The Specification of Background Error Variances in the ECMWF Variational Analysis System

M. Fisher

ECMWF

Introduction

Error variances form an important component of the statistical model of background errors in any data assimilation system. The optimal interpolation (OI) analysis system which was in operational use at ECMWF before September 1996 included an explicit calculation of the variances of analysis error. These were inflated using a simple model of the growth of forecast errors to give an estimate of the variances of background error for the next analysis cycle. With the introduction of a variational data assimilation system at ECMWF, a replacement for the background error calculation of the OI scheme was also required. This paper describes the method which was implemented.

Estimation of Analysis Errors

It is well known that in variational assimilation, subject to certain simplifying assumptions, the Hessian matrix of the cost function at the minimum is the inverse of the covariance matrix of analysis errors (Thépaut and Moll, 1990). This relationship allows information about the variances and covariances of analysis error to be extracted.

Rabier and Courtier (1992) estimated the Hessian matrix as the empirical covariance matrix of a set of gradients obtained for different sets of randomly perturbed observations. Fisher and Courtier (1995) demonstrated two methods of extracting information about the covariance matrix of analysis error from the gradients generated during minimization of the cost function. These estimates correspond respectively to the estimates of the inverse of the Hessian matrix used (implicitly or explicitly) during conjugate gradient or quasi-Newton minimization, and are asymptotically more efficient than the random gradient method. Based on tests in a highly simplified analysis system, Fisher and Courtier concluded that, of the methods they evaluated, an algorithm based on conjugate gradient minimization was the most promising. It is this algorithm which has been adopted for the estimation of analysis errors in the ECMWF variational analysis system.

By way of an introduction to the rest of the paper, the forecast error calculation is briefly summarized below. The component parts of the algorithm are described in greater detail in subsequent sections.

Summary of the Algorithm

The main component of the forecast error calculation is a combined conjugate gradient and Lanczos algorithm. This algorithm calculates eigenvectors and eigenvalues of the Hessian matrix of the analysis cost function. It is described in the next section.

The eigenvectors and eigenvalues provide an approximation to the Hessian matrix of the analysis cost function. The approximate Hessian is inverted and transformed from the preconditioned space of the minimization to the space of physical variables. It is also transformed from a spectral to a gridpoint representation. The resulting matrix is an approximation to the covariance matrix of analysis error, expressed in terms of physical variables.

The diagonal of the matrix is easily computed, and contains approximate variances of analysis error at the gridpoints of the model.

The approximate variances of analysis error are inflated using a simple error growth model to provide estimates of the variances of short-term forecast error for use in the next analysis cycle. This part of the algorithm is similar to that used in the OI system, although the details are somewhat different.

The algorithm summarized above is currently used to provide estimates of short-term forecast error variances for most variables used in the analysis. However, variances of specific humidity are produced using an entirely separate algorithm due to Unden which parameterizes the variances as a function of the background specific humidity, temperature, pressure and land-sea mask. In the revised formulation of the background cost function (Bouttier *et al.*, 1997), variances of background error for the unbalanced components of temperature, surface pressure and divergence are also required. Currently, these are specified as fixed functions of model level.

The Combined Conjugate Gradient and Lanczos Algorithm

Under the tangent linear approximation, the cost function is quadratic. In any given direction in control space, the gradient of the cost function varies linearly. It is therefore sufficient to know the gradient at two points to completely determine the gradient at all points along a line. Given an initial gradient and search direction, it is sufficient to calculate a single gradient at some trial point along the search line to determine both the location of the line minimum of the cost function and the gradient at the line minimum. This allows a conjugate gradient minimization algorithm with exact line searches to be implemented at the cost of one gradient calculation per iteration. The computational cost is the same as a quasi-Newton algorithm in this case. Moreover, Nocedal (1980) shows that for quadratic functions with exact line searches, the quasi-Newton minimization algorithm used at ECMWF (Gilbert and Lemaréchal, 1989) produces identical iterates to a conjugate gradient scheme.

It is well known that conjugate gradient minimization is closely related to the Lanczos algorithm. Specifically, for exact line searches, the gradient vectors generated during conjugate gradient minimization of a quadratic function are proportional to the Lanczos vectors generated if the Lanczos algorithm is applied to the Hessian matrix of the function. Thus, by saving the gradient vectors from the minimization, good approximations to the leading eigenvectors of the Hessian matrix may be determined at minimal extra cost.

The main complication in combining the conjugate gradient and Lanczos algorithms arises from the fact that in any practical implementation, the Lanczos vectors must be explicitly orthogonalized. This is necessary to prevent the algorithm from repeatedly discovering the same eigenvectors. However, it is not clear how to apply the usual orthogonalization methods to the conjugate gradient algorithm. The Lanczos algorithm deals only with gradient vectors and generates each new vector using a single three-term recurrence equation. The conjugate gradient algorithm, on the other hand, generates search directions and control vectors as well as gradients. The gradients are calculated using a pair of coupled two-term recurrence equations. After some trial and error, the following method was found to work satisfactorily.

