APPLICATION OF KALMAN FILTERING
TO NUMERICAL WEATHER PREDICTION

F Bouttier
European Centre for Medium-Range Weather Forecasts
Shinfield Park, Reading, UK

Summary : Despite its theoretical advantages, the Extended Kalman Filter (EKF) algorithm cannot be used as such in data assimilation for operational numerical weather prediction (NWP). However, practical experience shows that a suitably simplified version should be useful if it is conveniently interfaced with a more conventional assimilation scheme such as variational assimilation (3D-Var or 4D-Var). One implementation problem is the representation of flow-dependent error covariances, and this paper discusses a solution based on eigenvalue decompositions.

1. KALMAN FILTERING AND NUMERICAL WEATHER PREDICTION

1.1 The Extended Kalman Filter

The Kalman filter equations have already been derived and discussed in several publications on meteorology and oceanography (e.g. Ghil 1989, Bouttier 1995) as well as in other chapters of this volume, so that they will be only briefly recalled here in their Extended Kalman Filter (EKF) discrete version. The notations used are those advocated in Ide et al (1995) :

x^t	true fluid state
x^a, x^f	analysis and forecast vectors, respectively
$\mathbf{P}^a$, $\mathbf{P}^f$	analysis and forecast error covariances
M	prediction operator as defined by the forecast model
$\mathbf{M}$	tangent linear prediction operator
η	random model error
$\mathbf{Q}$	error covariance matrix of the model error
y^o	vector of observed values
ε	random observation error
$\mathbf{R}$	error covariance matrix of the observation error
H	observation operator
$\mathbf{H}$	linearized observation operator
$\mathbf{K}$	analysis gain matrix
t_i	time indexed by i

The EKF relies on the Kalman filter hypotheses :

(i) model error : The relationship between the model forecast and the real flow dynamics can be written as $\mathrm{x}^t(t_{i+1}) = M(t_{i+1}, t_i)\mathrm{x}^t(t_i) + \eta(t_i)$, where the random model errors η have covariances $\overline{\eta(t_i)\eta^{\mathrm{T}}(t_i)} = \mathbf{Q}(t_i)$

(ii) observation error : The relationship between the true fluid state and the observed values can be written as $\mathrm{y}_i^o = H_i\mathrm{x}^t(t_i) + \varepsilon(t_i)$, where the so-called observation errors ε have covariances $\overline{\varepsilon(t_i)\varepsilon^{\mathrm{T}}(t_i)} = \mathbf{R}(t_i)$

(iii) no bias : $\overline{\eta} = 0, \overline{\varepsilon} = 0$

(iv) no serial error correlation : $\overline{\eta(t_i)\eta^{\mathrm{T}}(t_j)} = 0$ and $\overline{\varepsilon(t_i)\varepsilon^{\mathrm{T}}(t_j)} = 0$ if $i \neq j$

(v) no cross-correlation between observation and model error : $\overline{\eta\varepsilon^{\mathrm{T}}} = 0$

The EKF differs from the plain Kalman filter in that the model and the observation operator are not assumed to be linear, so that the optimality of the filter is obtained only to the extent that the following linearization hypotheses are fulfilled :

(vi) tangent linear hypothesis on forecast model : if $\delta x(t_i)$ is an estimation error,

$$\delta \mathrm{x}(t_{i+1}) \simeq \mathbf{M}(t_{i+1}, t_i)\delta \mathrm{x}(t_i)$$

(vii) tangent linear hypothesis on observation operator :

$$H_i[\mathrm{x}(t_i) + \delta x(t_i)] - H_i[\mathrm{x}(t_i)] \simeq \mathbf{H}_i\delta \mathrm{x}(t_i)$$

Some numerical experiments (e.g. Lacarra and Talagrand 1988) have shown that the tangent linear hypotheses are good approximations for the synoptic-scale dynamics at ranges up to 48 hours, except in some particular cases (e.g. in convective situations or in the boundary layer) ; for some mesoscale phenomena the approximation is liable to break down after a few hours (Vukicevic 1991).

The EKF equations are essentially those of the Kalman filter in which the forecast model and the observation operators are replaced by their tangent linear counterparts wherever errors are involved, the linearization being carried out in the vicinity of the current estimate of the model state :

$$\begin{array}{ll}
\text{(a)} & \mathrm{x}^f(t_{i+1}) = M(t_{i+1}, t_i)\mathrm{x}^a(t_i) \\
\text{(b)} & \mathbf{P}^f(t_{i+1}) = \mathbf{M}(t_{i+1}, t_i)\mathbf{P}^a(t_i)\mathbf{M}^{\mathrm{T}}(t_{i+1}, t_i) + \mathbf{Q}(t_i) \\
\text{(c)} & \mathbf{K}_i = \mathbf{P}^f(t_i)\mathbf{H}_i^{\mathrm{T}}[\mathbf{H}_i\mathbf{P}^f(t_i)\mathbf{H}_i^{\mathrm{T}} + \mathbf{R}_i(t_i)]^{-1} \\
\text{(d)} & \mathrm{x}^a(t_i) = \mathrm{x}^f(t_i) + \mathbf{K}_i[\mathrm{y}_i^o - H_i\mathrm{x}^b(t_i)] \\
\text{(e)} & \mathbf{P}^a(t_i) = [\mathbf{I} - \mathbf{K}_i\mathbf{H}_i]\mathbf{P}^f(t_i)
\end{array}$$

The equations (a) and (d) are the well-known operations of model state forecast and linear analysis, whereas (b) and (e) are their counterparts for the estimation error covariances. They are derived by multiplying (a) and (d) by their respective transposes, taking the expectation and simplifying the results using the hypotheses. The optimal analysis weights $\mathbf{K}$ are derived

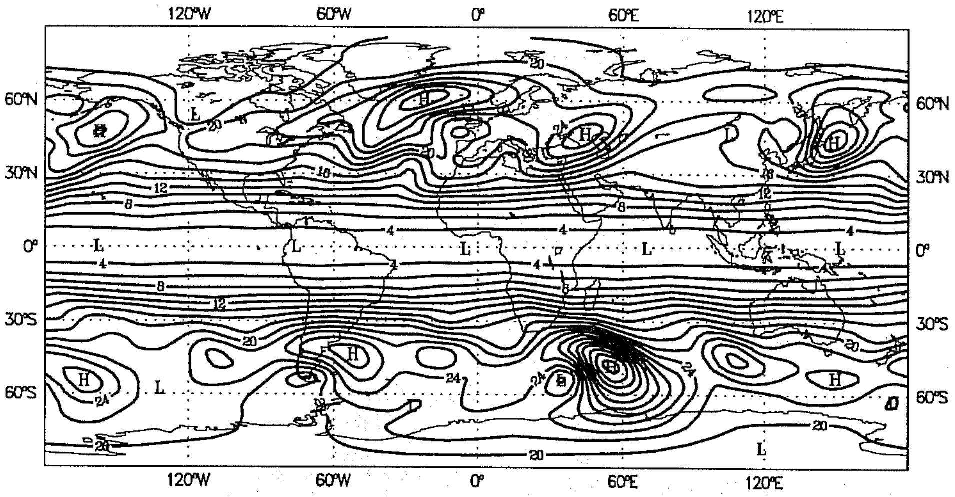

Figure 1: 24h-forecast of the 500hPa height standard error field (in m), starting with a zonally uniform standard errors, an homogeneous and isotropic height correlation model, without model error, using as dynamics the tangent-linear of a T21 barotropic vorticity equation model in the vicinity of an operational forecast. Notice how the forecast standard error increases in some areas along the extratropical jet.

in (c) according to the classic Best Linear Unbiased Estimate theory. Other forms of these equations are better suited for some applications ; they are mentioned in other contributions to this volume.

Numerous implementations of the EKF in simplified oceanic or atmospheric assimilation systems have been described in the litterature, e.g. : Parrish and Cohn 1985, Miller 1986, Dee 1991, Bouttier 1994 ; the EKF has never been tested with realistic primitive-equation models of the atmosphere, fortunately it is approximately equivalent to 4D-Var assimilation, so the published results on the structure functions of 4D-Var apply to the EKF as well, see Thépaut et al (1993) and Thépaut et al (1994). The most important features of an EKF assimilation are the sensitivity of standard error fields and correlations structures to the flow dynamics (figure 1) and to the structure of the observing network (figure 2). The complex vertical structure of the correlations and standard errors in the vicinity of an extratropical cyclone are shown in fig. 3 which was generated using simulated observations at the end of the 24-h period of a 4D-Var assimilation system.

1.2 Existing applications of the EKF to NWP

The EKF is a general tool in signal processing ; it has already been applied successfully to the adaptative statistical adaptation of forecast temperatures from numerical weather prediction models. There are some promising attempts to apply the EKF to data assimilation in reduced models of the atmosphere or the ocean (e.g. assimilation of wind from tracer information, see Daley (1995)). Here we discuss the potential for applying the EKF directly to state-of-the-art assimilation systems.

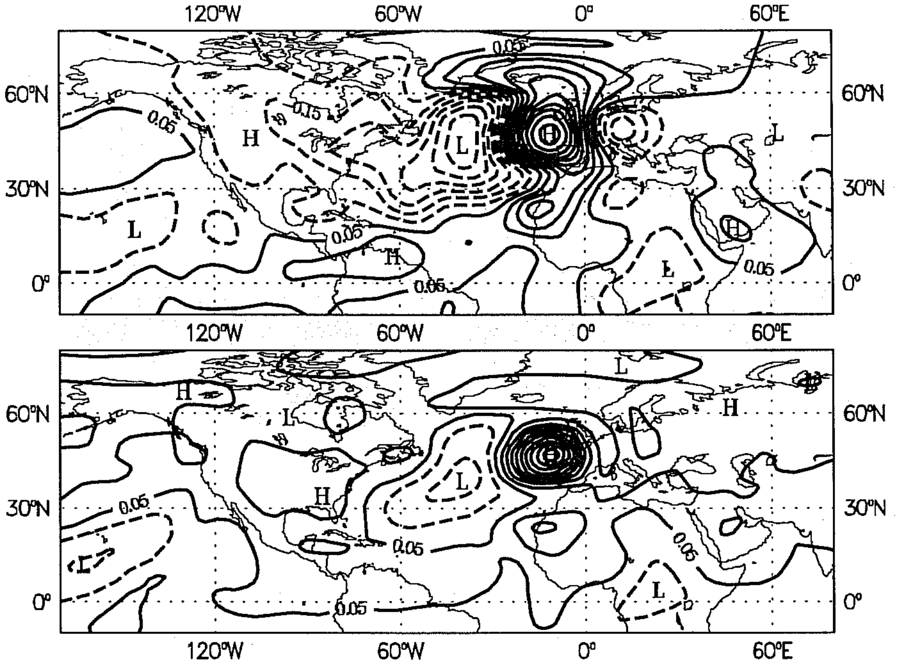

Figure 2: 500hPa field autocorrelation with a particular point, at an analysis time of an EKF based on a T21 barotropic vorticity equation model and 500hPa radiosonde observations of U,V and z. Top panel : background error correlation ; bottom panel : analysis error correlations. Notice how the background error structures have a flow-dependent tripolar structure along the jet, whereas the analysis correlations return to a more homogeneous and isotropic structure in the presence of dense, mutually uncorrelated observations over the continent.

