bottom) for the experiment (blue) in which the background error variance was inflated in a subspace defined by initial-time singular vectors. The 3D-Var control is shown in red.

5-24 October 1999 results from a cancellation in the mean of significant positive and negative impacts for individual cases.

Two experiments using Hessian singular vectors were run for the same period. In one experiment, the subspace covariance matrix was generated during the singular vector calculation. This is the usual RRKF configuration. In the other experiment, the covariance matrix was defined using $\alpha^2 = 10$. The background fields for each analysis cycle were again taken from the control experiment. Forecast scores for both cases were very similar, and were slightly positive with respect to the control experiment. For the standard RRKF configuration, analysis differences for 500hPa geopotential height were somewhat larger than those obtained using energy singular vectors. They were larger still when the covariance matrix was defined using $\alpha^2 = 10$. Analysis differences for temperature were smaller than those obtained using energy singular vectors. This reflects the fact that Hessian singular vectors have a less baroclinic structure than initial-time energy singular vectors. A cross section of the difference in analysed temperature for the standard RRKF configuration is shown in figure 13 for the line 140E to 160W at 65N. (The cross section is taken along a different line from

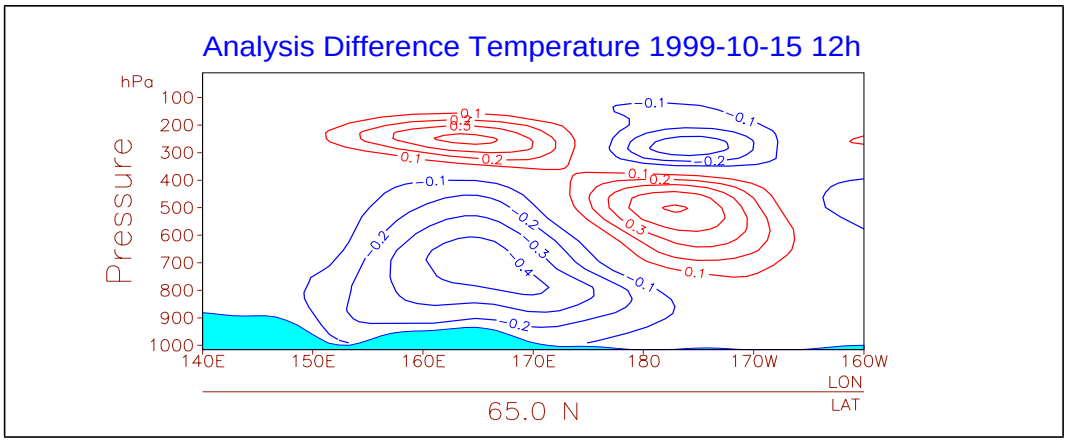

Fig 13: The effect on the temperature analysis of the standard RRKF, 19991015-12 UTC. The contour interval is 0.1 K with positive differences shown in red and negative blue.

that shown in figure 11 because the leading Hessian singular vectors were not in the same locations as the leading energy singular vectors.)

Fig. 13 shows that the structures provided by the Hessian-based RRKF are relatively barotropic, and broad in the vertical, with a change of sign at the tropopause. Futhermore, the Hessian singular vectors have a tendency to appear in the most data sparse areas: in the Arctic, northern Siberia and Central Pacific (Fig. 14, left panel) - with maximum amplitude at the tropopause level (Fig. 14, right panel). The figure shows very little evidence of singular vector amplitude in the storm-tracks in the North Pacific and the North Atlantic The reason is that Hessian-derived analysis error tends to be small where there are at least some observational data available.

To elucidate further on the reason why Hessian SVs are relatively less frequent in baroclinic regions of the Mid-latitudes (other than the central Pacific) error variances were cycled using Eq. (11) (Appendix A) in the case of a rapidly developing storm. A baroclinic wave had formed in the North Atlantic and was intensifying as it was approaching Ireland. Fig. 15 shows the resulting Hessian-based ensemble-estimate of prediction error at the time when the developing storm has reached the Irish Sea. The estimated error shows a local *minimum* at the location of the storm. The reason for this initially perplexing result can be understood through study of the leading Hessian eigenvectors. In this case they show a pattern which partly coincides with the structure of the storm and partly coincides with the data density distribution over north-west Europe. This illustrates that there is a strong dynamical influence on the Hessian-based cycling of errors which can sharply

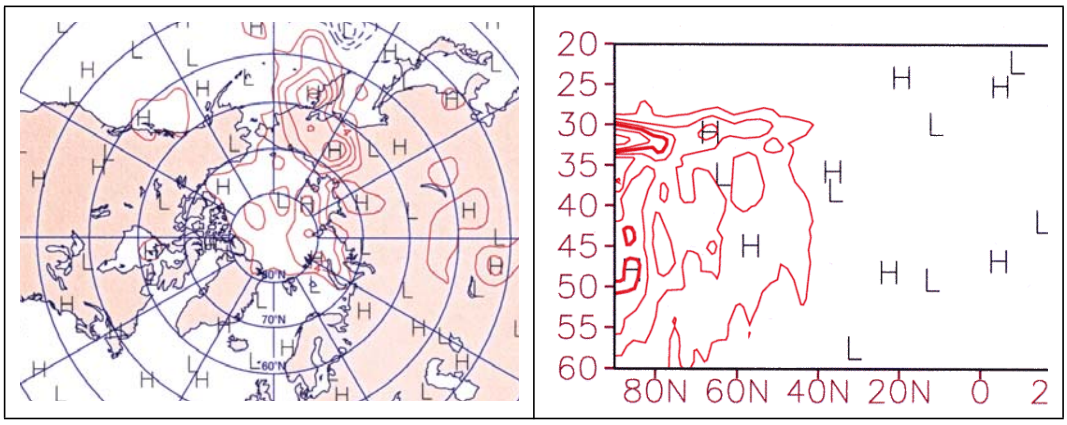

Fig 14: Change in effective background error standard deviations (diagnosed using the randomisation method, Appendix A). Polar map (left) and zonal mean cross section (right).

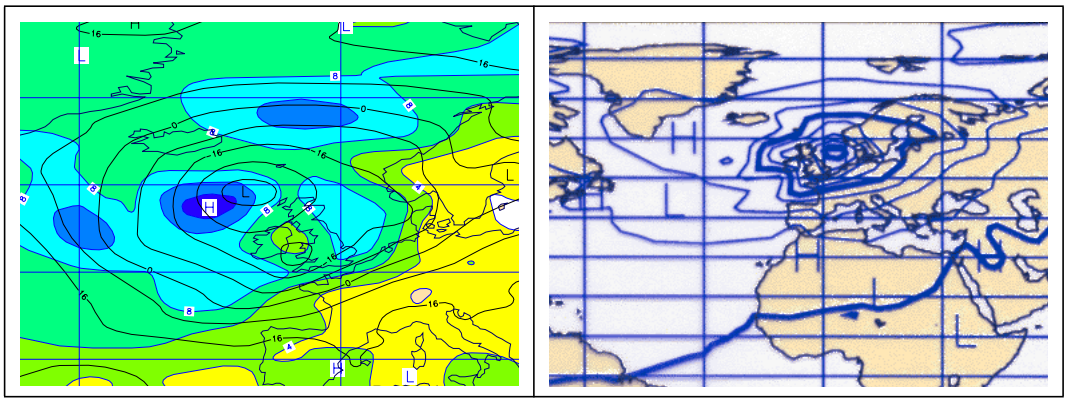

Fig 15: Flow dependent cycling of error variances for 20001029-15 UT +12h using the method in Appendix A. The panel on the left shows 1000 hPa geopotential (contoured) and the corresponding estimated 12-hour prediction error, colour shaded from yellow (2-4m) to blue (13-14m), in steps of 2m. The panel on the right shows the leading Hessian eigenvector, propagated by 12 hours.

decrease the analysis error of the dynamically most active features in the analysis, where there is good data coverage.

An experiment was conducted in which the subspace consisted of the single vector dubbed the “key analysis error” by Klinker *et al.* (1998). That is, the direction was defined by a truncated minimization of 2-day forecast error with respect to the initial conditions and an energy inner product. The RRKF analysis with this subspace was essentially unmodified compared with the control analysis. Analysis differences were large-scale patterns with an amplitude of a few tenths of a metre. This is in marked contrast to the experiments which defined the unstable subspace using singular vectors. A further experiment was run in which the variances were inflated by a factor of $\alpha^2 = 100$, and the number of iterations of minimization was doubled to compensate for the resulting degradation in the numerical conditioning of the minimization problem. This too had little impact on the analysis.

The inability of the analysis to draw in the direction of the “key analysis error” casts serious doubt on the interpretation of this pattern as an analysis error. However, it does not rule out the possibility that in localized regions the perturbation may coincide with analysis error, but with different amplitudes and signs in different geographical regions, and with the addition of decaying structures which may have little to do with analysis

error. The degree to which perturbations based on sensitivity calculations represent analysis error is an area which has received insufficient attention, but which has important implications for the RRKF. We intend to address this question in the near future.

3.3 Discussion

We have seen that the RRKF in its current form does not have a significant overall impact on forecast scores, on average over large samples. This is true both of the original Hessian-based formulation and the several variations described above. This is surprising because the RRKF specifically modifies the analysis in sensitive, unstable regions (as shown in Fig. 8 and Fig. 10). It is also contrary to the usual experience that changes to the formulation of the background error covariance matrix tend to have large effects on the accuracy of the analyses and on forecast skill.

The interpretation of the results is difficult. One possibility is that the actual analysis error has little projection on the unstable subspaces we have tried so far. For example, the actual analysis errors may initially have very small amplitude in the directions of the leading initial-time Hessian singular vectors, so that it is not until they have grown to several times their initial magnitude that they can be observed by the existing observing network. By this time, their structure will no longer correspond to that of the initial-time singular vectors.

The neutral result can also be interpreted as an indication that both the energy and the Hessian initial norms are poor approximations of the actual analysis error covariance, to the extent that the RRKF subspace is unsuccessful in describing a substantial part of the *likely* short-range forecast error evolution. Improvements in the characterisation of analysis error and alternatives to the standard singular vector approach will be explored in future work. We have seen that Hessian singular vectors tend to appear in those areas that are least well observed (the Arctic, northern Siberia and central North Pacific), i.e. where the estimated analysis error is large and approximately equal to the background error. The spatial structure of the Hessian singular vectors in those regions is therefore predominantly determined by the static $\mathbf{B}$, which favours isotropic and barotropic structures of certain horizontal and vertical scales. The RRKF to date has not been able to pin-point the most relevant (often small scale and tilted) components of analysis error which experience tells us may occur anywhere in the baroclinic areas at mid-latitudes. A further possible explanation is that covariance evolution is less important than we have hitherto supposed. This is the topic of the next section.