Given an initial control vector, search direction and gradient:

i) Make a trial step along the search direction and calculate a gradient.

ii) Using the initial gradient and the gradient at the trial point, determine the line minimum as the point along the line at which the gradient is orthogonal to the search direction.

iii) Orthogonalize the gradient at the line minimum against all converged eigenvectors or against all previous gradients.

iv) Normalize the gradient for use in the Lanczos algorithm.

v) Recalculate approximate eigenvectors using all the normalized gradients generated so far.

vi) Determine a new search direction using the Fletcher-Reeves formula.

vii) Repeat from step 1

Note that orthogonalization is performed at each iteration. This is because a reliable estimate of the loss of orthogonality of the Lanczos vectors has not been derived for the combined conjugate gradient and Lanczos algorithm. In any case, the cost of orthogonalization is small in comparison with the cost of a gradient calculation.

Calculation of Analysis Error Variances

The minimization is performed with respect to a control variable, χ which is related to the variables of the model, x, via a change of variable. In matrix form:

$$\chi = \mathrm{L}(\mathrm{x} - \mathrm{x}_b) \tag{1}$$

where x_b is the background.

The change of variable is chosen such that the background cost function is well approximated by

$$J_b \approx \frac{1}{2}\chi^T \chi \tag{2}$$

(In the revised formulation of the background cost function (Bouttier *et al.,* 1997), the background term is *defined* by the change of variable. In this case, equation 2 is exact.)

The change of variable serves two purposes. First, it acts as a preconditioner for the minimization. Second, it allows the Hessian matrix of the analysis cost function to be approximated by

$$J'' \approx \mathrm{I} + \sum_{i=1}^{K} (\lambda_i - 1)\mathrm{v}_i \mathrm{v}_i^T \tag{3}$$

where λ_i and $\mathrm{v}_i (i = 1 \ldots K)$ are those eigenvalues and eigenvectors of the Hessian which are determined by the combined conjugate gradient and Lanczos algorithm. Typically, these correspond to the K leading eigenvalues of the Hessian.

The approximate Hessian matrix is easily inverted to give the following approximation to the analysis error covariance matrix in the space of the control variable:

$$(J'')^{-1} \approx \mathrm{I} + \sum_{i=1}^{K} (\lambda_i^{-1} - 1)\mathrm{v}_i \mathrm{v}_i^T \tag{4}$$

Applying the inverse change of variable gives an approximation to the analysis error covariance matrix for model variables:

$$\mathrm{P}^a \approx \mathrm{L}^{-1}(\mathrm{L}^{-1})^T + \sum_{i=1}^{K} (\lambda_i^{-1} - 1)(L^{-1}\mathrm{v}_i)(L^{-1}\mathrm{v}_i)^T \tag{5}$$

The diagonal elements of this matrix are (approximate) variances of analysis error.

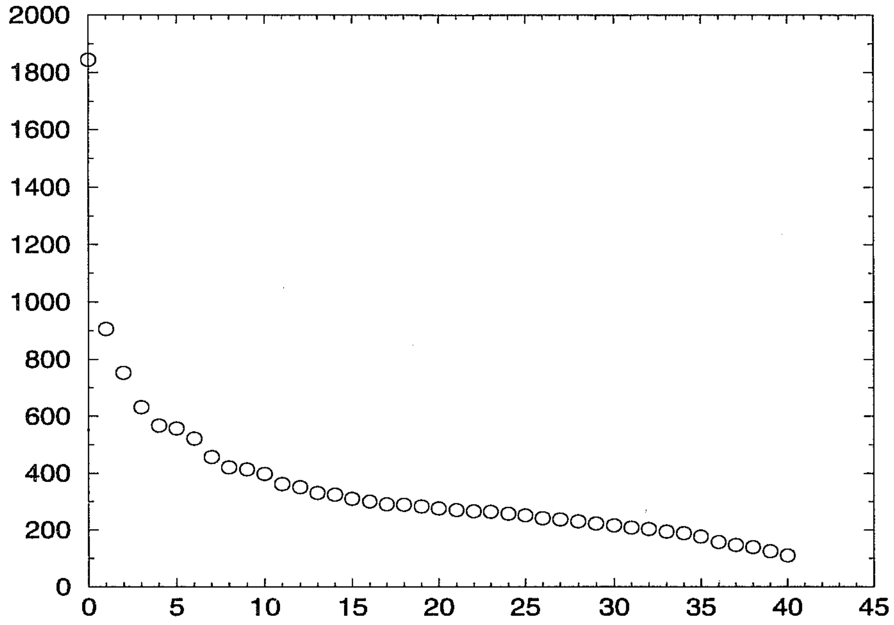

Figure 1: Eigenvalues of the analysis Hessian for a 3dVar analysis for 12UTC on 31 July 1997

The diagonal elements of the matrices $(L^{-1}v_i)(L^{-1}v_i)^T$ are the squares of the elements of the vectors $(L^{-1}v_i)$. These are easily calculated. The change of variable was chosen so that the matrix $L^{-1}(L^{-1})^T$ is an approximation to the covariance matrix of background error. The diagonal elements of this matrix are therefore approximately equal to the variances of background error. Thus, it is straightforward to calculate an approximation to the diagonal of P^a.