The equations (a) and (d) are already used for operational NWP almost everywhere with the best available model M, observation operator H and observation error statistics $\mathbf{R}$. The main degrees of simplification are found in the weight computation (c), whereas the handling of the error covariances has received little attention so far.

In the optimal interpolation (OI), the weight computation (c) has been simplified by a kind of banded approximation on $\mathbf{P}^f$ which obviates the need for any big matrix inversion ; only a set of relatively small linear systems has to be solved because, for each model variable, the OI analysis uses only a limited set of observed variables defined by an *ad hoc* data selection algorithm. The $\mathbf{P}^f$ correlations follow a parametric model which is itself constrained by the data selection ; usually they are not flow-dependent. The error forecast (b) is approximated so that only the variances are managed, using a cheap empirical formula to approximate the error growth during the forecast steps of the assimilation. The error analysis (e) is performed for the variances only, too, using the approximate OI analysis weights.

In 3D-Var (Courtier 1993), the weights $\mathbf{K}$ are not explicitly computed, but the model analysis problem (c)-(e) is replaced by the determination of the analysis x^a as the solution of a variational problem which is exactly equivalent. Now the fundamental approximation lies in the accuracy to which the relevant cost-function $J(\mathrm{x})$ is minimized — experience shows that it is excellent with a limited number of iterations of a suitable minimization algorithm. How-

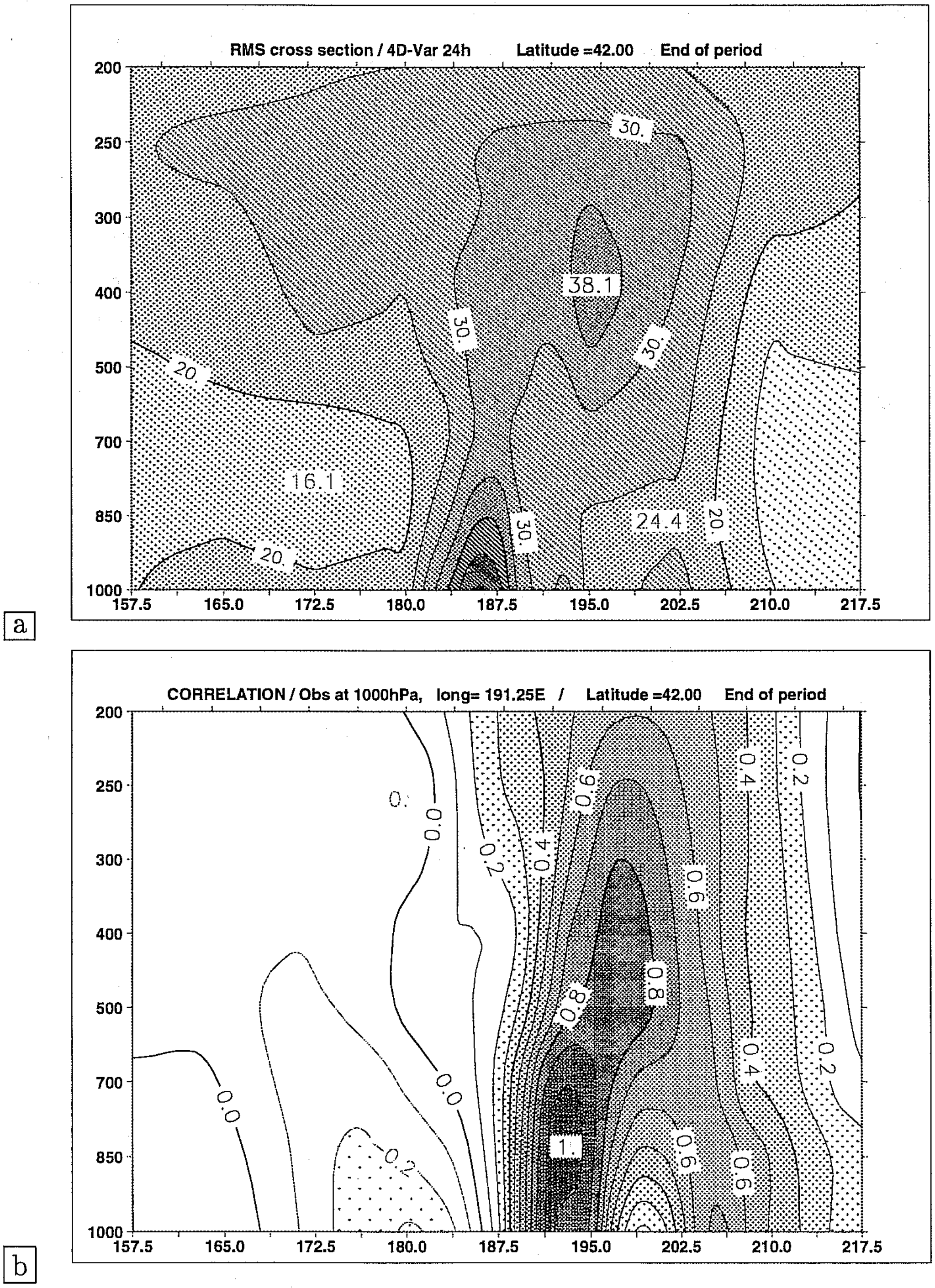

Figure 3: Vertical E-W cross-section of the effective background height standard errors (a) and autocorrelations with a particular surface height observation (b) in a 24-h 4D-Var assimilation of an extratropical low-pressure system (figures reproduced from Thépaut et al, 1994). The standard errors are maximum at the minimum of surface pressure, and along the upper-level jet at 350hPa. The correlations have a baroclinic structure, and tend to follow the shape of the air masses.

ever, in the existing operational 3D-Var systems, the handling of equations (b) and (e) is as approximate as in OI, and this implies that the analysis suffers from the approximations made on $\mathbf{P}^f$. It is not possible to access the analysis weights for computing (e), but $\mathbf{P}^a$ can in principle be estimated using the inverse of the Hessian (or second derivative) J'' of the analysis cost-function. This method is described later in this paper ; it has been applied successfully to the ECMWF 3D-Var system (Fisher and Courtier 1995).

In 4D-Var, (Talagrand and Courtier 1987, Thépaut and Courtier 1991) the minimization time span is limited by the tangent-linear hypothesis (among other problems), so that a 4D-Var-based NWP assimilation system will consist of a sequence of 4D-Var assimilations. Each 4D-Var solves the EKF equations approximately over its own period, assuming there is no model error and the initial forecast error covariances $\mathbf{P}^f$ are known. 4D-Var is known to produce analysis increments similar to the EKF for observations situated near the end of the assimilation interval (typically 12 to 24h, see fig.3), but it behaves basically like 3D-Var near the beginning. Actually, 4D-Var is algorithmically close to 3D-Var, so that it compares similarly to the EKF : there are approximations on $\mathbf{P}^f$ at the beginning of each 4D-Var period, the Hessian can be used to estimate $\mathbf{P}^a$ at the same time, but the situation is currently the same as in OI for the evaluation of equation (b) across different 4D-Var assimilations.

This explains why, although the move from OI to 3D-Var and then to 4D-Var will bring considerable improvement to the quality of the assimilation systems, there is a weakness in the specification of the forecast error covariances that 4D-Var will only solve to a certain extent. Although the EKF is very complex both theoretically and technically, it makes sense now to try to implement at least an approximation of it for operational NWP systems.

1.3 Practical requirements for an operational EKF

The implementation of the EKF on top of one of the existing operational data assimilation systems can be divided into 3 main tasks, and each of them relies on the preceding ones :

Estimation of the analysis error covariances. This is equation (e). Even if the remainder of the EKF is not implemented, a good knowledge of $\mathbf{P}^a$ is necessary in order to provide sensible statistics to the next analysis. By itself, a realistic estimate of the analysis error variances would already be a valuable by-product of the assimilation (it is an indication of the quality of the analyzed fields). The inverse of the matrix $\mathbf{P}^a$ is related to the second derivative of the cost-function in the case of a variational analysis, so that it may be useful for preconditioning the subsequent analyses if one can assume some degree of stationarity in the characteristics of the minimization problem. Of course, if $\mathbf{P}^a$ is to represent the real error statistics of the analysis, its computation should not only rely on the design of the analysis, it should also account for the weaknesses in the analysis algorithm itself, including flaws in the estimation of $\mathbf{P}^f$ and $\mathbf{R}$. Such flaws usually imply extra errors in the analysis.

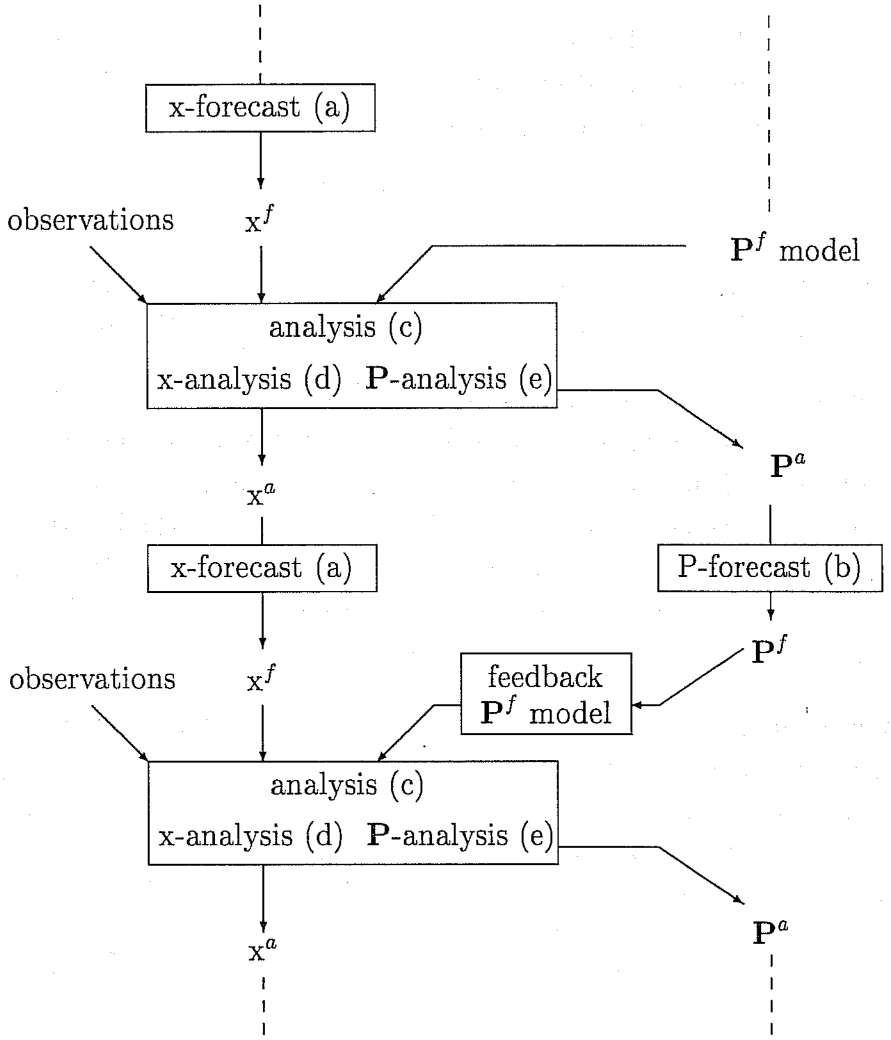

Figure 4: Organization of a simplified EKF around a preexisting sequential data assimilation system.