4. Covariance evolution

The significantly better performance of 4D-Var compared with 3D-Var is widely attributed to the implicit use in 4D-Var of an evolving, flow-dependent covariance matrix of background error. There is, however, an alternative explanation for the improvement, as will be explained in the following.

4.1 A first-order contribution to analysis error

Consider the 4D-Var cost function:

$$J = (\mathbf{x}-\mathbf{x}_b)^T\mathbf{B}^{-1}(\mathbf{x}-\mathbf{x}_b) + (\mathbf{y}-\mathbf{H}\mathbf{x})^T\mathbf{R}^{-1}(\mathbf{y}-\mathbf{H}\mathbf{x}) \quad (5)$$

where, for convenience of notation, the vector $\mathbf{y}$ is taken to include all the observations used in the analysis, and the observation operator $\mathbf{H}$ includes the model integrations required to propagate the initial model state $\mathbf{x}$ to the times of the observations. The analysis is given by setting the gradient of the cost function to zero:

$$\mathbf{x}_a = (\mathbf{B}^{-1} + \mathbf{H}^T\mathbf{R}^{-1}\mathbf{H})^{-1}(\mathbf{B}^{-1}\mathbf{x}_b + \mathbf{H}^T\mathbf{R}^{-1}\mathbf{y}) \tag{6}$$

Let us define the true state as $\mathbf{x}_*$, and the true values of the observed quantities as $\mathbf{y}_*$. We will assume that the observation error ε_o and the background error ε_b are unbiased, and seek an expression for the analysis error ε_a. Straightforward substitution into equation 6 gives the following:

$$\begin{aligned}\varepsilon_a = &-\mathbf{x}_* + (\mathbf{B}^{-1} + \mathbf{H}^T\mathbf{R}^{-1}\mathbf{H})^{-1}(\mathbf{B}^{-1}\mathbf{x}_* + \mathbf{H}^T\mathbf{R}^{-1}\mathbf{y}_*) \\ &+ (\mathbf{B}^{-1} + \mathbf{H}^T\mathbf{R}^{-1}\mathbf{H})^{-1}(\mathbf{B}^{-1}\varepsilon_b + \mathbf{H}^T\mathbf{R}^{-1}\varepsilon_o)\end{aligned} \tag{7}$$

Taking the expectation of equation 7, the last term vanishes, and we arrive after a little rearrangement at:

$$\langle\varepsilon_a\rangle = (\mathbf{B}^{-1} + \mathbf{H}^T\mathbf{R}^{-1}\mathbf{H})^{-1}\mathbf{H}^T\mathbf{R}^{-1}(\mathbf{y}_* - \mathbf{H}\mathbf{x}_*) \tag{8}$$

Both 3D-Var and 4D-Var tacitly assume that $\mathbf{y}_* = \mathbf{H}\mathbf{x}_*$, so that the expected analysis error is zero. However, this assumption is likely to be much more accurate in 4D-Var than in 3D-Var due to the inclusion in $\mathbf{H}$ of the propagation by the model of the initial state to the time of the observation. As a consequence, the expected error of a 4D-Var analysis is likely to be smaller than that of a 3D-Var analysis.

The presence in 3D-Var of a mean analysis error means that we cannot unequivocally assign the better performance of 4D-Var to its supposedly better covariance statistics. To further emphasize this point, we present in Appendix D a simple theoretical example for which the covariance matrices of analysis error for 3D-Var and 4D-Var are *identical*, but for which 4D-Var is nevertheless demonstrably more accurate than 3D-Var.

Bouttier (personal communication) noted that, for a linear model and observation operators, the mean analysis error given by equation 8 vanishes for the variant of 3D-Var known as 3D-FGAT. This is an incremental algorithm which replaces $\mathbf{Hx}$ in equation 5 by $H(\mathbf{x}_b)+\mathbf{H}(\mathbf{x}-\mathbf{x}_b)$, and retains the propagation of the initial state by the model to the time of the observations in H, but not in $\mathbf{H}$. 3D-FGAT has been shown to be superior to 3D-Var, and for this reason is being used for the ECMWF 40-year re-analysis (ERA-40). This suggests that elimination of mean error may indeed be an important factor in explaining the superiority of 4D-Var over 3D-Var.

The absence of a mean analysis error in 3D-FGAT does not imply that the better performance of 4D-Var is necessarily due to improved covariance statistics. To see this, we rewrite equation 6 in a form which applies to both 4D-Var and 3D-FGAT:

$$\mathbf{x}_a = \mathbf{x}_b + (\mathbf{B}^{-1} + \mathbf{H}^T\mathbf{R}^{-1}\mathbf{H})^{-1}\mathbf{H}^T\mathbf{R}^{-1}(\mathbf{y} - H(\mathbf{x}_b))$$

We see that, in 4D-Var, the analysis increment $(\mathbf{x}_a-\mathbf{x}_b)$ is determined by two separate flow-dependent effects. First, the scaled observation departure $\mathbf{R}^{-1}(\mathbf{y} - H(\mathbf{x}_b))$ is propagated back in time to the start of the analysis window by the action of $\mathbf{H}^T$. This propagated departure is then acted on by the flow-dependent analysis error covariance matrix, $\mathbf{P}^a = (\mathbf{B}^{-1} + \mathbf{H}^T\mathbf{R}^{-1}\mathbf{H})^{-1}$. In 3D-FGAT, neither of these flow-dependent effects occurs, since $\mathbf{H}$ does not contain the tangent linear model dynamics.

In general, it is difficult to determine which of the two flow-dependent effects is dominant in 4D-Var. However, in the example given in Appendix D it is trivial to separate them, since the example is explicitly constructed so that, even in 4D-Var, the analysis error covariance matrix is not flow-dependent. In this case, the demonstrable advantage of 4D-Var over 3D-FGAT comes purely from the action of $\mathbf{H}^{\mathrm{T}}$ in propagating the background departure to the start of the analysis window. In the next section we present an experiment which suggests that this is also the dominant flow-dependent effect in the full ECMWF 4D-Var analysis.

The improved performance, at LT511/T159 resolution (see Appendix B), of 12h 4D-Var compared with 6h 4D-Var provides a counter-argument to the hypothesis that covariance propagation may be relatively unimportant in 4D-Var. Once again, we may appeal to mean analysis error to explain the difference. The current observing network contains important classes of observations which report at 12 hourly intervals. This class includes large numbers of radiosondes. With a 6 hour analysis cycle, only alternate analyses contain these data. The intervening analyses do not. This leads to a 6 hour oscillation in the mean analysis error which results from biased observations. It was observed that this "flip-flop" effect was greatly reduced when 12h 4D-Var was introduced. Moreover, the largest impact of increasing the analysis cycle to 12 hours was over east Asia, where the "flip-flop" effect is large. It is entirely possible that a reduction of analysis bias is sufficient to explain the improved performance of 12h 4D-Var compared with 6h 4D-Var.

4.2 Extended window 4D-Var

To try to quantify the importance of covariance propagation in 4D-Var, we ran a 12h 4D-Var analysis experiment in which the initial time of the analysis window was moved back in time by 9 hours. We call this system "extended-window 4D-Var". No observations were assimilated during the initial 9 hours of each assimilation window, so that the observation cost function was identical to that of the usual 12h 4D-Var. Background fields were taken from the appropriate time step of the preceding cycle's 4D-Var analysis. Note that, since a 4D-Var analysis is a model trajectory, the background trajectory for extended-window 4D-Var is no less accurate than in a normal 12h 4D-Var analysis. (In practice, the 4D-Var analysis is not exactly a model trajectory due to the way in which the surface fields are analysed. Also, the low resolution trajectory which provides the linearization state for the tangent linear and adjoint models during the minimization may be somewhat less accurate than for a normal 12h 4D-Var. Neither of these effects are thought to have had a significant impact on the performance of the extended-window analysis system.)

At the start of the analysis window, the background error covariance matrix in a 4D-Var analysis is equal to the static covariance matrix specified in the background cost function. This covariance matrix is then implicitly propagated forward in time according to the tangent linear dynamics to generate flow-dependent "structure functions" at later times during the assimilation window. In the extended-window analysis, the background error covariance matrix is propagated over an additional 9 hours. Except for minor differences due to the surface analysis and the accuracy of the low resolution linearization trajectory, the additional covariance evolution is the only difference between extended-window 4D-Var and the conventional 4D-Var analysis. Fig. 16 illustrates extended-window 4D-Var schematically. Note in particular that covariance evolution effectively takes place in the entire control space. Thus, the extended-window analysis sidesteps questions about the choice of subspace to be propagated and the choice of inner product with which to define projection onto the subspace.

Three analysis experiments were conducted with the extended-window system. These corresponded to different specifications of the static covariance matrix at the start of the analysis window. In the first

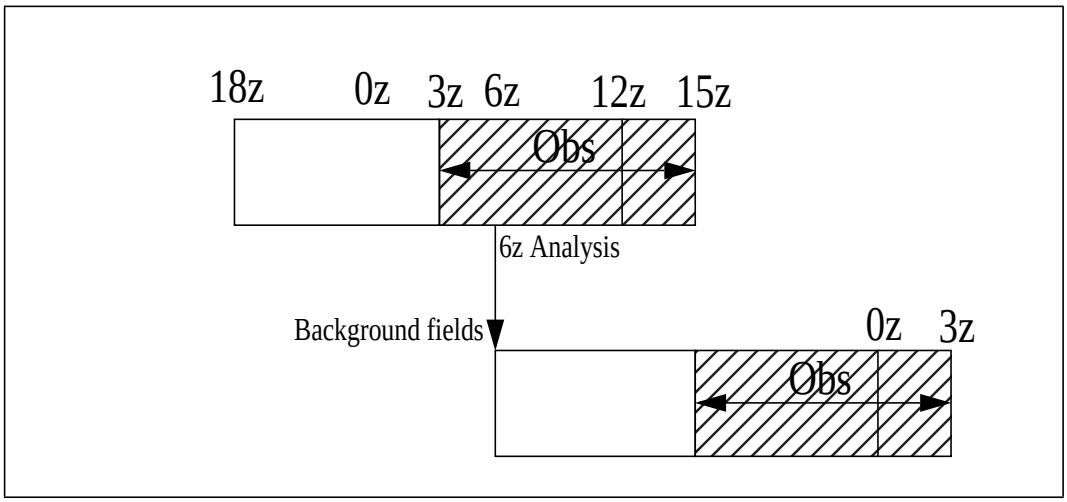

Fig 16: Schematic representation of extended window 4D-Var. The analysis cycle which produces the 0z analysis (bottom) takes its background fields from the 6z analysis of the preceding cycle (top). However, as in a normal 12h 4D-Var analysis, observations are assimilated only during the 12h period from 15z to 3z.

experiment, the **B** matrix was the same as is used in 12h 4D-Var. In the second experiment, the **B** matrix was multiplied by a factor of 0.84 to account approximately for the growth of forecast error variance over 9 hours. In the third experiment, the structure of the **B** matrix was kept the same, but the statistics were calculated from differences between analyses from an ensemble of analyses. In other words, the covariance matrix was a static model of analysis error. Mean forecast scores for these experiments are shown in Fig. 17. There is essentially no impact of extending the analysis window, which demonstrates that covariance evolution may not be the dominating effect that determines 4D-Var performance.