As a consequence of the choice of the change of variable, the leading eigenvalues of the Hessian matrix are larger than one. (In the case that equation 2 is exact, *all* eigenvalues of the Hessian are greater than or equal to one.) The diagonal elements of the matrices $(\lambda_i^{-1} - 1)(L^{-1}v_i)(L^{-1}v_i)^T$ are therefore negative. Since the number of eigenvectors of the Hessian determined by the Lanczos algorithm is much smaller than the number of non-unit eigenvalues, it is clear that equation 5 must overestimate the variances of analysis error.

Figure 1 shows a typical set of eigenvalues determined by the combined conjugate gradient algorithm for a 3d-Var analysis for 12UTC on 31 July 1997. In this case, 41 eigenvalues were determined. The largest and smallest eigenvalues were 1845 and 110.

Referring to equation 5, we see that the contribution made by an eigenvector to the estimated analysis error covariance matrix depends both on the weight, $(\lambda_i^{-1} - 1)$, and on the effect of the change of variable on the eigenvector. In particular, the contribution to the estimated variances of analysis error will be small if $\|L^{-1}v_i\|$ is small. Figure 2 shows the area-weighted norm of the vorticity component at model level 18 (close to 500hPa) for the eigenvectors whose eigenvalues are plotted in figure 1. The ordering of eigenvalues along the horizontal axis is the same in both figures.

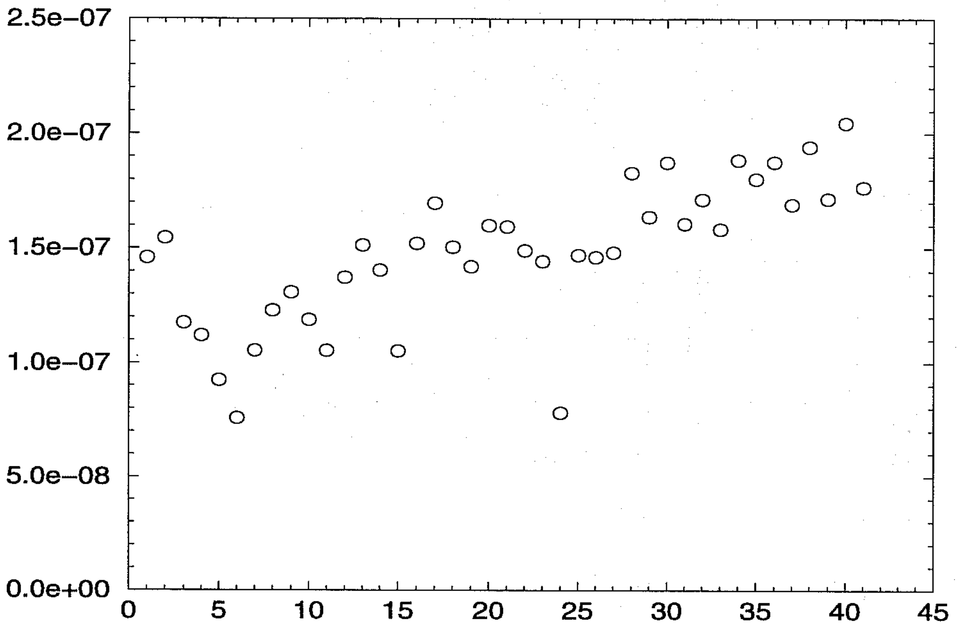

Figure 2: The norm of the vorticity component on model level 18 for the known eigenvectors.

It is clear from the tendency for the norm of the transformed eigenvector to increase as the eigenvalue decreases that a significant fraction of the difference between the variances of background and analysis error must be explained by eigenvectors outside the known part of the spectrum. However, it is difficult to estimate the size of this contribution. The number of unknown eigenvalues is very large. On the other hand, extrapolation of the tail of the spectrum shown in figure 1 suggests that a relatively small number of these eigenvalues will be sufficiently different from one that they will contribute with significant weight to the estimated variance of analysis error.