Estimation of the forecast error covariances. This is equation (b). The quality of the result is of course going to depend on the precision of the estimation of $\mathbf{P}^a$. The computation of $\mathbf{P}^f$, or even only some of its variances, provides indications about the short-range predictability of the atmospheric flow. Actually, it can be shown that the eigenvectors of $\mathbf{P}^f$ associated to the largest eigenvalues are exactly the singular vectors (SVs) which are used in the ensemble prediction at ECMWF (Buizza et al 1992). Equation (b) contains $\mathbf{Q}$, a representation of the errors caused by defaults in the model formulation. A realistic estimate of $\mathbf{P}^f$ should also account for weaknesses in the tangent linear hypothesis and in the estimation of $\mathbf{P}^a$.

Feed-back of the forecast error covariances into the analysis. Algebraically this is trivially the substitution of $\mathbf{P}^f$ computed by equation (b) into (c) and (e) for the next analysis, the operation is a key feature of a true Kalman filter. Meteorologically speaking, it implies that the structure functions and weights used in the analysis are going to depend on the past history of the flow and of the observing network ; in 4D-Var this information is *a priori* not taken into account for the observations near the beginning of the minimization time interval. This is supposed to improve the quality of the analysis, hence it will have an impact on the next $\mathbf{P}^a$, and then on all the subsequent error covariance estimates. One can expect some useful information to be accumulated into the covariance matrices in the long run because of this 'spin-up' effect ; on the other hand, unless some climatological information about the error covariances is used, one may experience slowly growing problems in the error covariances as the assimilation goes on.

In the complete system, one would obtain a coupled system of analyses and forecasts for the model state and for the error covariance matrices, as shown in the diagram in figure 4.

Any implementation of some or all of these components into a high-resolution assimilation system will have to be a compromise between these expectations and the technical constraints. One could argue that a full-fledged EKF will never be implemented operationally despite the increase in computing power, because the models keep improving whereas the theoretical cost of the EKF grows much faster than the complexity of the models. What is important in practice is that we are beginning to have enough computing power to design and implement some approximations to the EKF which are meteorologically useful.

An inspection of the EKF equation reveals that they imply huge matrix operations (multiplications and/or inversions), with some matrices containing of the order of 10^{12} coefficients. Some matrix evaluations can be avoided by using operators such as $\mathbf{M}$ in (b) (the tangent linear forecast operator), or by seeking a variational formulation of the problem, but then one has to face the cost of many short-range integrations of the model (typically 10^6 times) and of even more evaluations of the cost-function of the variational analysis. Also, because of the recurrent nature of the EKF, sooner or later one has to store a covariance matrix, and that creates even worse problems of memory, I/O and data storage requirements.

Because of the lack of experience with the EKF, it is difficult to assess to what extent such costs would be justified by a improvement of the background error covariances. However, common sense suggests that the EKF equations are extremely wasteful because they do not take advantage of the physical nature of the underlying system. First, the covariance matrices are very large because the EKF allows any couple of model variables to have correlated errors ; it is dubious that any meaningful correlation exists between e.g. two synoptic-scale meteorological phenomena located very far from each other. Second, the analysis equations are complex because it is assumed that the whole model state is affected by the observations ; we know from operational experience and low-resolution EKF experiments that there are large data-sparse areas in which it is almost impossible to infer any sensible correction to the background. Third, the forecast equation for the covariances (b) is extremely costly, not only because of the sheer size of the covariance matrices, but also because it is assumed that any error in the initial state may lead to errors in the whole forecast state ; we know that information does not propagate instantly in the real atmosphere, and that localized analysis errors usually remain in a small portion of the model domain during a forecast step (i.e. the group velocity of errors is not too large), at least in a global model : this means that the propagator $\mathbf{M}$ is a relatively sparse operator. To summarize, the EKF assumes that there can be a lot more information in the covariance matrices than we will ever be able to calculate, partly because of the comparatively simple structure of the forecast model and of the observing network, and partly because of the inherent uncertainty in the model design and in the observed data, which dooms any attempt to build error covariance matrices of more than a limited complexity. This was pointed out by Dee (1991), among others.

This is good news in some sense : it means we must try and simplify the EKF algorithm rather than just wait for computers to be powerful enough for a brute-force implementation of a full EKF. Several simplification techniques have been proposed by various authors, and most of them are based on a restriction of the space in which the covariances and/or the tangent-linear model are used :

- The model can be truncated to a low resolution. It implies that the resolved part of the covariances is truncated accordingly. This is technically straightforward, but it is not obvious how optimal this strategy is for meteorological applications ;
- The covariances can be constrained to obey a predefined modelization, in order to obviate the need for handling large explicit matrices. This does not necessarily imply that the model is simplified. It has been proposed to keep the correlations fixed and equal to a simple model, and to compute the evolution of the variances only. Some cleverly defined balance constraints (like geostrophy) between the variables can be used to summarize the information. Perhaps a completely parametric model of the covariances could be designed, but so far only fairly restrictive and unrealistic models have been proposed. If the covariances were expressed as a function of a reduced set of parameters, then it would be possible to rewrite equation (b) in an efficient form which only computes the evolution of those parameters, however this problem is algebraically very difficult (see e.g. Thompson 88), best solved using adjoint techniques.
- It is possible to refine further the definition of the "interesting" subspace of the errors according to what one believes to be important in the assimilation problem. As we will see below, the EKF can be restricted to solve only for the eigenspace of $\mathbf{P}^f$ related to the largest or smallest errors (Ehrendorfer and Tribbia 1996), or for the most unstable (or stable) subspace of the model dynamics : breeding vectors, or singular vectors (Cohn and Todling 1996). For regional prediction, one could as well solve only for the errors affecting an area of particular interest.

It is still unclear how these numerous ideas can be blended into a single efficient system. The scientific issue is not only the nature of the error covariances and of the forecast instabilities (and model errors), we need to account for the way the analysis and forecast steps fit into each other from the point of view of the errors : are the short-range forecast errors dominated by the atmospheric instabilities during the forecast itself (this is what is assumed by the ensemble prediction system at ECMWF), or are they dominated by the analysis errors ? The answer obviously depends on the place, the parameter and the meteorological situation ; it will determine how we should balance the computational effort between the modelization of the analysis error covariances and the atmospheric dynamics. A more technical question, but equally important, is how to design a background error covariance operator for the analysis, which is carried out at the resolution of the operational model, using the information coming from the EKF in some subspace. This requires mixing the EKF covariances with a more static error model. Last but not least, some corrections to the algorithm will have to be implemented in order to account for important missing features : modelization errors, non-linear effects, suboptimality of the analysis, in particular. This question is quite open, and it is similar to the problem of designing physical parameterizations in the early times of NWP.

1.4 Specifications for operational use

From the experience gathered with various meteorological problems indirectly linked to the EKF, it is possible to predefine some characteristics any implementation should have in order to be reasonably realistic. The specifications below are for global weather prediction at a few days' range, with emphasis on the extratropics.

- In equation (b) we want at least that the synoptic-scale atmospheric instabilities are represented. This means that we shall use a tangent-linear model of resolution at least T42, and perhaps even more in dynamically active areas. This is suggested by the spectra of forecast error sensitivity patterns (Rabier et al 1994).
- The dimension of the unstable subspace shall be truncated to no less than 50 or so. This comes from the experience with the ECMWF ensemble prediction system (number of singular vectors with significant growth rates, Buizza et al (1992)).
- the trajectory used for linearization must be as realistic as possible : it can be provided by a high-resolution model with state-of-the art physics, such as those already used for operational weather prediction.
- the tangent linear model shall have a reasonably good physical package ; it is already recognized that surface drag, horizontal and vertical diffusion are absolutely necessary to prevent spurious instabilities from developing (Buizza 1993). In the tropical areas and in stormy weather systems, a representation of water condensation and convection is probably necessary too, this is also important if one wants to improve the analysis of humidity.
- a careful interface must be designed between the EKF model states and error covariances, and the higher-resolution analysis and forecast system used to make the operational predictions. Spectral truncation may not be enough, and it may be necessary to take into account vertical interpolations as well as changes of orography. Even with an incremental technique, we know that the computation of the increments needs to be carried out at truncation T63 at least.
- the current computer technology means that we can aim for an algorithm which involves running and temporarily storing about 500 low-resolution model runs (tangent linear or adjoint) per analysis cycle.

Of course, the final requirement will be that implementing the EKF in the assimilation system shall have a positive impact on the the forecasts ; an impact is expected on both deterministic and stochastic forecasts (i.e. ensemble prediction), since the EKF is supposed to produce not only analyses, but also estimates of the short-range predictability.

The implementation of the EKF in a different framework may lead to different requirements. Depending on the application, one may need to consider additional requirements, such as the simulation of some specific physical phenomena ; on the other hand, the algorithm may be tailored to put the emphasis on some particular observations or forecast areas and parameters, leading to more efficient implementations ; this is discussed later in the section on 'special applications'.

2. THE REPRESENTATION OF ERROR COVARIANCES

In this section we are going to dwell specifically on the problem of representing the error covariances in a tractable yet satisfactory way ; it is a central problem in the implementation of a simplified EKF.

2.1 Design constraints

As covariance matrices, the error covariances must meet some basic mathematical criteria which are quite fundamental for the numerical computation of the analysis step. Some of their characteristics are directly related to well-known features of the atmosphere, and it is obviously important to ensure that they make sense.

- an error covariance matrix (or, more, generally, a covariance tensor) is by definition the expectation of the squared difference between the model vector state and the analogous vector representing the "true" atmospheric state : $\mathbf{P}^f = \overline{(\mathrm{x}^f - \mathrm{x}^t)^2}$. Thus, it needs to be defined for every single variable of the model. If some variables are not in $\mathbf{P}^f$, it means that those variables in x^f will not be corrected by the analysis, or, equivalently, that one assumes that their forecast is perfect (which is usually not true).
- In the actual assimilation algorithm, the simplifications that are made imply that we do not actually use every single coefficient of the covariance matrices, but rather that we need them as an operator $\mathrm{x} \mapsto \mathbf{P}\mathrm{x}$.
- The covariances must be symmetric positive definite. It means that the associated bilinear operator must be a quadratic form. Although this is mathematically obvious, it is important to check that it is still numerically true in the EKF implementation itself.
- In 3D- or 4D-Var, we need the inverse of $\mathbf{P}^f$, rather than $\mathbf{P}^f$ itself. This is special to the variational formulation — other algorithms (like PSAS, or the representer method, see Bennett and Thornburn 1992) use $\mathbf{P}$ directly.
- In a variational analysis, the preconditioning is usually done using the metric defined by the background term. Again, this is special to a particular formulation of the analysis, but

it means that a symmetric square root of $\mathbf{P}^f$ (or its inverse) is required, and computing it may not be trivial.

- The analysis relies on the $\mathbf{P}^f$ covariances themselves, and also on their mapping into covariances with the observed variables ($\mathbf{P}^f\mathbf{H}^{\mathrm{T}}$), which is how the structure functions of the analysis are built. Hence it is good practice to ensure that the background error formulation is reasonable in terms of all the variables in the model and in the observations.[1]
- The diagonal of $\mathbf{P}^f$ defines the background standard errors, in other words the weight of the background field in the analysis. It is a fundamental feature of the analysis which needs to be realistic in all the relevant spaces (gridpoint and spectral).
- Similarly, the $\mathbf{P}^f$ correlations imply the shape of the structure functions, and the filtering properties of the analysis in data-rich areas. If they are sharp, isolated observations will generate only a very localized increment, hence the observed difference with the background will only have a limited impact on the analysis. If they are broad, there will be a lack of small-scales structures in the increments over data-rich areas, and possibly spurious increments generated by distant observations over data-poor areas.
- The multivariate cross-correlations can be interpreted as an effective balance constraint on the increments. The hydrostatic and geostrophic equilibria are well-known, but there are certainly other complicated relationships in the atmosphere, and it is hoped that the EKF will allow for a good representation of them, thus providing a powerful reconstruction of meteorological structures from limited observing systems. The ability to reconstruct baroclinic waves can been clearly demonstrated using the equivalence with 4D-Var (Rabier and Courtier 1992).