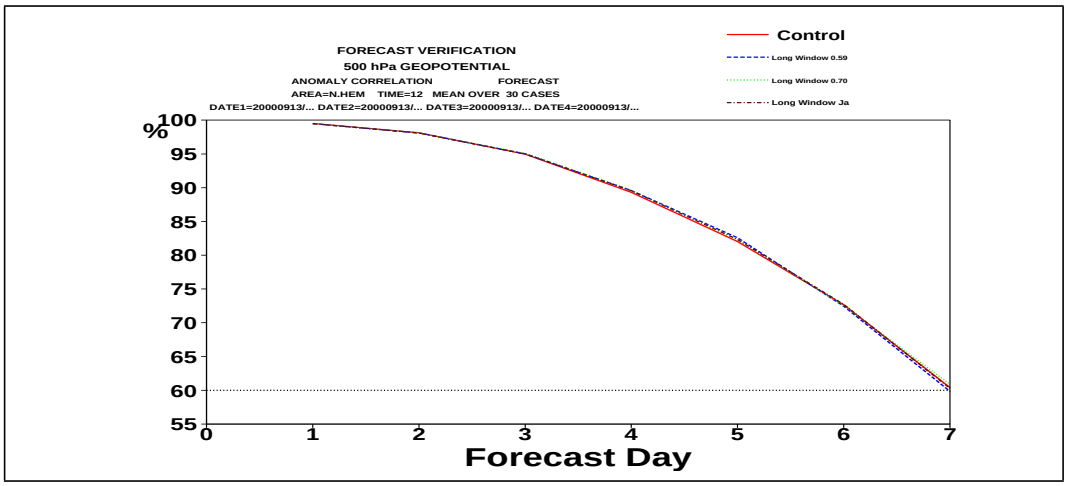

Fig 17: Northern Hemisphere forecast scores for extended-window 4D-Var.

5. Developments in 4D-Var

The discussions in Section 3.3 emphasised the importance of the static **B** matrix in its influence on the 4D-Var Hessian, and therefore also on the RRKF. Improvements to the J_b-formulation are of great significance to the performance of 4D-Var (e.g. Derber and Bouttier 1999), and its continued development has remained a priority.

5.1 Background error formulation

Variational background terms are commonly formulated in spectral space for reasons of computational efficiency. Isotropic and homogeneous covariances are spectrally represented simply by a diagonal **B** matrix. Non-separability between horizontal and vertical scales can also be incorporated relatively easily (Courtier *et al.* 1998). However, Andersson *et al.* (1998) found that the very significant advantages of non-separability (in 3D-Var) were largely offset by equally significant disadvantages due to the poor representation of the regional variations of background error statistics (in comparisons with an Optimum Interpolation scheme with a grid-point **B**). Current J_b -formulation (Derber and Bouttier 1999) allows some latitudinal variation in temperature (but not vorticity) background error statistics through its varying mass/wind coupling.

The wavelet-J_b described in Appendix C allows the advantages of non-separability to be combined with a degree of regional variation in the statistics. Fig. 18 shows the effective wavenumber-averaged vertical correlation matrix for vorticity background error implied by the wavelet J_b for points in North America and over the Equatorial Pacific. The differences between the two diagrams reflect differences in tropopause height and boundary layer depth in the two regions. Fig. 19 shows the horizontal correlation of background error for vorticity implied by the wavelet J_b for North America and for the Equatorial Pacific. The wavelet J_b produces significantly more large-scale correlations in the equatorial region (dashed line) than in North America (full line). The latitudinal variation in horizontal length scales is a prevalent feature in background error statistics (Ingleby 2001) which (for vorticity) has so far been neglected in ECMWF's 3D and 4D-Var.

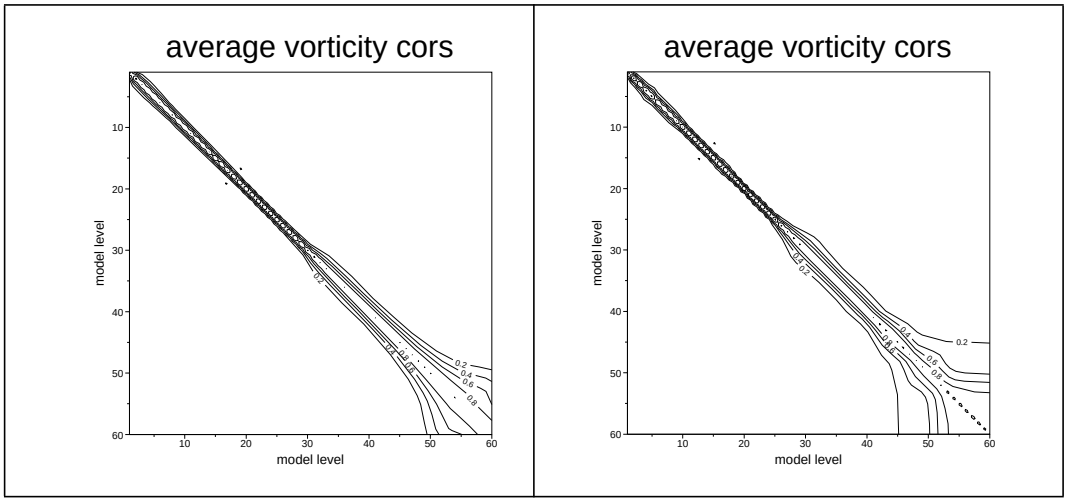

Fig 18: Effective wavenumber-averaged vertical correlation matrices for vorticity for wavelet J_b. The panel on the left shows North America, and Equatorial Pacific is on the right. Model level 30=202 hPa, 45=728 hPa and 50=884 hPa.

5.2 Calculation of J_b statistics

A new method for the calculation of background error statistics has been developed. It relies on an ensemble of data assimilation experiments, in which the members differ because of random noise added to the observations, in accordance with the assumed observation errors. A detailed description of the method with a compilation of results is currently in preparation. The main feature of the ensemble statistics is that vertical as well as horizontal length-scales are reduced, compared to statistics based on lagged forecast differences (the "NMC-method"). A **B** -matrix based on a 3D-Var ensemble was implemented in operations (version labelled 21r4) in October 1999. There was an important beneficial forecast impact associated with this change, as shown in Fig. 20. A second ensemble has recently been completed, this time using 4D-Var, with perturbations

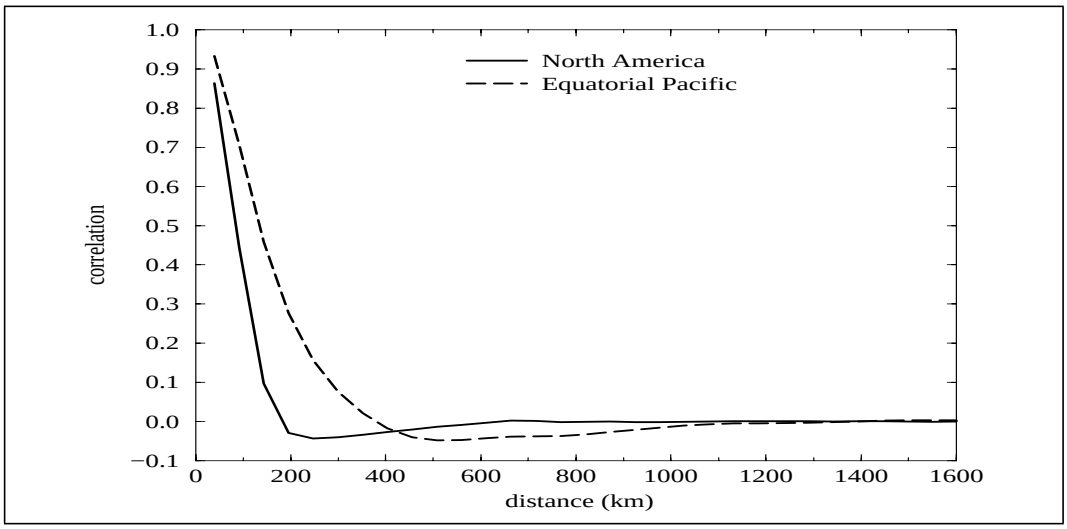

Fig 19: Effective horizontal correlations of vorticity background error implied by wavelet J_b *for model level 39 (~500hPa).*

added also to the data used in sea-surface temperature and soil-wetness analyses. A recent change to the vertical discretization of the forecast model and the recent increase in ATOVS data usage have also been incorporated in the new ensemble, with noticeable effects on the obtained background error statistics. Pre-operational testing is currently under way.