2.2 Static error covariances

The most basic component of an error covariance modelization is its time- and space-averaged component, which can be regarded as the 'climate' of the errors. This 'static' part depends both on the model and on the observing system used. It provides a reference against which to check more sophisticated models of covariances.

Static covariances are about all we can calibrate using objective statistics on real data. The so-called "Hollingsworth/Lönnberg method" (Hollingsworth and Lönnberg 1986, Lönnberg and Hollingsworth 1986) is a practical calibration algorithm in which there are assumptions of uncorrelated observation errors, no biases, and homogeneity of the errors in space or time. It only allows one to calculate convincing statistics on rather homogeneous and isotropic components of the error covariances. Being based on histograms of empirical covariances, it requires a lot of data, so that it is intrinsically limited by the size of existing datasets of observation departures.

[1] For instance, the background error variances are artificially reduced when observations are interpolated far from the model grid, whereas one would expect just the opposite : a larger background variance because of interpolation errors. This comes on top of the representativeness errors in the observation operator.

A more flexible (and very popular) method is called the "NMC method" for historical reasons (Parrish and Derber 1992). It obviates the need for handling observations, and provides error statistics directly in model space. However, it relies on a fundamental hypothesis (that differences of forecasts valid at the same time have the same correlations as short-range forecast errors) which has no rigorous justification, and has only been tested against the Hollingsworth/Lönnberg method in a few cases. This means no-one knows the limits of its validity, so it can be used safely only to calibrate average characteristics of the covariances, e.g. global spectra or domain-averaged correlations.

Statistical methods are technically cumbersome and they are always restricted by the amount of data necessary to reduce the sampling errors to an acceptable level. An alternative approach is to specify covariance models externally, using theoretical arguments. A variety of comprehensive covariance models have been derived by several authors (Phillips 1986, Bartello and Mitchell 1992, Balgovind et al 1983) using hypotheses of equipartition of energy in various senses. Although they provide a useful reference framework to understand some properties of the real covariances, they have not had much practical application in NWP so far.

A last method would be to use an EKF in research mode with a convenient configuration, so that the time-averaged covariances of the EKF could be identified with those of a genuine assimilation system. Some components of the EKF would necessarily contain some arbitrariness (the model error term and the initial covariances in particular), but one can hope to use this technique for improving the static covariances beyond the inherent limits of the statistical methods. The additional information would come from the direct representation of the observing network and the model dynamics. For instance, a low-resolution EKF could be useful to calibrate the very large-scale covariances, for which the sampling is very poor.

The current objective knowledge of the static error covariances reduces to the following :

- Maps of forecast standard errors of all fields,
- Error covariance spectra, i.e. the homogeneous isotropic part of the horizontal correlations,
- Vertical error covariance matrices for a given scale (see figure 5) or for a given geographical domain (see figure 6).
- Balance diagnostics (mainly a verification of the geostrophic balance in the extratropics), with some dependency on the horizontal and vertical scales.

This has actually been the basis of operational NWP for years ; how it can be used to design a coherent background error covariance model is described in detail in Rabier et al (1996).

The implementation of such ideas in a covariance operator relies on its splitting into a sequence of simple operators. Variances can be separated from correlations matrices ; the multiplication by standard errors is a diagonal operator. Large correlation matrices can be built economically by assuming that they are tensor products of simpler correlation models, e.g. for

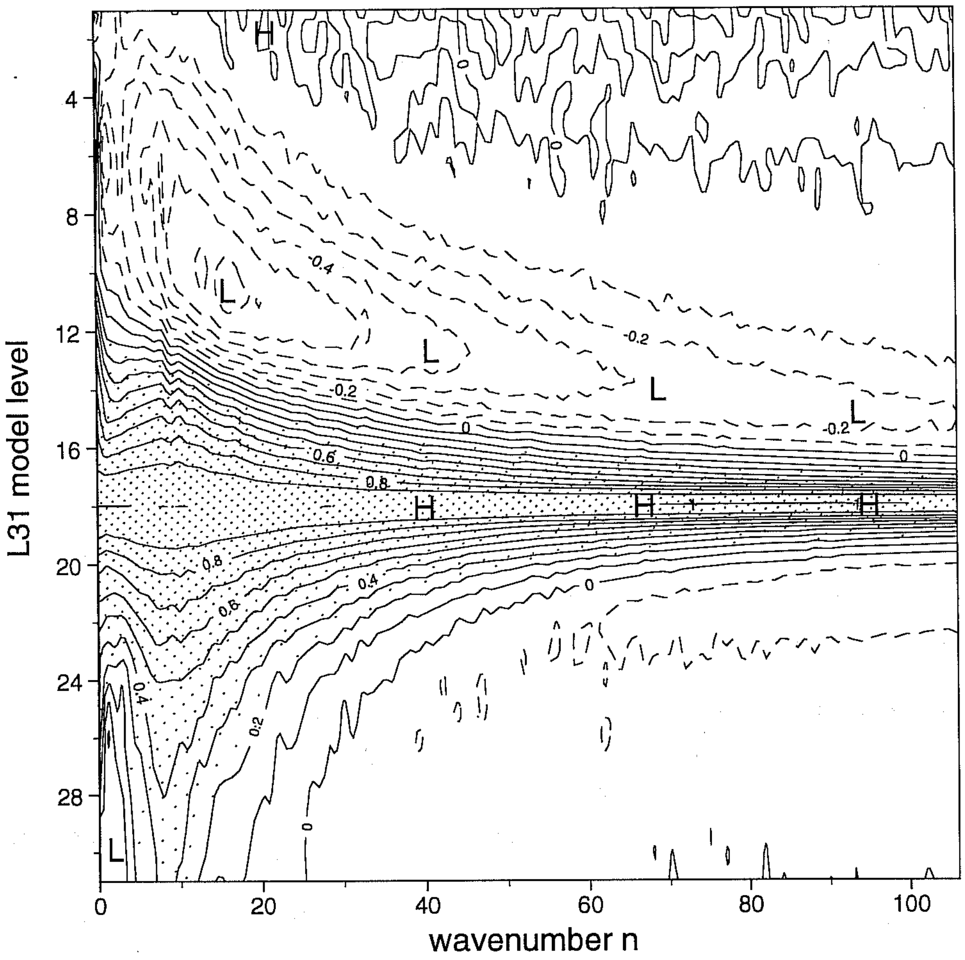

Figure 5: Vertical autocorrelation of temperature forecast errors with the temperature at 500hPa, diagnosed by the NMC method, as a function of total wavenumber, in the ECMWF operational model. Notice how the correlations are sharper for smaller scales.

balanced and unbalanced variables, or for horizontal and vertical separation, or for vertical correlations on different scales. This usually restricts the number of degrees of freedom in the correlation model, e.g. balanced and unbalanced errors are usually assumed to be uncorrelated to each other. The remaining correlations are then simple enough to be implemented : vertical correlations matrices and horizontal correlation spectra can be represented explicitly. Horizontal correlations can be implemented as cheap numerical filters (Lorenc 1992).

Several covariance models can be mixed into a single one, by a linear combination of projections of different covariance matrices, the projectors reflecting changes in the behaviour of the variables. This can be used to manage dynamically balanced versus unbalanced parts of the errors, correlations changes with geographical area, or, as we will see below, subspaces in which the covariances are supplied by an external source like the EKF. Mathematically, if we denote p_1 the orthogonal projector into a subspace H_1 in which we assume that the covariance matrix obeys a model $\mathbf{P}_1$, whereas another model $\mathbf{P}_2$ applies to the orthogonal space, the combination of those models has the form

$$\mathbf{P} = p_1^{\mathrm{T}}\mathbf{P}_1 p_1 + (\mathbf{I} - p_1)^{\mathrm{T}}\mathbf{P}_2(\mathbf{I} - p_1)$$

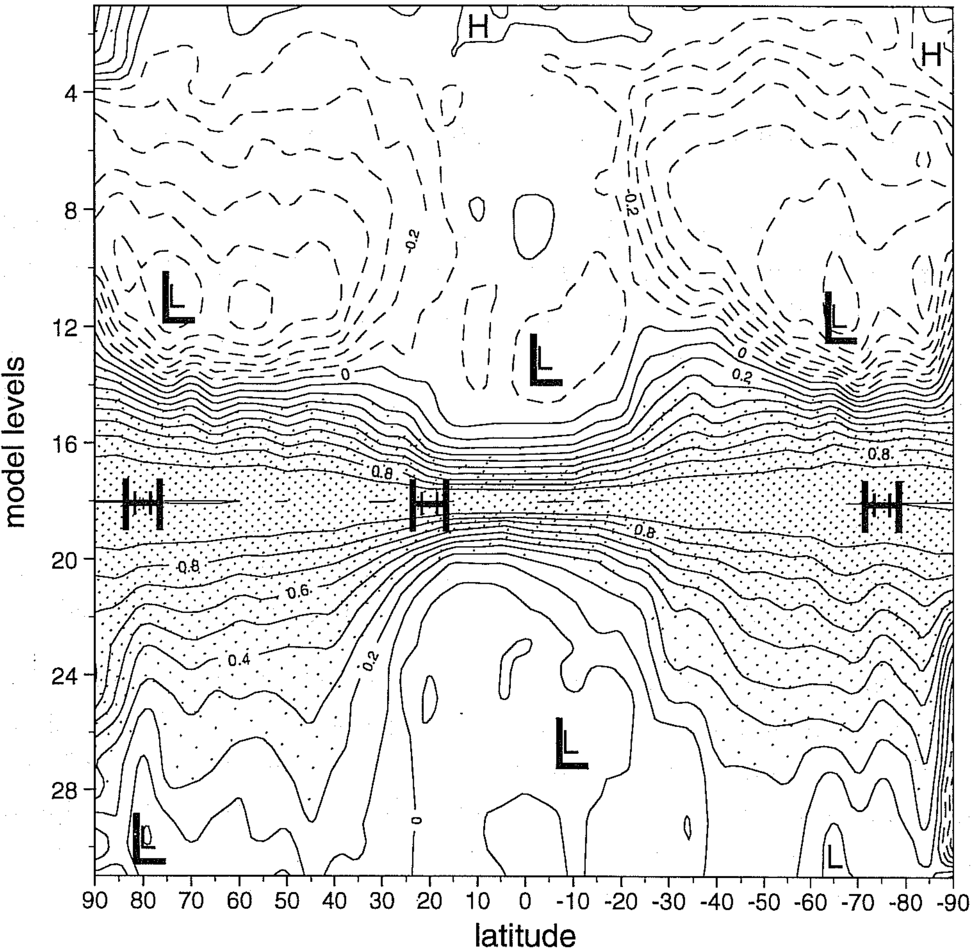

Figure 6: Vertical autocorrelation of temperature forecast errors with the temperature at 500hPa, diagnosed by the NMC method, as a function of latitude, in the ECMWF operational model. Notice how the correlations are sharper in the tropical regions.

which implies that there is no cross-correlation between spaces H_1 and $(H_1)^{\perp}$; this may have some non-trivial consequences on the structure functions of the analysis.