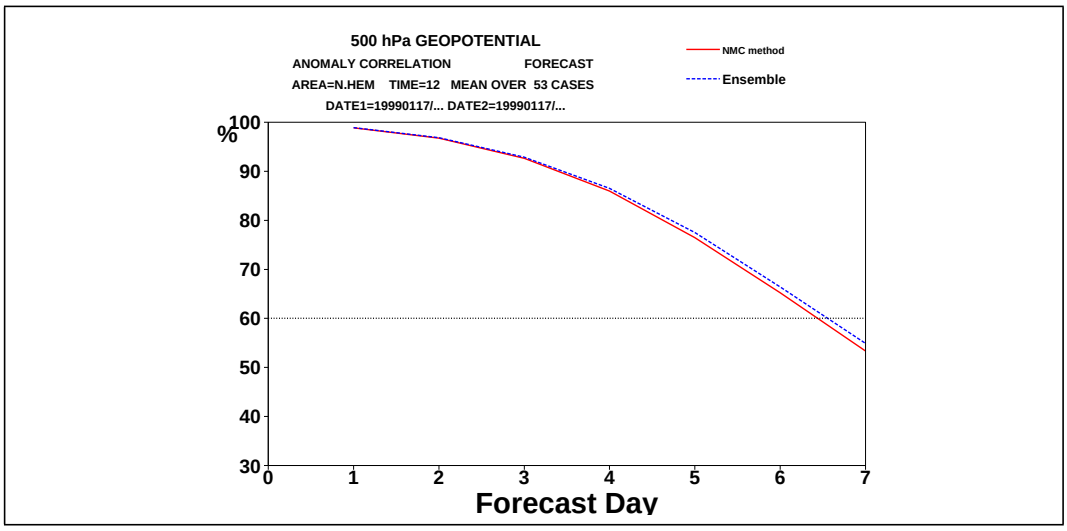

Fig 20: Northern Hemisphere forecast impact of the ensemble-based $\mathbf{B}$*-matrix introduced operationally in October 1999, averaged over 53 cases in three separate periods.*

5.3 Preconditioning

The iterative solution of the 4D-Var estimation problem has so far been preconditioned according to the background term. This is achieved through a variable transformation $\chi = \mathbf{L}^{-1}(\mathbf{x} - \mathbf{x}_b)$, with $\mathbf{L}$, defined such that $\mathbf{B} = \mathbf{L}\mathbf{L}^T$. A consequence of this choice is that the presence of very dense or particularly accurate observational data may deteriorate the conditioning and slow down the rate-of-convergence of the minimisation procedure. Andersson *et al*. (2000) investigated a case of poor convergence (which was found to be due to a combination of dense Meteosat radiance data and unrealistically large humidity background

errors) and derived an expression for the 4D-Var condition number as a function of data density and the background-to-observation error ratio.

More recently, in experiments with additional AMSU-A data, the data coverage in the Arctic stratosphere, where orbits overlap, became excessively dense so that conditioning, and thus the rate-of-convergence, were severely affected. A solution to these difficulties has now been provided through the Hessian eigenvector preconditioning presented in Appendix B. It is expected that this new feature will be highly relevant for the successful assimilation of future high-density satellite data.

The new preconditioning procedure makes the 4D-Var algorithm significantly more efficient, and is already benefiting the 40-year re-analysis (ERA-40). Its effect on the cost function and gradient reduction during the minimisation is illustrated in Fig. 21.

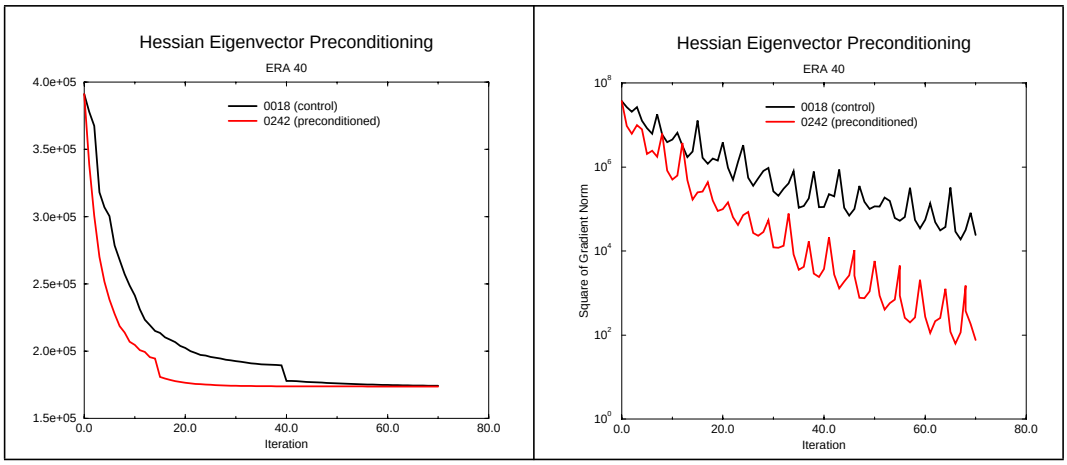

Fig 21: Cost function (left) and its gradient norm (right) as a function of the iteration count, during minimisation using the M1QN3 optimization algorithm (Gilbert & Lemarechal, 1989). The example is taken from the 40-year re-analysis (ERA-40) which uses 3D-Var. The benefit of Hessian eigenvector preconditioning (red lines) compared to J_b -preconditioning (black) is clear from the faster decrease in cost function and by the steeper reduction in gradient norm.

6. Conclusions

We have presented results showing that the RRKF, as currently formulated, has an entirely neutral impact on forecast scores. Moreover, this result is insensitive to the dimension of the resolved unstable subspace, and to changes in the subspace produced by varying the optimization time or the initial inner product used in the singular vector calculation. The neutral result can be interpreted as an indication that both the energy and the Hessian initial norms are poor approximations of the actual analysis error covariance, to the extent that the RRKF subspace is unsuccessful in describing a substantial part of the fast-growing short-range forecast errors in the assimilation. Attempts to use so-called “key analysis error” perturbations to define the analysis subspace cast doubt on the interpretation of these perturbations as analysis errors. Experiments using extended-window 4D-Var cast doubt on the conventional explanation that 4D-Var’s superior performance (relative to 3D-Var) results from its implicit dynamical propagation of error covariance.

6.1 Comments on the future development of 4D-Var

The work on preconditioning as well as RRKF make extensive use of the 4D-Var Hessian. Recent study of the leading Hessian eigenvectors has provided four results of strategic importance for the continued development of 4D-Var:

- **They are relatively large scale.** We can therefore be confident that further increases of inner-loop resolution can be achieved without deteriorating the conditioning of the problem. An inner-loop resolution of T399 is envisaged within the next four years, or so.
- **They reflect high data density.** Appropriate Hessian pre-conditioning will therefore be a required ingredient for the successful assimilation of future high-density satellite data in the coming years. Issues relating to the information content in new data types will require further study. This will include channel selection, data thinning and modelling of observation error correlations $\mathbf{R}$.
- **They are modulated by the tangent-linear physics.** It is therefore expected that further improvements in the linearization of physical processes (Janiskova, 2001) will be valuable for the data assimilation performance. More extensive use of the linearized physics within 4D-Var will be exploited, when the computational resources required become available in 2003.
- **They are influenced by the model dynamics** (through covariance propagation) and by the static background error covariances. This confirms that the J_b-term remains crucially important also with 12-hourly cycling. Continued development of the J_b formulation and relatively frequent re-calibrations of the statistics will take place.

6.2 The future of the RRKF

We have hypothesized that covariance evolution may be less important than expected in explaining the superiority of 4D-Var over 3D-Var. This does not necessarily imply that a well-formulated Kalman filter will not bring substantial improvements in the accuracy of the analysis (although it is clearly now a priority to quantify the potential benefits). Instead, it may indicate that we are attempting to propagate an approximate covariance matrix of analysis error which is such a poor approximation to the true covariance matrix that the propagated matrix is a no more realistic representation of forecast error covariance than the static $\mathbf{B}$ matrix. Pertinent to this suggestion is the question of the overlap of initial and partially evolved Hessian singular vector subspaces.

At each analysis cycle of the RRKF, the Hessian singular vector calculation implicitly propagates the projection of the analysis error covariance matrix onto initial-time singular vectors. The propagated covariance matrix valid 12 hours later forms the background error covariance matrix for the unstable subspace. This subspace is defined by the 12-hour-evolved singular vectors. Covariances of background error for directions orthogonal to the subspace are provided by the static $\mathbf{B}$ matrix.

Leutbecher has shown (personal communication) that the projection of 12-hour-evolved Hessian singular vectors onto initial-time singular vectors is rather small (less than 30%). This implies that much of the flow-dependent covariance information contained in the analysis Hessian at a given cycle of analysis is not propagated to the next cycle, since this information is known only for a space which is nearly orthogonal to the initial-time singular vectors. As a consequence, it is largely the static covariance information which is propagated. A recent paper by Reynolds *et al.* (2001) explains (their figure 16) that this lack of overlap is due to a phase difference between the evolved singular vectors (which propagate with the group velocity) and the initial-time singular vectors (which tend to follow the individual developing storm systems, i.e. the phase-

speed). Extended-window 4D-Var demonstrates that there is no benefit to forecast skill in propagating a static approximation to the covariance matrix.

Recently, a new type of subspace based on Hankel singular vectors, has been proposed by Farrell and Ioannou (2001a; 2001b). Their subspace balances, in an optimal way, the initial perturbations and the evolved responses of the forecast model. In effect, the projection of the initial covariance matrix of analysis error remains largely within the subspace as it evolves. By using this approach, we expect that the subspaces used at successive analysis cycles would overlap significantly, so that flow-dependent covariance information would be propagated from one cycle to the next.

We propose a two-pronged attack. The first objective will be to quantify, in a reasonably realistic environment, the benefits which should be expected from a Kalman filter. To do this, we will collaborate with Dr Ehrendorfer of the University of Vienna, who will compare full and reduced-rank Kalman filters with 4D-Var in a T21 3-level quasi-geostrophic system. Observations will be taken from a “truth run” of the quasi-geostrophic model, so that analysis errors may be quantified exactly. The second line of attack will be to evaluate the use of approximate Hankel singular vectors to define the subspace in which covariance information is evolved.

Acknowledgements
We are grateful to the members of the Data Assimilation section for their contributions towards the developments presented here, and to Anthony Hollingsworth, Adrian Simmons, Tim Palmer, Francois Bouttier, Jan Barkmeijer and Martin Leutbecher for in-depth discussions of the results. We thank Els Kooij-Connally for editing and formatting the manuscript.

Appendix A An ensemble-based Kalman filter for the propagation of variances

Following suggestions by Fisher and Courtier (1995) the analysis error covariance $\mathbf{P}^a$ is estimated using the combined Lanczos/conjugate gradient algorithm which finds approximately the leading eigenvectors v_k of the 4D-Var Hessian and the associated eigenvalues λ_k. The leading eigenvectors describe the directions in control-vector space in which the information from observations is most important. By applying the change of variable operator to each eigenvector, an estimate $\mathbf{A}$ of the analysis error covariance in model space is obtained:

$$\mathbf{A} = \mathbf{B} + \sum_{k=1}^{M} (\lambda_k^{-1} - 1)(\mathbf{L}v_k)(\mathbf{L}v_k)^T \tag{9}$$

where M is the number of computed eigenvectors. Only the variances, i.e. the diagonal elements of $\mathbf{A}$, are computed.

The randomisation method. A randomisation method can be used to calculate a low-rank estimate of $\mathbf{B}$, in terms of model variables (Fisher and Courtier 1995). In particular the diagonal of $\mathbf{B}$ can be estimated by

$$\mathbf{B} \approx \frac{1}{N} \sum_{i=1}^{N} (\mathbf{L}\xi_i)(\mathbf{L}\xi_i)^T \tag{10}$$

where ξ_i is a set of N random vectors in control-vector space, drawn from a population with zero mean and unit Gaussian variance. Variances produced by randomisation are somewhat noisy. The amplitude of the noise decreases as N is increased.

Propagation in time. The simple error growth model of Savijärvi (1995) used so far in 4D-Var represents exponential error growth of small errors and the asymptotic behaviour of large errors towards a climatological variance (Fisher 1996). It lacks the dynamical i.e. flow-dependent effects on error growth. From Kalman Filter theory (Eq. (2)) we have an expression for the evolution of the prediction error covariance matrix, $\mathbf{P}^f = \mathbf{MAM}^T + \mathbf{Q}$, where $\mathbf{M}$ is the tangent linear of the forecast model and $\mathbf{Q}$ is the model error covariance. Inserting the approximate forms for $\mathbf{A}$ and $\mathbf{B}$ from Eq. (9) and Eq. (10) into Eq. (2), we have:

$$\mathbf{P}^f \approx \frac{1}{N} \sum_{i=1}^{N} (\mathbf{ML}\xi_i)(\mathbf{ML}\xi_i)^T + \sum_{k=1}^{M} (\lambda_k^{-1} - 1)(\mathbf{ML}v_k)(\mathbf{ML}v_k)^T \tag{11}$$

Eq. (11) provides an expression for the evolution of error variances to any future time within the range of validity of the tangent linear approximation. In the current operational context around 90 v_k-vectors are computed. By setting $N = 50$ the additional cost is 90+50=140 12-hour integrations of the adiabatic tangent linear model $\mathbf{M}$, at low resolution (e.g. T_L95). It is hoped that this method could replace the current simple error-growth model and introduce the previously lacking flow-dependent effects on error growth. The viability of the method has been demonstrated by Andersson and Fisher (1999), where example illustrations can also be found.

Appendix B Hessian-Eigenvector Preconditioning

The exact rate of convergence of the minimization in 4D-Var depends in a complicated way on the details of the algorithm used and on the distribution of the eigenvalues of the Hessian matrix of the analysis cost function. Fisher (1998b) shows that for a quadratic cost function and conjugate gradient minimization, the following upper bound provides a good estimate of the actual convergence rate in the ECMWF analysis:

$$\left\|\mathbf{e}^{(n)}\right\|_{J''}^2 \leq 2\left(\frac{\sqrt{\kappa}-1}{\sqrt{\kappa}+1}\right)^n \left\|\mathbf{e}^{(0)}\right\|_{J''}^2. \tag{12}$$

Here, $\left\|\mathbf{e}^{(n)}\right\|_{J''}^2$ is a measure of the error remaining in the solution after n iterations of minimization, and κ is the condition number (i.e. the ratio of the largest to the smallest eigenvalue) of the Hessian. By expanding this expression as a power series in $1/\sqrt{\kappa}$, and truncating to first order, we arrive at an estimate of the number of iterations of minimization required to reduce the Hessian-norm of the error by a factor ε :

$$n \approx \frac{1}{2}\sqrt{\kappa}\ln\left(\frac{2}{\varepsilon}\right) \tag{13}$$

(Currently, the cost function in the ECMWF analysis is not quadratic, and the quasi-Newton minimization algorithm M1QN3 (Gilbert and Lemarechal, 1989) is used. As a consequence, the number of iterations required to achieve a given error reduction is roughly twice the estimate given above.)

Consider a general inner product $\langle ., . \rangle_{\mathbf{P}}$, defined by:

$$\langle \mathbf{p}, \mathbf{q} \rangle_{\mathbf{P}} = \mathbf{p}^{\mathrm{T}}\mathbf{P}\mathbf{q} \tag{14}$$

where $\mathbf{P}$ is a positive-definite symmetric matrix.

Expressing the analysis cost function as a Taylor expansion with respect to this inner product, we have:

$$J(\mathbf{x}_0 + \delta\mathbf{x}) = J(\mathbf{x}_0) + \langle \delta\mathbf{x}, \nabla_{\mathbf{P}} J \rangle_{\mathbf{P}} + \langle \delta\mathbf{x}, J''_{\mathbf{P}}\ \delta\mathbf{x} \rangle + O(\delta\mathbf{x}^3) \tag{15}$$

Both the gradient and the Hessian are dependent on the choice of inner product. In particular, the Hessian with respect to the P-inner product is related to the Hessian with respect to the Euclidean inner product, J''_2 , via:

$$J''_{\mathbf{P}} = \mathbf{P}^{-1} J''_2 \tag{16}$$

The quasi-Newton minimization algorithm M1QN3 allows the user to precondition the minimization by specifying the inner product to be used during the minimization. The optimal choice is $\mathbf{P} = J''_2$, since the Hessian with respect to the P-inner product is then the identity matrix, which has the smallest possible condition number of one. However, it is not possible to use the Hessian matrix itself as a preconditioner, since the minimization algorithm requires that $\mathbf{P}$ is easily inverted. We therefore choose an approximation to the Hessian.

Fisher and Courtier (1995) show that an approximation to the Hessian of the cost function may be constructed from its leading eigenvalues and eigenvectors, λ_k and $\mathbf{v}_k$. This approximation is already used in the analysis

system to estimate the variances of analysis error. The leading eigenvalues and eigenvectors are determined using a Lanczos algorithm (Lanczos 1950). Moreover, since the leading eigenvectors are large-scale patterns, they may be determined accurately, but cheaply, at low horizontal resolution.

Hessian eigenvector preconditioning (Fisher and Courtier, 1995) defines the minimization inner product as:

$$\mathbf{P} = \mathbf{I} + \sum_{k=1}^{K} (\mu_k - 1)\mathbf{v}_k\mathbf{v}_k^T \tag{17}$$

Here, $\mathbf{v}_k$ are the leading eigenvectors of J_2'' . The coefficients μ_k have yet to be defined.

The inverse of $\mathbf{P}$ is given by replacing μ_k by $1/\mu_k$ in equation 17. (This is a consequence of the orthonormality of the eigenvectors.) So, substituting for $\mathbf{P}^{-1}$ in equation 16, and replacing J_2'' by its full eigen-decomposition (with eigenvalues arranged in descending order) we have:

$$J_{\mathbf{P}}'' = \left(\mathbf{I} + \sum_{k=1}^{K}\left(\frac{1}{\mu_k} - 1\right)\mathbf{v}_k\mathbf{v}_k^T\right)\left(\sum_{k=1}^{N}\lambda_k\mathbf{v}_k\mathbf{v}_k^T\right) \tag{18}$$

Using the orthonormality of the eigenvectors, this may be written as:

$$J_{\mathbf{P}}'' = \sum_{k=1}^{K}\left(\frac{\lambda_k}{\mu_k}\right)\mathbf{v}_k\mathbf{v}_k^T + \sum_{k=K+1}^{N}\lambda_k\mathbf{v}_k\mathbf{v}_k^T \tag{19}$$

That is, the preconditioned Hessian $J_{\mathbf{P}}''$ has the same eigenvectors as the un-preconditioned Hessian J_2'' , but the leading K eigenvalues are reduced by factors $\mu_1 \ldots \mu_K$. By choosing these factors so that $\lambda_k/\mu_k < \lambda_{K+1}$, we produce a preconditioning which reduces the condition number of the Hessian by a factor λ_{K+1}/λ_1 . (Note that the preconditioning effect is relatively insensitive to the precise choice of the parameters $\mu_1 \ldots \mu_K$, provided that they respect the bound $\mu_k > \lambda_k/\lambda_K$. However, too-large values of $\mu_1 \ldots \mu_K$ may make the computations which use the preconditioner ill-conditioned and subject to excessive rounding error.)

Clearly, the effectiveness of the preconditioner increases monotonically with the number of vectors used. However, this must be offset against the computational cost of calculating, manipulating and storing the eigenvectors. The eigenvalue spectrum of the Hessian for the ECMWF analysis decreases rapidly for the first few eigenvalues, so that two or three eigenvectors are sufficient to reduce the condition number by a factor of two. However, the spectrum flattens out, so that typically around 25 vectors will reduce the condition number by a factor of six or seven. In the ECMWF system 25 vectors represents a reasonable compromise between effective preconditioning and the additional cost of manipulating the preconditioner. The factor of six or seven decrease in the condition number corresponds to a decrease by a factor of roughly 2.5 in the number of iterations of minimization required to achieve a given level of accuracy in the minimization. This has also been demonstrated in practice as shown in Fig. 21.

Appendix C Wavelet J_b Formulation

It is well known (Phillips 1986, Bartello and Mitchell 1992) that the vertical and horizontal scales of background error covariance are non-separable: large horizontal scales tend to have deeper vertical correlations than small horizontal scales. It is essential to retain this property in the covariance model for background error in order to achieve a correct description of the covariance structures for both wind and temperature (Rabier *et al.* 1998; Andersson *et al.*, 1998). Courtier *et al.* (1998), Derber and Bouttier (1999) achieve a non-separable model of background error covariance matrix by specifying different vertical correlation matrices for each total spherical wavenumber n. However, for variables which are unmodified by the balance operator (in particular, for vorticity), the covariance model is isotropic and homogeneous.

It is also well known that horizontal and vertical correlations vary geographically (Lönnberg 1988). Horizontal scales tend to be broader in the tropics than at high latitudes, as a consequence of atmospheric dynamics (Ingleby 2001). Correlation scales have also been shown to be influenced by variations in data density (Bouttier 1993).