As an example we can briefly recall the 3D-Var J_b formulation in operational use at ECMWF since February 1996 (Courtier et al 1991, Rabier et al 1996) : the inverse of the cost-function is split into three components defined by the Hough mode balance. Each component A is defined by its symmetric square root S, itself built as the multiplication D by a standard error field, the tensor product of horizontal C_h and vertical C_v correlation operators (as in fig. 5), and a set of ad hoc weights W to penalize the geostrophically unbalanced components of the increments. This can be loosely summarized by the following equations :

$$\begin{aligned}\mathbf{P}^f &= (A_{bal}^{-1} + A_{unbal}^{-1} + A_{univ}^{-1})^{-1} \\ A &= S^{\mathrm{T}} S \\ S &= W C_v C_h D\end{aligned}$$

The implied covariance model is completely static, except for the background standard errors which are calculated from the previous analysis errors using a simple relaxation towards climatology.

2.3 Situation-dependent covariances : the remapping technique

The "static" covariances are essentially stationary in time, except perhaps for a slow variation to account for the seasonal variations. Few attempts have been made so far to include more flow-dependency into the covariances without going to a sophisticated method such as 4D-Var or the EKF. The only widely implemented feature is the variation of the standard errors with the observing network, because this is an unexpensive by-product of OI. Some authors have advocated the use of air mass-dependent correlation models, because it is obvious that spatially averaged structures are not well suited to meteorologically important systems like fronts. Although this sounds like an obvious thing to do, the local modification of a covariance model is very difficult because of all the constraints to meet simultaneously : smoothness of the structures, positive definiteness of the resulting covariance matrix and balance constraints.

A possible solution to these problems is a technique we will call 'remapping'. The idea is that if we are not able to build a complex covariance model, we can start from a simple covariance matrix, and distort it conveniently using a geometry transform which defines a remapping between the model space and the space in which the covariance model is built, This allows one to translate, stretch and magnify at will the structure functions, provided the remapping operator is smooth enough. Mathematically this can be written as

$$\mathbf{C}_2 = \mathbf{U}\mathbf{C}_1\mathbf{U}^{\mathrm{T}}$$

where $\mathbf{C}_1$ is a simple covariance model (e.g. homogeneous and isotropic), and $\mathbf{U}$ is a geometry transform, or "remapping" operator. This implies a distortion of the standard errors which is linked in a simple way to the local map factor of $\mathbf{U}$. The resulting correlation matrix $\mathbf{C}_2$ may be algebraically much more complex than $\mathbf{C}_1$. The idea of remapping errors was initially suggested as a diagnostic tool called the "distortion representation" (Hoffman 1995). It has been tested at Météo-France using the the Schmidt transform (Moll and Bouttier 1996) and semi-geostrophic coordinates (Desroziers 1996). At ECMWF, this is being tried for local corrections to the vertical correlation structures in the tropics, as suggested by figure 6.

3. IMPLEMENTATION OF A SIMPLIFIED EKF

In this section we are discussing a possible algorithm for the implementation of a simplified Extended Kalman Filter, on top of an assimilation system which uses a variational 3D-Var analysis. This is the case of the ECMWF operational assimilation system. The algorithm would be the same with an intermittent 4D-Var assimilation.

3.1 Analysis error covariances

The estimation of the analysis error covariances with an EKF-style method has been implemented operationally at ECMWF in September 1996 for the estimation of the analysis standard errors. It uses the algorithm described below, developed by Fisher and Courtier (1995).

The estimation of $\mathbf{P}^a$ in a 3D-Var analysis can be made using the equivalence between the Hessian of the cost function J (i.e. the matrix of its second derivatives) and $(\mathbf{P}^a)^{-1}$ (for a demonstration of the equivalentce see e.g. Barmeijer et al (1996)). The Hessian operator can be obtained from finite differences of the gradient ∇J, assuming J is quadratic. Thus, we use a quadratic approximation to the 3D-Var analysis, which is believed to be good enough for this application :

- the analysis is incremental, i.e. we are considering a low-resolution EKF (T42L31),
- the observation operators are linearized,
- the variational quality control of the observations is switched off,
- we use a close approximation $\mathbf{P}^f$ to the background term J_b of the operational analysis, so that we can use it in a factorized form : $(\mathbf{P}^f)^{-1} = \mathbf{L}^{\mathrm{T}}\mathbf{L}$.

The incremental 3D-Var relies on the definition of the analysis increments as the solution of an optimization problem :

$$\mathrm{x}^a = \mathrm{x}^f + \arg\min J(\delta\mathrm{x})$$

where the cost function is written in terms of the departures $\delta\mathrm{x}$ from the background, and it includes a background term which is involves the assumed background error covariances $\mathbf{P}^f$:

$$\begin{aligned} J(\delta\mathrm{x}) &= J_b(\delta\mathrm{x}) + J_o(\delta\mathrm{x}) \\ &= \delta\mathrm{x}^{\mathrm{T}}(\mathbf{P}^f)^{-1}\delta\mathrm{x} + J_o(\delta\mathrm{x}) \end{aligned}$$

The cost function may be rewritten as a function of a variable χ defined using the operator $\mathbf{L}$ defined above :

$$\begin{aligned} \chi &= \mathbf{L}\delta\mathrm{x} \\ J(\chi) &= \chi\chi^{\mathrm{T}} + J_o(\mathbf{L}^{-1}\chi) \end{aligned}$$

The second derivative of the function $J_\chi = J \circ \mathbf{L}^{-1}$ is connected to the analysis error covariance matrix in terms of variable χ :

$$\begin{aligned} \frac{1}{2}J''_\chi &= \mathbf{I} + \mathbf{L}^{-\mathrm{T}}\mathbf{H}^{\mathrm{T}}\mathbf{R}^{-1}\mathbf{H}\,\mathbf{L}^{-1} \\ &= (\mathbf{P}^a\chi)^{-1} \end{aligned}$$

The proposed simplification of this EKF step is to assume that most of the useful information in $\mathbf{P}^a\chi$ can be summarized the eigenspace $\{\mathrm{v}_i\}$ associated to its largest eigenvalues $\{\lambda_i\}$, and by the identity matrix $\mathbf{I}$:

$$\frac{1}{2}J''_\chi \simeq \mathbf{I} + \sum_i(\lambda_i - 1)\mathrm{v}_i\mathrm{v}_i^{\mathrm{T}}$$

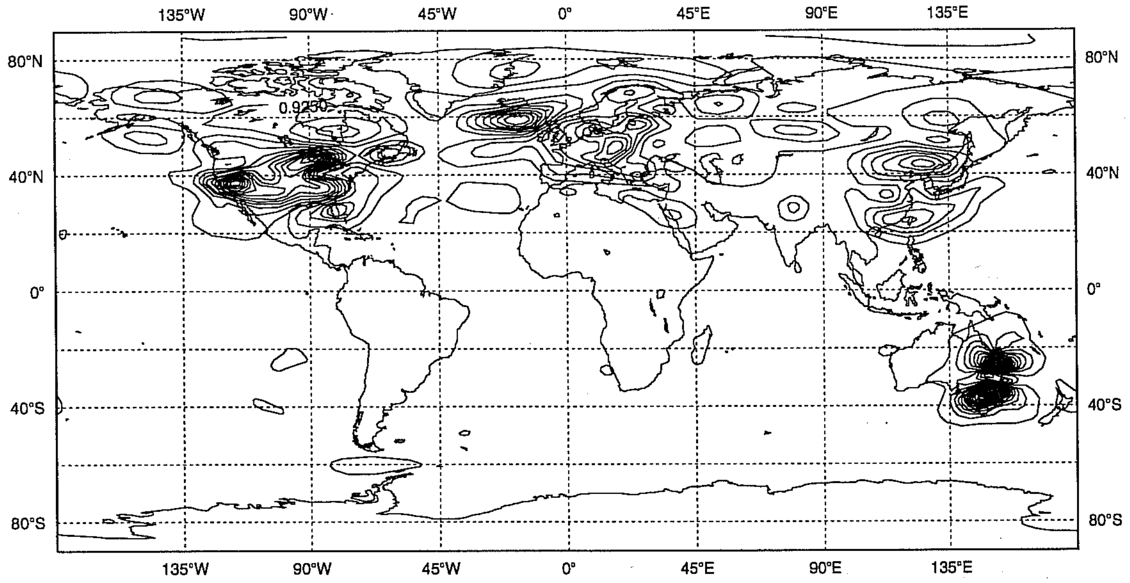

Figure 7: Map of the differences between the background and analysis standard errors of temperature at 500hPa in the ECMWF analysis of 21 Nov 1996, 12h. The difference has been computed using 30 eigenvectors with a Lancoós method on the Hessian of the 3D-Var analysis, as explained in the text.

The term **I** in χ space is equivalent to the background error covariances $\mathbf{P}^f$ in physical space (where the variable is δx). The other term represents an approximation of the information brought by the observation term J_o : all the eigenvalues λ_i are strictly positive, which is consistent with the fact that observations cause the estimation errors to decrease. The largest eigenvalues are expected to be associated to the error patterns which are best observed, i.e. to the most important differences between $\mathbf{P}^f$ and $\mathbf{P}^a$, in other words, to the places where the analysis error variances are going to be the smallest.

The approximation in this step of the EKF consists of computing only a small number of eigenvectors $(\mathrm{v}_i, \lambda_i)$. A Lanczós-type numerical method is used to compute efficiently a few eigenvectors associated to the largest eigenvalues. The quality of the convergence towards the the exact $\mathbf{P}^a$ matrix is linked in a non-trivial way to the steepness of its eigenspectrum, but numerical experimentation at ECMWF has demonstrated that very convincing error patterns are obtained when one computes about 30 eigenvectors (figure 7). It is yet unclear how many eigenvectors would be required to obtain a very good approximation of the complete analysis error matrix. We compute only a fraction of the eigenvalues of $J''_\chi/2$, and we approximate the rest by 1, which is an underestimation of the true values. Therefore, truncating the eigenspectrum leads to overestimating the analysis variances[2], and to correlation structures which are too close to those assumed in J_b. The variance estimate is nevertheless optimal given the constraint of solving only for an eigenspace of limited dimension (Ehrendorfer and Tribbia 1996).

[2] Actually the change of variable does not have to be the square root of J_b, it could be any preconditioner that brings the Hessian close to the identity. The use of **L** in this presentation simplifies the algebra, and guarantees that the most important eigenvectors in χ space are associated to the largest eigenvalues. With a different preconditioner, the smallest eigenvalues may turn out to be important, and then comes the problem of distributing the computational effort between the determination of the largest and the smallest eigenvalues.

In order to obtain an unbiased estimate of the true analysis error variances, it is necessary to reduce empirically those provided by the method. Other empirical corrections are necessary, as explained later in this paper.

3.2 Forecast error covariances

The estimation of the forecast error covariances has been examined indirectly in the context the generation of optimal perturbations for ensemble prediction systems, or EPS (Buizza et al 1992, Toth and Kalnay 1993) ; EKF experimentation with simplified models (Houtekamer 1993, Cohn and Todling 1996) demonstrates that indeed it seems to be the right framework for an efficient simplification of the EKF. Like in the analysis step, one needs to pay attention to the most significant eigenvectors of the error covariance matrices. They are believed to contain the most important flow-dependent structure functions needed for the analysis, this is clearly supported by 4D-Var experiments (Thépaut et al 1993).

In the forecast equation for covariances, (the time indices have been dropped for the sake of clarity)

$$(b) \qquad \mathbf{P}^f = \mathbf{M}\mathbf{P}^a\mathbf{M}^{\mathrm{T}} + \mathbf{Q}$$

we need on input $\mathbf{P}^a = \frac{1}{2}\mathbf{L}\, J''^{-1}\mathbf{L}^{\mathrm{T}}$, not its inverse, so there is the problem of inverting the Hessian J'' of the variational analysis. This may be extremely costly. On output, the variational analysis needs for its J_b term the operator $(\mathbf{P}^f)^{-1}$, not $\mathbf{P}^f$, so that there is again a problem of matrix inversion. The latter problem disappears if an algorithm like PSAS is used, but this is not currently the case.

In this step of the EKF we make, again, the approximation that the useful information in $\mathbf{P}^f$ can be summarized in a subspace of a small predefined dimension. The rms error of the approximation is minimized if the subspace is the one associated with the largest eigenvalues of $\mathbf{P}^f$; this is a classical algebraic result (Ehrendorfer and Tribbia 1996) which reflects the intuitive fact that only a limited number of weather systems are active at a given time in the atmosphere, giving rise to a comparable number of notable forecast error patterns, as in figure 1. There is some suspicion that very stable structures (i.e. those associated with the smallest eigenvalues) may also be worth taking into account, because they point to areas where the forecast is supposed to be good, but they are probably difficult to handle, for instance because we know so little about the structure of modelization errors.

In the following discussion we will temporarily drop the model error term $\mathbf{Q}$ from eq. (b). On top of the inversion problems outlined above, there is the issue of the representation of $\mathbf{P}^a$. If there is an approximation in $\mathbf{P}^a$ itself, it should be such that it has a minimum impact on $\mathbf{P}^f$. There are two possible strategies which are outlined below.

Integrated solution. The first option, which can be called the "integrated" approach, is to get rid of the $\mathbf{P}^a$ problem and to concentrate on $\mathbf{P}^f$ as an single operator encompassing simultaneously the analysis and the forecast steps. With a little algebra one can rewrite the

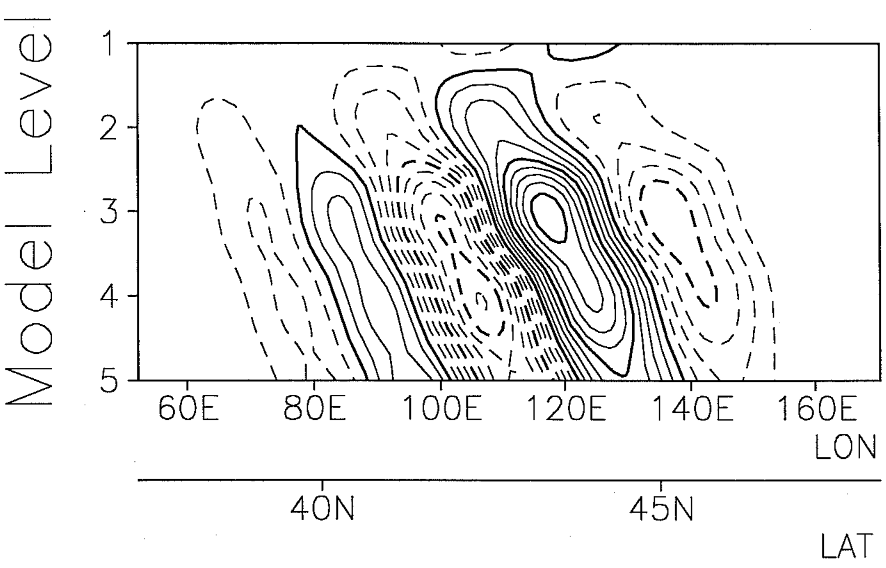

Figure 8: Vertical cross-section of the largest eigenvector found in $\mathbf{P}^f$ using a T21L5 approximation of the ECMWF 3D-Var analysis and a Jacobi-Davidson method (data provided by J. Barkmeijer, personal communication).

following eigenvalue problem on $\mathbf{P}^f$:

$$\mathbf{M}\mathbf{P}^a\mathbf{M}^\mathrm{T}x_i = \lambda x_i$$

as a generalized eigenvalue problem :

$$\mathbf{M}^\mathrm{T}\mathbf{M}y_i = \lambda_i(\mathbf{P}^a)^{-1}y_i \quad \text{with } x_i = \mathbf{M}y_i$$

It is interesting to note the similarity of the equation above with the determination of the singular vectors of a short-range forecast, and there is indeed a large similarity between the EKF and EPS problems ; more comments on that subject can be found in Bouttier (1996). The important feature in the above form is that the inverse of $\mathbf{P}^a$ is now used, so it can be directly replaced by the Hessian of the analysis cost-function. This obviates the need for approximating $\mathbf{P}^a$. The determination of a few eigenvectors of this problem can be computed efficiently using methods like the Jacobi-Davidson algorithm ; the eigenvectors have the expected baroclinic structure in dynamically unstable regions (see figure 8, and the contribution of T. Palmer in this volume). This technique is much more expensive than a Lanczós algorithm for the same number of eigenvectors, because there is an approximate inversion of the $(\mathbf{P}^a)^{-1}$ operator for each trial vector (J. Barkmeijer et al 1996). Thus, the cost of this way of looking for eigenvectors of $\mathbf{P}^f$ may well be dominated by that of evaluating gradients of the cost function, especially if there are a lot of observations going into the analysis. It may sound inefficient to use a lot of observations if the EKF computations are done at a relatively low resolution, but unfortunately we do not know how to suitably approximate the structure of the observing network, although there is some suspicion that the impact of observations on the error covariances may be quite simple in data-rich areas (see figure 2).

Two-step approximation. An alternative method consists of separating the analysis and forecasts steps, so that we never have to solve a generalized eigenvalue problem. The first step is simply the EKF approximate analysis described in the previous section ; the eigenvector-based approximation of $\mathbf{P}^a$ is well-suited for exact inversion using the SWM (Sherley-Woodbury-Morrisson) formula, which is very cheap :

$$\begin{aligned}
\frac{1}{2}J''_\chi &= \mathbf{I} + \sum_i (\lambda_i - 1)\mathrm{v}_i\mathrm{v}_i^{\mathrm{T}} \\
\mathbf{P}^a_\chi(\frac{1}{2}J''_\chi)^{-1} &= \mathbf{I} + \sum_i (\frac{1}{\lambda_i} - 1)\mathrm{v}_i\mathrm{v}_i^{\mathrm{T}} \\
\mathbf{P}^a &= \mathbf{L}^{-1}[\mathbf{I} + \sum_i (\frac{1}{\lambda_i} - 1)\mathrm{v}_i\mathrm{v}_i^{\mathrm{T}}]\mathbf{L}^{\text{-T}}
\end{aligned}$$

As one can see, $\mathbf{P}^a$ is conveniently expressed as a chain of simple operators, so that the eigenproblem $\mathbf{M}\mathbf{P}^a\mathbf{M}^{\mathrm{T}}x_i = \lambda x_i$ can be solved using a classical Lanczós method, at a price dominated by the cost of running the adjoint and tangent linear models.

For a given number of eigenvectors of $\mathbf{P}^f$, this is much cheaper than the integrated method, but there is no guarantee that the approximation made on $\mathbf{P}^a$ will be optimal ; there is no proof that the short-range forecast instabilities are unaffected by the use of approximations on $\mathbf{P}^a$ in the orthogonal of the $\{\mathrm{v}_i\}$ space. There is not even a proof that the result will converge to the exact eigenvectors as the number of $\mathbf{P}^a$ eigenvectors is increased, because the set of the largest eigenvectors is a non-continuous function of the coefficients of the operator.

On the other hand, one may argue that the lower cost of the method will allow for a much larger set of v_i to be determined (assuming vector storage is not a technical issue), and that may offset the approximate character of the method. Moreover, it may make sense to compute more eigenvectors of $\mathbf{P}^f$ in order to improve the J_b structures in the next analysis, and that in turn may benefit more to the quality of the next $\mathbf{P}^f$ than the sophistication of the integrated method. The fact is, it is almost impossible to answer those questions without numerical experimentation, because we know so little about the relative structures of analysis and forecast errors. This can be appreciated only in a framework that includes both a realistic model and a comprehensive observing network. Determining the number of eigenvectors to compute and their resolution is another problem to solve. However, an encouraging aspect of this question is that it brings the hope of merging the EKF with the computation of singular vectors for the EPS, so that it can rely on more computer resources.

Like in the analysis step, the estimated $\mathbf{P}^f$ needs to be corrected. First, we estimate $\mathbf{P}^f$ only in a small eigenspace, and it is necessary to provide something in the orthogonal in order to have a well-defined covariance matrix ; this is discussed in the next section. Then, something needs to be added to represent the model error term $\mathbf{Q}$, and a correction must be made in order to account for non-linearities ; this usually leads to reducing the estimated variances to reflect the non-linear saturation of large-amplitude perturbations (Bouttier 1994). The estimation of $\mathbf{Q}$ is quite difficult, because it is obviously cannot be calibrated (even on the average) using the model alone ; indirect methods must be used, such as cross-validation with independent error diagnostics (figure 9), comparison with ensemble predictions using non-linear forecasts with physics, and possibly model-error estimates given by appropriate assimilation algorithms such as the representer method (Bennett and Thornburn 1992) or 4D-PSAS (Courtier 1996).

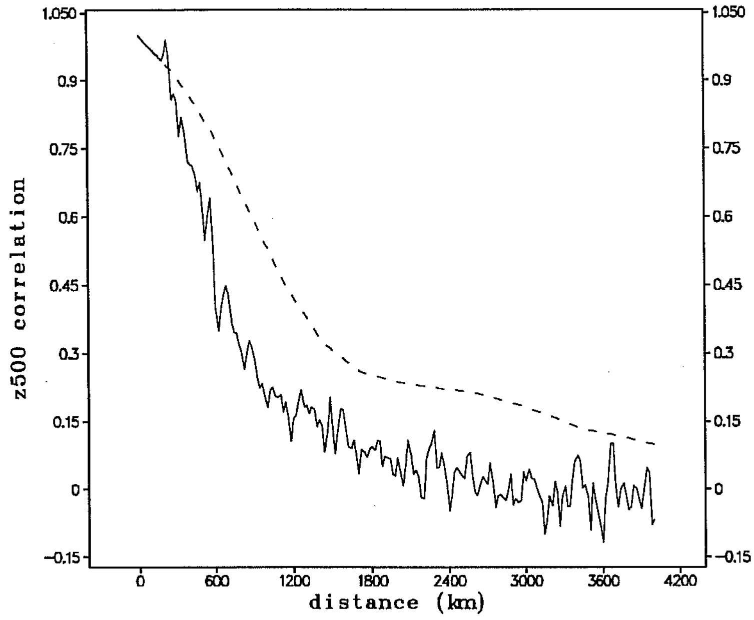

Figure 9: Average radial structure of the background error autocorrelation of the 500hPa geopotential in an experimental EKF (dashed line) and as diagnosed using the Hollingsworth/Lönnberg method (plain line) on the same couples of points. In this case the EKF forecast error covariances are too broad because the model error has been neglected ; model error is expected to have a maximum on the scales at which the model is truncated, giving rise to sharper correlations than implied by the model dynamics alone.