Derber and Bouttier's formulation (the current J_b) may be seen as one end of a spectrum. It allows full resolution of the variation of vertical correlation with horizontal scale (as measured by n), but it allows no horizontal variability of the vertical correlations. At the other end of the spectrum is the separable formulation which allows full horizontal variation of the vertical correlations (we may specify a different vertical covariance matrix for each horizontal grid point), but has no variation of vertical correlation with horizontal scale. The wavelet J_b achieves a compromise between these two extremes and allows a degree of variation of vertical correlation with both wavenumber and horizontal location. Moreover, it also allows horizontal variation of horizontal correlation.

The multivariate aspects of the wavelet J_b are identical to those described by Derber and Bouttier (1999). Cross-correlations between mass and wind are described by a statistically-derived "balance operator" which subtracts from the temperature, log(surface pressure) and divergence, the components which can be explained by the vorticity. The residual "unbalanced" fields, together with vorticity, specific humidity and ozone, are treated univariately. The differences between the wavelet J_b and the current J_b formulation lie in the different treatment of these univariate covariances, which we now describe.

The wavelet J_b is based on a wavelet expansion for the sphere. In recent years, there have been several attempts to formulate such an expansion. Frequently, an orthogonal basis is sought. However, any such orthogonal basis will necessarily manifest some form of "pole problem", making it unattractive for use in a spectral model. Recently, Freeden and Windheuser (1996) have suggested defining spherical wavelets in terms of convolutions with radial basis functions (i.e. functions of great-circle distance, r). Such wavelets are free from "pole problems", but are non-orthogonal.

Freeden and Windheuser's idea is to generate wavelets from a family of bell-shaped radial basis functions $\varphi_j(|r|)$. Freeden and Windheuser define the functions in terms of a single generator function. This has advantages when considering infinite expansions of functions on the sphere, since it makes the wavelet expansion easier to handle mathematically. In the case of a finite wavelet expansion, this restriction is not necessary. Instead, we may define a sequence of "cutoff" wavenumbers N_j (with $N_j < N_{j+1}$) and require that the spectral transform of $\varphi_j(|r|)$ satisfy

$$\begin{aligned} \hat{\varphi}_j(n) &= 0 \qquad \textbf{for } n > N_j \\ \hat{\varphi}_j(n) &= 1 \qquad \textbf{for } n \le N_{j-1} \end{aligned} \tag{20}$$

The wavelets are defined in terms of $\hat{\varphi}_j(n)$ as $\hat{\psi}_j(n) = \sqrt{\hat{\varphi}_j^2(n) - \hat{\varphi}_{j-1}^2(n)}$. Note that $\hat{\psi}_j$ is zero outside that the range $N_{j-1} \le n \le N_{j+1}$. So, convolution with $\psi_j(|r|)$ results in a spectral bandpass filtering. Furthermore, $\psi_j(|r|)$ is itself a radial basis function which decays with great-circle distance. So, convolution with $\psi_j(|r|)$ may also be thought of as a localized spatial averaging. In other words, convolution with $\psi_j(|r|)$ achieves the simultaneous localization in space and wavenumber which is the hallmark of a wavelet transform.

The non-orthogonal wavelet expansion of a function on the sphere $f(\lambda,\theta)$ is defined as the set of functions $f_j = f \otimes \psi_j$ (where $\otimes$ denotes convolution). These functions have finite spectral truncations. They may therefore be represented exactly on grids of appropriate resolution. The value of f_j at a given grid point is a localized spatial average of f at nearby points.

An important property of the wavelet transform is that f may be reconstructed from its transform by further convolutions with $\psi_j(|r|)$:

$$f = \sum_j f_j \otimes \psi_j \tag{21}$$

This is easily proved by taking the spectral transform of the right hand side of equation 21:

$$\begin{aligned} \sum_j \hat{f}_j(m,n)\hat{\psi}_j(n) &= \sum_j \hat{f}(m,n)\hat{\psi}_j^{\,2}(n) \\ &= \sum_j \hat{f}(m,n)(\hat{\varphi}_j^2(n) - \hat{\varphi}_{j-1}^2(n)) \end{aligned} \tag{22}$$

The final sum collapses to leave only:

$$\hat{f}(m,n)(\hat{\varphi}_J^2(n) - \hat{\varphi}_0^2(n)) \tag{23}$$

The result holds if we define $\hat{\varphi}_J(n) = 1 \ \forall n$ and $\hat{\varphi}_0(n) = 0 \ \forall n$.

This property (equation 21) of the wavelet transform means that to reconstruct the function f from its transform $\hat{f}_j$, we apply localized spatial averaging to each of the functions $\hat{f}_j$ before summing over j. It is this property which allows us to construct the wavelet J_b.

The wavelet J_b, like the current J_b formulation, defines the background error covariance matrix implicitly via a change of variable, $\mathbf{L}$. The minimization is carried out in terms of a control variable χ for which the background cost function is $J_b = \frac{1}{2}\chi^T\chi$. The departures of model variables from the background are given by $\mathbf{x}\text{-}\mathbf{x}_b = \mathbf{L}\chi$. In the wavelet J_b, the control vector is defined in the wavelet space as a set of functions χ_j, and the background cost function is

$$J_b = \frac{1}{2}\sum_j \chi_j^T\chi_j \tag{24}$$

The univariate part of the change of variable consists of the following steps. First, each vertical column of χ_j for each j is multiplied by the symmetric square-root of a square matrix $\mathbf{C}_j(\lambda, \theta)$ whose columns have dimension equal to the number of levels. In principle, there is one such covariance matrix for each grid point for each j. In practice, to reduce the computer memory required to store them, the matrices are stored on coarser grids and interpolated to the grid points as needed.

Next, the functions $\mathbf{C}_j^{1/2}(\lambda, \theta)\chi_j$ are convolved with the wavelet functions $\psi_j(|r|)$ and summed according to equation 21. The matrices $\mathbf{C}_j(\lambda, \theta)$ account for both the vertical and horizontal correlations of background error. So, the reconstituted function must be multiplied by the standard deviations of background error, and operated on by the balance operator to restore mass-wind cross-correlation. These last two steps of the change of variable are identical to the corresponding steps of the current J_b formulation.

It is convenient to write the change of variable in matrix form:

$$\mathbf{L} = \mathbf{K}\mathbf{S}_J\Sigma_b\mathbf{S}_J^{-1}(\Psi_0\mathbf{S}_0\mathbf{C}_0{}^{1/2},\Psi_1\mathbf{S}_1\mathbf{C}_1{}^{1/2},\ldots,\Psi_J\mathbf{S}_J\mathbf{C}_J{}^{1/2}) \tag{25}$$

where $\mathbf{K}$ is the balance operator and Σ_b is the diagonal matrix of standard deviations of background error. Subscripts j indicate the sequence of wavelet functions. For $j = 0$ the wavelet function is broad and its spectral transform is limited to the lowest wavenumbers. For $j = J$, the wavelet function is narrow, and its spectral transform is limited to the highest resolved wavenumbers. The matrices $\mathbf{S}_j$ represent spherical transforms from the grid appropriate to the spectral truncation imposed by convolution with $\psi_j(|r|)$. The convolutions are represented by the matrices Ψ_j, which are diagonal (in spectral space). The matrices $\mathbf{C}_j$ are block diagonal with one block for each variable and for each horizontal grid point. These blocks are the matrices $\mathbf{C}_j(\lambda, \theta)$. The background error covariance matrix is given by $\mathbf{B} = \mathbf{L}\mathbf{L}^T$. That is,

$$\mathbf{B} = \mathbf{K}\mathbf{S}_J\Sigma_b\mathbf{S}_J^{-1}\left(\sum_j \Psi_j\mathbf{S}_j\mathbf{C}_j\mathbf{S}_j^T\Psi_j\right)\mathbf{S}_J^{-T}\Sigma_b\mathbf{S}_J^T\mathbf{K}^T \tag{26}$$

To understand the effect of the wavelet J_b, we will consider the matrix represented by the sum (in parentheses) in equation 26, since the remainder of the matrix is identical to the current J_b formulation. We note also that the current J_b formulation corresponds to replacing the term in parentheses in equation 26 by a block diagonal matrix whose blocks vary with total wavenumber. Each block in this formulation is a matrix $\mathbf{H}_n\mathbf{V}_n\mathbf{H}_n^T$, where $\mathbf{V}_n$ is the vertical correlation matrix for each total wavenumber n, and $\mathbf{H}_n$ is a diagonal matrix whose elements are the modal variances at each level. ($\mathbf{H}_n$ defines the horizontal structure functions.)

For the wavelet J_b, consider first the case in which the matrices $\mathbf{C}_j(\lambda, \theta)$ are independent of latitude and longitude. Then, $\mathbf{C}_j$ commutes with the spherical transform, and we arrive at the following expression for the correlation between coefficients for different wavenumbers (n_1 and n_2) and levels (k_1 and k_2):

$$\sum_j \hat{\psi}_j(n_1)\hat{\psi}_j(n_2)(C_j)_{k_1, k_2} \tag{27}$$

Note that the wavelet J_b does not assume that different total wavenumbers are uncorrelated. For the particular case $n_1 = n_2 = n$, we have

$$\sum_j \hat{\psi}_j^2(n)(C_j)_{k_1,\,k_2} \tag{28}$$

Now, $\hat{\psi}_j$ is zero outside that the range $N_{j-1} \le n \le N_{j+1}$. So, for $n \ge N_j$, only the terms involving $\hat{\psi}_j$ and $\hat{\psi}_{j+1}$ contribute to the sum, whereas for $n < N_j$, only the terms involving $\hat{\psi}_j$ and $\hat{\psi}_{j-1}$ contribute to the sum. Furthermore, it is easy to show that $\hat{\psi}_j^2(n) + \hat{\psi}_{j+1}^2(n) = 1$ for $N_j \le n \le N_{j+1}$, and $\hat{\psi}_j^2(n) + \hat{\psi}_{j-1}^2(n) = 1$ for $N_{j-1} \le n < N_j$.

So, equation 28 represents an interpolation between a pair of matrices C_j, each of which may be thought of as equivalent to the matrix $\mathbf{H}_{N_j}\mathbf{V}_{N_j}\mathbf{H}_{N_j}^T$ of the current formulation. Thus, the wavelet J_b retains the non-separability property of the current formulation, but at a reduced spectral resolution determined by the widths of the wave number ranges $[N_j, N_{j+1}]$.

Note that it is the choice of the bandwidths $[N_j, N_{j+1}]$ which determines the trade-off between spectral and spatial resolution. If the bands are narrow, the corresponding wavelet functions are not spatially localized. In the limit of one band per wavenumber, the wavelet J_b is identical to the current J_b. On the other hand, a single band containing all wavenumbers corresponds to a delta function on the sphere. That is, full horizontal variability of vertical correlation is allowed, but no variation of vertical correlation with horizontal scale is possible. (The horizontal correlations would also be very poorly described in this case.) A mathematically precise “uncertainty principle” for radial basis functions on the sphere is given by Freeden et al. (1998).