3.3 Background error covariances

In this last step of the EKF we want to take advantage of the approximate yet costly information gathered on $\mathbf{P}^f$ in the analysis. It is clearly not sufficient to feed the analysis with covariances in a small subspace, because the actual background errors need not be confined to that subspace, it is necessary to merge that information with a more comprehensive background error covariance model. For the sake of simplicity, we shall only discuss the mixing of "exact" EKF-generated covariances in an eigenspace $[\mathrm{v}_i]$ with a "static" preexisting $\mathbf{P}^f_s$ model. It is also possible to mix with covariance information which is not in an eigenspace (like singular vectors of the subsequent forecasts, or covariances with a given model parameter or observation), but this is technically more complicated.

The crux of the operation is to only mix covariances defined in subspaces which are mutually uncorrelated ; this guarantees that the merged structure functions will be seamless. For instance, if one were to mix two covariance models defined on complementary geographical regions, it would not be acceptable to build a block-matrix covariance model based on a projection onto each region, because it would imply that neighbouring points separated by the domain boundary would have zero correlation, hence generating discontinuities in the structure functions. If both covariance models are defined in a more complicated way (namely, in some predefined subspaces), similar problems may arise in a more subtle way.

The matrix $\sum_i \lambda_i \mathrm{v}_i \mathrm{v}_i^{\mathrm{T}}$ defines covariances in an eigenspace $[\mathrm{v}_i]$ of the "real" $\mathbf{P}^f$: by construction it is uncorrelated to the orthogonal space (eigenspaces of a diagonal matrix are orthogonal to each other). However, these spaces are a priori correlated in the $\mathbf{P}^f_s$ model, so there may be some distortion of the structure functions when the covariance models are merged ; experimentation is needed to see whether this is really a concern. Since the concept of orthogonality is just a property of the particular metric used, this little problem can be evacuated by a suitable change of variable : with the square root $\mathbf{L}_s$ of the static covariances $\mathbf{P}^f_s$, the static background term becomes the identity :

$$\begin{aligned} \chi &= \mathbf{L}_s \delta \mathrm{x} \\ (\mathbf{P}^f)^{-1} &= \mathbf{L}_s^{\mathrm{T}} \mathbf{L}_s \\ \mathbf{P}^f_{s\chi} &= \mathbf{I} \end{aligned}$$

If the eigenvalue problem for the EKF prediction of error covariances has been solved in χ space, rather than in $\delta\mathrm{x}$ space, then we have by construction zero correlations between the space spanned by $\{\mathrm{v}_i\}$ and its orthogonal, simultaneously in the EKF forecast error covariance and in the static covariance model. This is a consequence of the choice of metric, and it allows a completely seamless integration of the EKF covariances. Moreover, the resulting covariance model is readily invertible and factorizable in a form suitable for the next 3D-Var step, as demonstrated in the derivation below (p is the projector onto the subspace spanned by the v_i's) :

$$\begin{aligned} \mathbf{P}^f &= (\mathbf{I} - p)^{\mathrm{T}} \mathbf{P}^f (\mathbf{I} - p) \; + \; \sum_i \lambda_i p^{\mathrm{T}} \mathbf{L}_s^{\mathrm{T}} \mathrm{v}_i \mathrm{v}_i^{\mathrm{T}} \mathbf{L}_s p \\ &= \mathbf{L}_s^{-1} [I + \sum_i (\lambda_i - 1) \mathrm{v}_i \mathrm{v}_i^{\mathrm{T}}] \mathbf{L}_s^{-\mathrm{T}} \\ (\mathbf{P}^f)^{-1} &= \mathbf{L}_s^{\mathrm{T}} [I + \sum_i (\frac{1}{\lambda_i} - 1) \mathrm{v}_i \mathrm{v}_i^{\mathrm{T}}] \mathbf{L}_s \\ &= \mathbf{L}^{\mathrm{T}} \mathbf{L} \end{aligned}$$

with the new preconditioner :

$$\mathbf{L} = [I + \sum_i (\frac{1}{\sqrt{\lambda_i}} - 1) \mathrm{v}_i \mathrm{v}_i^{\mathrm{T}}] \mathbf{L}_s$$

which is exactly what is needed to run the next analysis step. The change of metric, however, may cause the calibrated $\mathbf{P}^f$ eigenvectors to be less useful as singular vectors for the EPS than those calibrated directly in physical space : this remains to be seen.

3.4 Covariance parameterization

It has been mentioned several times in this paper that the EKF-based algorithms need some empirical modifications in order to obtain realistic covariances, because of the intrinsic assumptions behind the EKF. In the $\mathbf{P}^a$ matrix provided by the analysis step, we need to account for the approximations in the analysis algorithm itself, as well as for weaknesses in the definition of the background error covariance matrix. In the $\mathbf{P}^f$ matrix provided by the forecast step, we need to account for modelization and linearization errors. In both steps there is also a need to correct for biases in the model state or in the observations : biased errors can distort considerably the computed covariances.

More precisely, in the analysis step we have mentioned that the $\mathbf{P}^a$ variances need to be reduced because they are overestimated with the eigenvector approximation, as it has been pointed out already. On the other hand, the analysis itself is suboptimal : there are approximations in $\mathbf{P}^f$, in $\mathbf{R}$, and we have approximated 3D-Var itself. This induces additional errors in $\mathbf{P}^a$, which mean extra analysis error variances, particularly in areas where they were supposed to be very low. As for the correlations, we know that they tend to be too smooth in the static part of the background error term, and they are likely to be reflected into $\mathbf{P}^a$, so they should be sharpened. This is because the static part of $\mathbf{P}^f$ usually stems from globally averaged statistics, which are not representative of the error structures in dynamically active areas ; the resulting $\mathbf{P}^a$ is particularly detrimental to the calibration of singular vectors of the subsequent forecast, because singular vectors precisely point to unstable areas, as demonstrated in Barkmeijer et al (1996). On the other hand, there may be correlated biases in the observations or in the background term (notably in the assumed balance) ; this implies that some analysis error correlations may be broader than they seem.

In the forecast, the concerns are very similar : in $\mathbf{P}^f$, the variances should be reduced in order to account for non-linear error saturation, while at the same time they should include a contribution of model error, due to e.g. missing physics or numerical truncation. This could be checked against a non-linear ensemble forecast (Evenssen and van Leeuwen 1995). The correlation may need to be sharpened because model design weaknesses suggest that the model cannot be trusted for predicting error wave propagation over large distances (Bouttier 1994). At the same time, failure to simulate large-scale circulations, like the Hadley cells, implies broad model error correlations. In order to increase the realism of $\mathbf{P}^f$, one should use the best possible static covariance model, preferably including some flow-dependency.

Because the EKF steps connect into each other, one expects a long-term behaviour to build up after a few assimilation cycles (fig. 10), and it is essential to watch out for any slow-growing undesirable features. This means checking for the boundedness, symmetry and positiveness of the covariance matrices, taking care of the dependence of the covariances to respect to the initial condition (as in Bouttier 1994), and examining the "climate" of the EKF covariances : it should make physical sense, and be consistent with the static model used.

3.5 Special applications

The above description of a simplified EKF is biased towards a particular class of applications : providing global operational analyses and forecasts. Here we review briefly some other important problems which give rise to different EKF implementation issues, and to other interesting applications as well.

- In a limited area model, the implementation is more complicated because the boundary

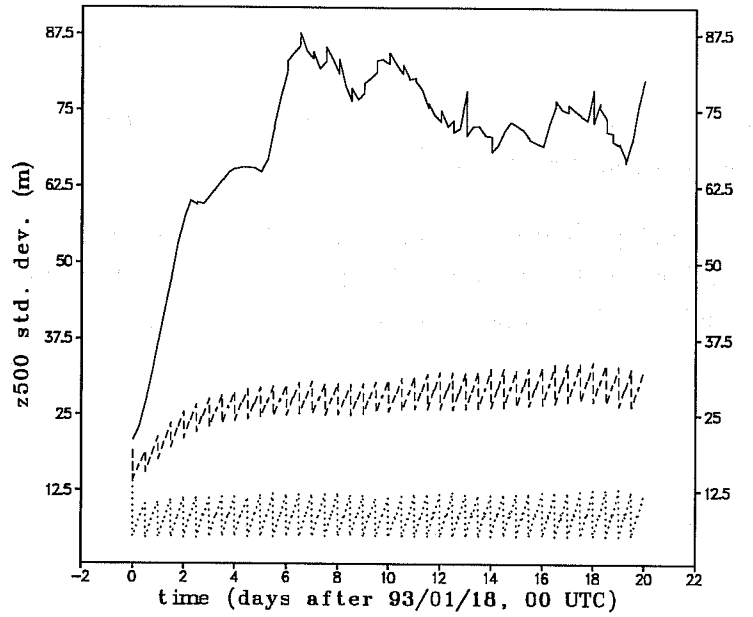

Figure 10: Evolution of the maximum, averaged and minimum of the global 500hPa height standard error field in a simplified EKF. After a spin-up period of a few days, a quasi-static state is achieved, over which the climatology of the standard errors should be consistent with that of the real ones.

forcing contains some errors, and the corresponding error covariances must be accounted for correctly. However, applying the EKF in this framework could provide some useful information about the way the predictions inside the domain depend on the boundary forcing at some earlier time (in the EKF it is straightforward to derive covariances between features of the analysis and of the forecast, the cross-correlations are given by $\mathbf{MP}^a$). The resolution and physical package in the tangent linear model must be realistic enough to provide a good simulation of the mesoscale phenomena which are important in the domain considered, e.g. orography, breezes or thunderstorms, to name but a few.

- In an observing system simulation experiment (OSSE), the informativity of the observations is usually evaluated using impact experiments. This tends to be costly and case-dependent. In the EKF, the knowledge of the analysis error covariances allows one, in principle, to evaluate directly the change in the analysis quality using the error variances, and that is supposed to summarize the impact of observation changes with all the possible error patterns in the model state, so it is presumably more objective and reliable than a small number of impact experiments. The technique has been applied to new observing system by Gauthier et al (1992) and Cohn and Parrish (1991).

- When making predictions for a particular area, it often happens that the atmospheric situation is dominated by one important phenomenon. Adjoint techniques allow one to compute the sensitivity of the prediction of parameters of interest with respect to the initial state (Hello et al 1996), and provide some guidance to the forecaster. They can be imbedded within the EKF so that the computational emphasis is put on these parameters for which the forecast error covariance matrices can be computed exactly (Bouttier 1993) : hence the assimilation can be made more optimal for the forecast of the phenomenon of interest.

- A similar technique can be used in field experiments : if a few observations are believed to be very important for the quality of the forecast, it is possible to coerce the EKF into computing exact covariances for these observations so that they will be used optimally.
- The EKF has direct applications to predictability studies, because $\mathbf{P}^a$ and $\mathbf{P}^f$ estimate the uncertainty of the analyses and short-range predictions. Hence it finds straightforward applications in fields such as singular vectors, EPS, breeding vectors, or adjoint prediction of local forecast error variances. The interested reader is referred to Bouttier (1995) for more details.

4. Conclusion : EKF tuning and adaptative filtering.

The lack of experience with real-size EKFs makes it difficult to draw a general conclusion. One can already predict that some efficient approximate algorithms are going to be designed, but a unique difficulty will always remain in the EKF : error covariances are never observed. This means that for the validation of $\mathbf{P}^a$, of $\mathbf{P}^f$, and for the specification of $\mathbf{R}$ and $\mathbf{Q}$, one must resort to indirect methods and do some constant adaptative retuning of the algorithm. This is not trivial because of the huge amount of information contained in the covariance matrices.

A basic sanity check of the assimilation is the quality of the subsequent forecasts, but it may be difficult to interpret in terms of individual features of the assimilation algorithm. The basic behaviour of the EKF can be monitored using maps of analysis and forecast standard errors, and averaged correlation structures like in figure 9. A more sophisticated diagnostic is provided by the eigenvectors of the forecasts error covariances, i.e. the singular vectors of the forecasts, which are supposed to explain the largest changes to the static error correlation structures ; this should also indicate what in the model dynamics is actually used by the EKF. A very important diagnostic is also provided by the fit of the background and analysis to the observations : effective values of forecast and analysis errors can be calculated easily. Finally, some interesting information may be gathered from the projection of forecast error sensitivity patterns (Rabier and Klinker 94, Errico and Vukicevic 92) onto the analysis error covariance matrices.

As was pointed out by Dee (1991), the calibration of the modelization error covariances is difficult, if not impossible ; however it is an important component of the EKF, and in itself it should be a useful summary of the weaknesses of the model. Hence it makes sense to invest into sophisticated methods in order to improve some components of matrix $\mathbf{Q}$. It has been proposed to diagnose model errors in the fit to the data of a 4D-Var assimilation (Dee 95) ; despite their large cost, other assimilation methods using the model equations as a weak constraint may provide directly this information.

Ideally, there should be a real-time algorithm which analyzes the performance of the EKF and automatically retunes its arbitrary components. This is the problem of adaptative filtering, for which a sound methodology remains to be developed. Daley (1992) has advocated the calibration of model error **Q** as the difference between the perceived forecast error covariances (provided by the differences between the background and the observations) and the covariance matrix generated by the EKF. Perhaps some efficient solutions can also be developed from other methods which are used in other fields, such as generalized cross-validation ; there are still many open problems which could lead to a better use of the observations in data assimilation.

Acknowledgements

Some important results presented in this seminar paper have been kindly provided by collaborators at ECMWF, mainly M. Fisher, J. Barkmeijer, F. Rabier, P. Courtier and J.-N. Thépaut. Some unpublished results have been communicated by G. Desroziers (CNRM, Météo-France). Thanks are also due to C. Edis and to the many developers of the joint IFS/Arpège project between Météo-France and ECMWF.

References

Balgovind, R., A. Dalcher, M. Ghil and E. Kalnay, 1983: A Stochastic-Dynamic Model for the Spatial Structure of Forecast Error Statistics. *Mon. Wea. Rev.*, **111**, 701-722.

Barkmeijer, J., M. van Gizjen and F. Bouttier, 1996: Singular vectors and the analysis error covariance metric. *To be submitted to Quart. J. Roy. Met. Soc.*

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

Bennett, A. and M Thornburn, 1992: The generalized inverse of a non-linear quasigeostrophic ocean circulation model. *J. Phys. Oceanogr.,* **3,** 213-230.

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

Bouttier, F., 1994: A dynamical estimation of error covariances in an assimilation system. *Mon. Wea. Rev.,* **122,** 2376-2390.

Bouttier, F., 1995: The Kalman filter. *Proceedings of the 1995 ECMWF seminar on predictability, 4-8 Sept 1995,* 221-245. Available from ECMWF, Shinfield Park, Reading RG2 9AX, GB.

Buizza, R., 1993: Impact of a Simple Vertical Diffusion Scheme and of the Optimisation Time Interval on Optimal Unstable Structures. *ECMWF Res. Dept. Tech. Memo. no.192,* available from ECMWF, Shinfield Park, Reading RG2 9AX, UK.

Buizza, R., J. Tribbia, F. Molteni and T. Palmer, 1992: Computation of Optimal Unstable Structures for a Numerical Weather Prediction Model. *Tellus,* **45A,** 388-407.

Cohn, S., and D. Parrish, 1991: The Behavior of Forecast Error Covariances for a Kalman Filter in Two Dimensions. *Mon. Wea. Rev.*, **119**, 1757-1785.

Cohn, S. and R. Todling, 1996: Approximate data assimilation schemes for stable and unstable dynamics. *J. Meteor. Soc. Japan,* **74,** 63-75.

Courtier, P., 1993: Introduction to numerical weather prediction data assimilation methods. *ECMWF seminar proceedings,* 6-10 Sept 1993, pp.189-207, available from ECMWF, Shinfield Park, Reading RG2 9AX, UK.

Courtier, P., 1996: Dual formulation of four-dimensional variational assimilation. *Proceedings of the HIRLAM workshop on variational data assimilation,* De Bilt, The Netherlands, 15-26.

Courtier, P., E. Andersson, W. Heckley, J. Pailleux, D. Vasiljevic, A. Hollingsworth, M. Fisher, F. Rabier, 1996: The ECMWF implementation of three-dimensional variational assimilation (3D-Var). Part 1: formulation. To be submitted to *Quart. J. Roy. Met. Soc.*

Daley, R., 1992: Estimating Model-Error Covariances for Application to Atmospheric Data Assimilation . *Mon. Wea. Rev.,* **120,** 1735-1746.

Daley, R., 1995: Estimating the wind field from chemical constituent observations: Experiments with a one-dimensional extended Kalman filter. *Mon. Wea. Rev.,* **123,** 181-198

Dee, D., 1991: Simplification of the Kalman filter for meteorological data assimilation. *Quart. J. Roy. Meteor. Soc.*, **117**, 365-384.

Dee, D., 1995: Testing the perfect-model assumption in variational data assimilation. *Proceedings of the 2nd WMO symposium on data assimilation,* Tokyo, 49-54.

Desroziers, G., 1996: A coordinate transformation for data assimilation of frontal structures. Submitted to *Mon. Wea. Rev.* Available at CNRM, Météo-France, 31057 Toulouse cedex, France.

Ehrendorfer, M., and J. Tribbia, 1996: Approximation of forecast error covariances by singular vectors. *Submitted to Tellus.*

Errico, R. and T. Vukicevic, 1992: Sensitivity Analysis using an Adjoint of the PSU-NCAR Mesoscale Model. *Mon. Wea. Rev.*, **120,** 1644-1660.

Evenssen, G., 1992 : Using the Extended Kalman Filter With a Multilayer Quasi-Geostrophic Ocean Model. *Jour. Geophys. Res.*, **99, C11,** 19995-19924.

Evenssen, G. and P. J. van Leeuwen, 1995: Advanced Data Assimilation Based on Ensemble Statistics. *Proceedings of the 2nd WMO symposium on data assimilation,* Tokyo, 153-163.

Fisher, M. and P. Courtier, 1995: Estimating the covariance matrix of analysis and forecast error in variational data assimilation. *ECMWF Res. Dept. Tech. Memo. no.220,* available from ECMWF, Shinfield Park, Reading RG2 9AX, UK.

Gauthier, P., P. Courtier and P. Moll, 1992: Assimilation of Simulated Wind Lidar Data with a Kalman Filter. *Mon. Wea. Rev;,* **121,** 1803-1820.

Ghil, M., 1989: Meteorological Data Assimilation for Oceanographers. Part I : Description and Theoretical Framework. *Dyn. of Atm. and Oceans*, **13**, 171-218.

Hoffman, R., Liu Zheng, J.-F. Louis and C. Grassotti, 1995: Distortion representation of forecast errors. *Mon. Wea. Rev.,* **123,** 2758-2770.

Hollingsworth, A., and P. Lönnberg, 1986: The statistical structure of short-range forecast errors as determined from radiosonde data. Part I: The wind field. *Tellus*, **38A**, 111-136.

Houtekamer, P., 1993: Global and local skill forecasts. *Mon. Wea. Rev.,* **121,** 1834-1846.

Ide, K., A. Bennett, P. Courtier, M. Ghil and A. Lorenc, 1995: Unified notations for data assimilation: operational, sequential and variational. *J. Met. Soc. Japan,* submitted.

Lacarra, J.-F., and O. Talagrand, 1988: Short-range evolution of small perturbations in a barotropic model. *Tellus,* **40A**, 81-95.

Lönnberg, P., and A. Hollingsworth, 1986: The Statistical Structure of short-range forecast errors as determined from radiosonde data. Part II: the covariance of height and wind errors. *Tellus*, **38A**, 137-161.

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

Lorenc, A., 1992: Iterative analysis using covariance functions and filters. *Quart. J. Roy. Met. Soc.,* **188,** 569-591.

Miller, R., 1986: Toward the application of the Kalman filter to regional open ocean modeling. *J. Phys. Oceanogr.,* **16,** 72-86.

Moll, P. and F. Bouttier, 1995: 3-D variational assimilation with variable resolution. *Proceedings of the 2nd WMO symposium on data assimilation,* Tokyo, 105-110.

Parrish, D., and S. Cohn, 1985: *A Kalman filter for a two-dimensional shallow-water model: formulation and preliminary experiments.* NMC Office note no.304 .

Parrish, D. and J. Derber, 1992: The National Meteorological Center's spectral statistical-interpolation analysis system. *Mon. Wea. Rev.*, **120,** 1747-1763.

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

Rabier, F., and P. Courtier, 1992: Four-dimensional assimilation in the presence of baroclinic instability. *Quart. J. Roy. Meteor. Soc.*, **118**, 649-672.

Rabier, F., E. Klinker, P. Courtier, A. Hollingsworth, 1994: Sensitivity of 2-day forecast errors over the Northern Hemisphere to initial conditions. *ECMWF Res. Dept. Tech. Memo. no.203,* available from ECMWF, Shinfield Park, Reading RG2 9AX, UK.

Rabier, F., A. McNally, E. Andersson, P. Courtier, P. Undén, J. Eyre, A. Hollingsworth and F. Bouttier, 1996: The ECMWF implementation of three-dimensional variational assimilation (3D-Var). Part II: structure functions. To be submitted to *Quart. J. Roy. Met. Soc.*

Talagrand, O. and P. Courtier, 1987: Variational assimilation of meteorological observations with the adjoint vorticity equation. I: Theory. *Quart. J. Roy. Meteor. Soc.*, **113,** 1311-1328.

Thépaut, J.-N. and P. Courtier, 1991: Four-dimensional data assimilation using the adjoint of a multilevel primitive-equation model. *Quart. J. Roy. Meteor. Soc.*, **117,** 1225-1254.

Thépaut, J.-N., P. Courtier, G. Belaud and G. Lemaître, 1994 : Dynamical structure functions in 4DVAR: a case study. *Proceedings of the 2nd WMO symposium on data assimilation,* Tokyo, 129-134.

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

Thompson, P., 1988: Stochastic-dynamic prediction of three-dimensional quasi-geostrophic flow. *J. Atmos. Sci.*, **45**, 2669-2679.

Toth, Z. and E. Kalnay, 1993: Ensemble forecasting at NMC : the generation of perturbations. *Bull. Amer. Meteor. Soc.,* **74,** 2317-2330.

Vukicevic, T., 1991: Nonlinear and linear evolution of initial forecast errors. *Mon. Wea. Rev.*, **119**, 1602-1611.