Next, for simplicity, assume that the grids used to represent the functions χ_j are the same for all values of j (i.e. $\mathbf{S}_j = \mathbf{S}_J$ for all j). This does not affect the covariance model, just the storage and calculation requirements. In this case, if we transform the matrix represented by the sum (in parentheses) in equation 26 to grid space, we get

$$\sum_j (\mathbf{S}^{-1}\Psi_j\mathbf{S})\mathbf{C}_j(\lambda,\theta)(\mathbf{S}^{-1}\Psi_j\mathbf{S})^T \tag{29}$$

The matrices $\mathbf{S}^{-1}\Psi_j\mathbf{S}$ represent convolution with the radial basis function $\psi_j(|r|)$ in grid space. So, each term in the sum represents a weighted horizontal averaging, or smoothing, of the matrices $\mathbf{C}_j(\lambda,\theta)$. The smoothing is appropriate to the scales represented by $\mathbf{C}_j(\lambda,\theta)$. If we consider a single horizontal grid point, we see that both the vertical correlation matrix and the horizontal structure function are determined by a localized average of nearby matrices $\mathbf{C}_j(\lambda,\theta)$. In other words, the wavelet J_b achieves horizontal variation of both the vertical correlations and the horizontal structure functions.

To compare the vertical and horizontal correlations of the wavelet J_b and the current J_b, it is convenient to construct “effective” matrices $\mathbf{H}_n\mathbf{V}_n\mathbf{H}_n^T$ from the wavelet J_b for a given latitude and longitude. We did this for figure 18 and figure 19 by first interpolating the matrices $\mathbf{C}_j^{1/2}(\lambda,\theta)$ for each j to the selected latitude and longitude using bilinear interpolation, and then constructing the effective $\mathbf{H}_n\mathbf{V}_n\mathbf{H}_n^T$ matrix by interpolating in wavenumber using the coefficients $\hat{\psi}_j^2(n)$. Note that bilinear horizontal interpolation was used for computational convenience. Strictly, to faithfully reproduce the wavelet structure functions, the horizontal interpolation should use weighted averaging of nearby matrices, with weights given by $\psi_j^2(|r|)$.

Appendix D There is More to 4D-Var than Covariance Evolution!

Consider a 4D-Var analysis at some time t_a of a linear system for a case in which we have a single observation of the entire state at some time $t_o > t_a$. With this simplification, the 4D-Var cost function becomes:

$$\mathbf{J_{4dVar}} = (\mathbf{x}-\mathbf{x}_b)^T\mathbf{B}^{-1}(\mathbf{x}-\mathbf{x}_b) + (\mathbf{y}-\mathbf{M}_{t_o}\mathbf{x})^T\mathbf{R}^{-1}(\mathbf{y}-\mathbf{M}_{t_o}\mathbf{x}) \tag{30}$$

where $\mathbf{M}_{t_o}$ is the resolvent of the dynamics for the interval $t_a \le t \le t_o$.

The analysis corresponds to the minimum of the cost function. Setting the gradient of the cost function to zero, and rearranging slightly, we find that the analysis is given by:

$$\mathbf{x}_a = (\mathbf{B}^{-1} + \mathbf{M}_{t_o}{}^T\mathbf{R}^{-1}\mathbf{M}_{t_o})^{-1}(\mathbf{B}^{-1}\mathbf{x}_b + \mathbf{M}_{t_o}^T\mathbf{R}^{-1}\mathbf{y}) \tag{31}$$

An equation for the analysis error ε_a is easily derived from equation 31 by writing $\mathbf{x}_a = \mathbf{x}_* + \varepsilon_a$, $\mathbf{x}_b = \mathbf{x}_* + \varepsilon_b$ and $\mathbf{y} = \mathbf{y}_* + \varepsilon_o$, where the subscript $*$ denotes the true state. Note that, for a perfect model, $\mathbf{y}_* = \mathbf{M}_{t_o}\mathbf{x}_*$. This allows us to simplify the equation for the analysis error to:

$$\varepsilon_a = (\mathbf{B}^{-1} + \mathbf{M}_{t_o}{}^T\mathbf{R}^{-1}\mathbf{M}_{t_o})^{-1}(\mathbf{B}^{-1}\varepsilon_b + \mathbf{M}_{t_o}^T\mathbf{R}^{-1}\varepsilon_o) \tag{32}$$

The covariance matrix of analysis error is therefore:

$$\mathbf{P}^a_{\mathbf{4dVar}}(t_a) = (\mathbf{B}^{-1} + \mathbf{M}_{t_o}{}^T\mathbf{R}^{-1}\mathbf{M}_{t_o})^{-1}. \tag{33}$$

A 3D-Var analysis for this system corresponds to replacing $\mathbf{M}_{t_o}$ by the identity matrix in equation 31. The corresponding equation for the 3D-Var analysis error is:

$$\varepsilon_a = (\mathbf{B}^{-1} + \mathbf{R}^{-1})^{-1}(\mathbf{B}^{-1}\varepsilon_b + \mathbf{R}^{-1}\varepsilon_o + \mathbf{R}^{-1}(\mathbf{M}-\mathbf{I})\mathbf{x}_t) \tag{34}$$

Note that the 3D-Var analysis error contains an additional, first order error, $\mathbf{R}^{-1}(\mathbf{M}-\mathbf{I})\mathbf{x}_t$, which is not present in 4D-Var. However, since the covariance matrix of analysis error is defined as $\langle(\varepsilon_a - \langle\varepsilon_a\rangle)(\varepsilon_a - \langle\varepsilon_a\rangle)^T\rangle$, the first order error does not contribute to the covariance matrix, which is given by:

$$\mathbf{P}^a_{\mathbf{3dVar}}(t_a) = (\mathbf{B}^{-1} + \mathbf{R}^{-1})^{-1} \tag{35}$$

Now, suppose that the resolvent of the dynamics is an orthogonal matrix, and that the covariance matrices of observation error are proportional to the identity matrix. Consider the evolution of the covariance matrices during the analysis. For a perfect model, the background error covariance evolves as:

$$\mathbf{P}^b(t) = \mathbf{M}_t\mathbf{B}\mathbf{M}_t^T \tag{36}$$

However, since $\mathbf{B}$ is proportional to the identity matrix, it commutes with $\mathbf{M}_t$. Furthermore, since $\mathbf{M}_t$ is orthogonal, we have $\mathbf{M}_t\mathbf{M}_t^T = \mathbf{I}$. Hence, in this contrived example, *the covariance matrix of background error does not evolve*:

$$\mathbf{P}^{b}(t) = \mathbf{B}. \tag{37}$$

Next, consider the covariance matrix of 4D-Var analysis error (equation 33). Again, since $\mathbf{M}_t$ is orthogonal and $\mathbf{R}^{-1}$ is proportional to the identity matrix, we find that:

$$\mathbf{P}^{a}_{\mathbf{4dVar}}(t_a) = (\mathbf{B}^{-1} + \mathbf{R}^{-1})^{-1}. \tag{38}$$

That is, *the covariance matrix of analysis error for the 4D-Var analysis is identical to that of the 3D-Var analysis*. Moreover, the covariance matrix of analysis error is itself proportional to the identity matrix. Now, for a perfect model, the covariance matrix of forecast error is $\mathbf{P}^{f}(t) = \mathbf{M}_t\mathbf{P}^{a}(t_a)\mathbf{M}_t^{T}$. But, since $\mathbf{M}_t$ is orthogonal and $\mathbf{P}^{a}(t_a)$ is proportional to the identity matrix, we find that *the covariance matrix of forecast error is constant*:

$$\mathbf{P}^{f}(t) = \mathbf{P}^{a}(t_a). \tag{39}$$

In summary, this example demonstrates an idealized system for which there is no covariance evolution, and for which the covariance matrices of analysis error for 3dVar and 4dVar are identical. It is neverthless the case that the 4dVar analysis is more accurate than the 3dVar analysis. This is in part because the 3dVar analysis contains a *first order* error which is not present in 4dVar. The extent to which reduction of first order error explains the superior performance of the ECMWF 4dVar, relative to the 3dVar system, remains an open question.

The first order error of 3dVar may be eliminated using the 3dFGAT method described in section 4.1. For the example presented here, the 3dFGAT analyis equation may be written as:

$$\mathbf{x}_a = \mathbf{x}_b + \mathbf{P}^{a}\mathbf{R}^{-1}(\mathbf{y} - \mathbf{M}_{t_0}\mathbf{x}_b)$$

Written in this form, the corresponding 3dVar and 4dVar analysis equations are respectively:

$$\mathbf{x}_a = \mathbf{x}_b + \mathbf{P}^{a}\mathbf{R}^{-1}(\mathbf{y} - \mathbf{x}_b) \qquad \text{(3dVar)}$$

and $$\mathbf{x}_a = \mathbf{x}_b + \mathbf{P}^{a}\mathbf{M}_{t_0}^{T}\mathbf{R}^{-1}(\mathbf{y} - \mathbf{M}_{t_0}\mathbf{x}_b) \qquad \text{(4dVar).}$$

Note that both 3dFGAT and 4dVar propagate the background to the time of the observation using the model. However, only 4dVar correctly propagates the background departure to the time of the analysis, using the adjoint dynamics. Thus, 4dVar is superior to 3dFGAT in this example. This superiority is not due to covariance propagation, since the covariance matrix of analysis error is identical for both analyses.

Fig. 22 illustrates the example given above for the case in which the state is a single two-dimensional vector, and $\mathbf{M}_t$ corresponds to rotation by an angle $\theta(t)$. For simplicity, the variances of background and observation error are assumed to be equal, and the observation and the evolved background are in the same direction.

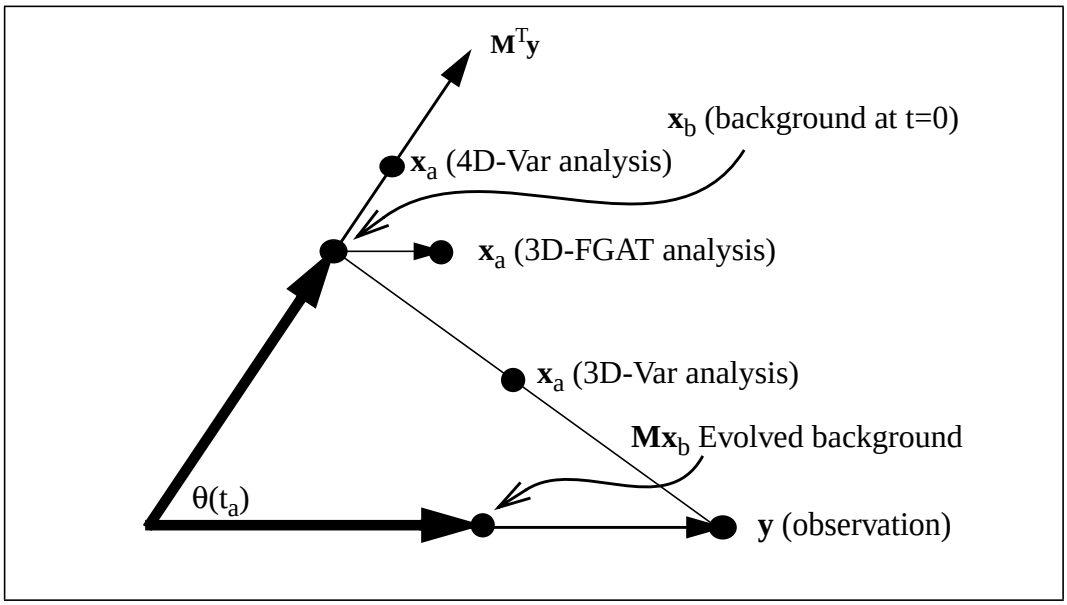

Figure 22: Schematic representation of an analysis system for which the covariance matrices of error for 3dVar and 4dVar are identical, but for which 4dVar is nevertheless superior to 3dVar.

The 4dVar analysis is optimal, and is given by equation 31, which for this example may be simplified to $\mathbf{x}_a = (\mathbf{x}_b + \mathbf{M}^T(\mathbf{y}\text{-}\mathbf{M}\mathbf{x}_b))/2$. The analysis is in the same direction as the background. The 3dVar analysis, on the other hand, is given by $\mathbf{x}_a = (\mathbf{x}_b + \mathbf{y})/2$. This averaging of an observation valid at $t = t_o$ with the background (valid at $t = 0$) produces an analysis which is incorrectly rotated with respect to the background. The 3dFGAT analysis is given by $\mathbf{x}_a = (\mathbf{x}_b + (\mathbf{y}\text{-}\mathbf{M}\mathbf{x}_b))/2$. This is also rotated with respect to the background. The error covariance matrices are identical for all three analyses.

References

Andersson, E., Haseler, J., Undén, P., Courtier, P., Kelly, G., Vasiljevic, D., Brankovic, C., Cardinali, C., Gaffard, C., Hollingsworth, A., Jakob, C., Janssen, P., Klinker, E., Lanzinger, A., Miller, M., Rabier, F., Simmons, A., Strauss, B., Thépaut, J-N. and Viterbo, P., 1998: The ECMWF implementation of three dimensional variational assimilation (3D-Var). Part III: Experimental results. *Q. J. R. Meteorol. Soc*., **124**, 1831-1860.

Andersson, E. and M. Fisher, 1999: Background errors for observed quantities and their propagation in time. Proc ECMWF Workshop on "Diagnosis of data assimilation systems", Reading, UK, 1-4 November 1998.

Andersson, E., M. Fisher, R. Munro and A. McNally, 2000: Diagnosis of background errors for radiances and other observable quantities in a variational data assimilation system, and the explanation of a case of poor convergence. *Q. J. R. Meteorol. Soc*., **126**, 1455-1472.

Barkmeijer, J., M. van Gijzen and F. Bouttier, 1998: Singular vectors and estimates of the analysis-error covariance metric. *Q. J. R. Meteorol. Soc*., **124**, 1695-1713.

Barkmeijer, J., R. Buizza and T. N. Palmer, 1999: 3D-Var Hessian singular vectors and their potential use in the ECMWF ensemble prediction system. *Q. J. R. Meteorol. Soc*., **125,** 2333-2351.

Bartello, P. and H. L. Mitchell, 1992, A continuous three-dimensional model of short-range forecast error covariances. *Tellus*, **44A,** 217-235.

Bouttier, F., 1993: The dynamics of error covariances in a barotropic model. *Tellus*, **45A,** 408-423.

Buizza, R. and T. N. Palmer, 1995: The singular-vector structure of the atmospheric global circulation. *J. Atmos. Sci.*, **52**, 1434-1456.

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", Reading 6-10 September 1993.

Courtier, P., Andersson, E., Heckley, W., Pailleux, J., Vasiljevic, D., Hamrud, M., Hollingsworth, A., Rabier, F. and Fisher, M., 1998: The ECMWF implementation of three dimensional variational assimilation (3D-Var). Part I: Formulation. *Q. J. R. Meteorol. Soc.,* **124**, 1783-1808.

Derber, J. and F. Bouttier, 1999, A reformulation of the background error covariance in the ECMWF global data assimilation system. *Tellus*, **51A**, 195-221.

Farrell, B. F. and P. J. Ioannou, 2001a: Accurate low dimensional approximation of the linear dynamics of fluid flow. To appear in *J. Atmos. Sci.*

Farrell, B. F. and P. J. Ioannou, 2001b: State estimation using a reduced order Kalman filter. To appear in *J. Atmos. Sci.*

Fisher, M. and P. Courtier, 1995: Estimating the covariance matrices of analysis and forecast error in variational data assimilation. ECMWF RD Tech. Memo. 220.

Fisher, M., 1996: The specification of background error variances in the ECMWF variational analysis system. Proc. ECMWF workshop on "Non-linear aspects of data assimilation", Reading, UK, 9-11 September 1996.

Fisher, M., 1998a Development of a Simplified Kalman Filter. ECMWF RD Tech. Memo 260.

Fisher, M., 1998b Minimization Algorithms for Variational Data Assimilation, Proc. ECMWF Seminar on "Recent Developments in Numerical Methods for Atmospheric Modelling", Reading 7-11 September 1998.

Freeden, W., T. Gervens and M. Schreiner, 1998: Constructive approximation on the sphere with applications to geomathematics. Clarendon Press, Oxford. ISBN 0 19 853682.

Freeden, W. and U. Windheuser, 1996: Spherical wavelet transform and its discretization. *Adv. Comput. Math.* **Vol. 5,** 51-94.

Gelaro, R., R. Buizza, T. N. Palmer and E. Klinker, 1998: Sensitivity analysis of forecast errors and the construction of optimal perturbations using singular vectors. *J. Atmos. Sci.*, ***55,*** 1012-1037.

Gilbert, J. C. and C. Lemarechal, 1989: Some numerical experiments with variable storage quasi-Newton algorithms. *Math. Prog.*, **B25,** 407-435.

Ide, K., P. Courtier, M. Ghil and A. C. Lorenc, 1997: Unified notation for data assimilation: operational sequential and variational. *J. Met. Soc. Japan*, **75**, 181-189.

Ingleby, B., 2001: The statistical structure of forecast errors and its representation in The Met. Office Global 3-D variational data assimilation scheme. *Q. J. R. Meteorol. Soc.*, **127**, 209-231.

Janiskova, M., 2001: Preparatory studies for the use of observations from the Earth Radiation Mission in numerical weather prediction. ESA Contract Report. Published by ECMWF.

Klinker, E., F. Rabier and R. Gelaro, 1998: Estimation of key analysis errors using the adjoint technique. *Q. J. R. Meteorol. Soc*., **124**, 1909-1933.

Lanczos, C., 1950: An iteration method for the solution of the eigenvalue problem. *J. Res. Nat. Bur. Standards*, **Vol. 45,** 255-282.

Lönnberg, P., 1988: Developments in the ECMWF analysis scheme. Proc. ECMWF Seminar on “Data assimilation and the use of satellite data”, Reading, 5--9 Sept. 1988, 75-119.

Mahfouf, J.-F. and F. Rabier, 2000: The ECMWF operational implementation of four dimensional variational assimilation. Part II: Experimental results with improved physics. *Q. J. R. Meteorol. Soc*., **126**, 1171-1190.

Molteni, F. and T. N. Palmer, 1993: Predictability and finite-time instability of the northern winter circulation. *Q. J. R. Meteorol. Soc.*, **119,** 269-298.

Molteni, F., R. Buizza, T. N. Palmer and T. Petroliagis, 1996: The ECMWF ensemble prediction system: Methodology and validation. *Q. J. R. Meteorol. Soc.*, **122,** 73-119.

Palmer, T. N., R. Gelaro, J. Barkmeijer and R. Buizza, 1998: Singular vectors, metrics and adaptive observations. *J. Atmos. Sci.*, **55,** 633-653.

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

Rabier, F., E. Klinker, P. Courtier and A. Hollingsworth, 1996, Sensitivity of Forecast Errors to Initial Conditions, *Q. J. R. Meteorol. Soc*., **122**, 121-150.

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 RD Tech. Memo. 240.

Rabier, F., McNally, A., Andersson, E., Courtier, P., Undén, P., Eyre, J., Hollingsworth, A., and Bouttier, F., 1998: The ECMWF implementation of three dimensional variational assimilation (3D-Var). Part II: Structure functions. *Q. J. R. Meteorol. Soc*., **124**, 1809-1829.

Rabier, F., Järvinen, H., Klinker, E., Mahfouf, J-F. and A. Simmons, 2000: The ECMWF operational implementation of four dimensional variational assimilation. Part I: Experimental results with simplified physics. *Q. J. R. Meteorol. Soc*., **126**, 1143-1170.

Reynolds, C.A., R. Gelaro and J.D. Doyle, 2001: Relationship between singular vectors and transient features in the background flow. *Q. J. R. Meteorol. Soc*., **127**, 1731-1760.

Savijärvi, H., 1995: Error growth in a large numerical forecast system. *Mon. Weather Rev.*, **123,** 212-221.

Thépaut, J.-N., Hoffman, R. N. and Courtier, P., 1993: Interactions of dynamics and observations in a four-dimensional variational assimilation. *Mon. Wea. Rev.*, **121**, 3393-3414.

Thépaut, J.-N., Courtier, P., Belaud, G. and Lemaître, G., 1996: Dynamical structure functions in four-dimensional variational assimilation: A case study. *Q. J. R. Meteorol. Soc*., **122**, 535-561.