Figure 3 shows the square root of the estimated difference between the background and analysis error variances for vorticity at model level 18. The contour interval is $10^{-6}s^{-1}$. A typical standard deviation of background error for vorticity at this level is around $1.5 \times 10^{-5}s^{-1}$ in northern mid-latitudes. Although underestimated, the leading 41 eigenvectors provide information about the difference between background and analysis error variance for all densely observed areas. In particular, there is a reduction in error variance over most of the mid-latitudes of the northern hemisphere.

It is tempting to try to compensate for the underestimated reduction in error variance by multiplying the reduction due to the known eigenvectors by a suitable empirical factor. This is dangerous since it may produce variances which are too small, or possibly negative, in data-dense regions. For this reason, the safer alternative of rescaling the estimated short-term forecast errors at each analysis cycle has been adopted in the ECMWF analysis system.

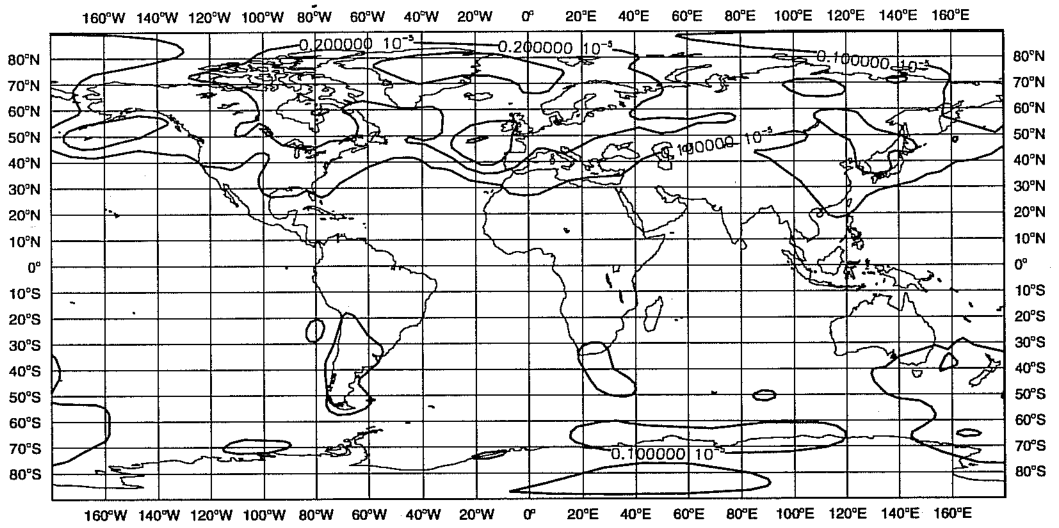
160°W 140°W 120°W 100°W 80°W 60°W 40°W 20°W 0° 20°E 40°E 60°E 80°E 100°E 120°E 140°E 160°E
80°N 70°N 60°N 50°N 40°N 30°N 20°N 10°N 0° 10°S 20°S 30°S 40°S 50°S 60°S 70°S 80°S
0.200000 10⁻⁵
0.200000 10⁻⁵
0.100000 10⁻⁵
0.100000 10⁻⁵
0.100000 10⁻⁵

Figure 3: The square root of the estimated difference between the background and analysis error variances for vorticity on model level 18 for 12UTC on 31 July 1997.

Estimation of Short-Term Forecast Errors

The standard deviations of forecast error are estimated from the standard deviations of analysis error by applying the error growth model of Savijarvi (1995). The evolution of the standard deviation of forecast error is given by

$$\frac{d\sigma}{dt} = (a + b\sigma)\left(1 - \frac{\sigma}{\sigma_\infty}\right) \tag{6}$$

Here, a represents error growth due to errors in the model; b represents the exponential growth rate of small errors; and σ_∞ is the standard deviation of forecast error at saturation.

At the time of implementation of 3d-Var, b was set to 0.4 day^{-1} while a and σ_∞ were set respectively to 0.8 and 10 times a prescribed vertical profile for each variable. In addition, the standard deviations of background error used in the analysis and in the estimation of analysis error variance were constructed by multiplying the mass-weighted vertical average of the standard deviation of 6 hour forecast error for vorticity by prescribed vertical mean profiles.

Since May 1997, three dimensional fields of σ_∞ have been used. These were calculated from the ECMWF re-analysis data. At the same time, the value of a was reduced, and background errors were no longer separated at each analysis cycle into the product of a horizontal pattern and a vertical profile.

As described in the preceding section, the variances of background error for each variable are normalized at each analysis cycle to have a prescribed global mean profile.

Figure 4 shows the estimated standard deviation of 6 hour forecast error for vorticity on model level 18 at 12UTC on 1 July 1997. There are several notable features. The reduction in analysis errors due to the dense observation network over Europe is clearly visible, as are the effects of AIREP observations over North America and south-eastern Australia. Error standard deviations are particularly low near the major airports in

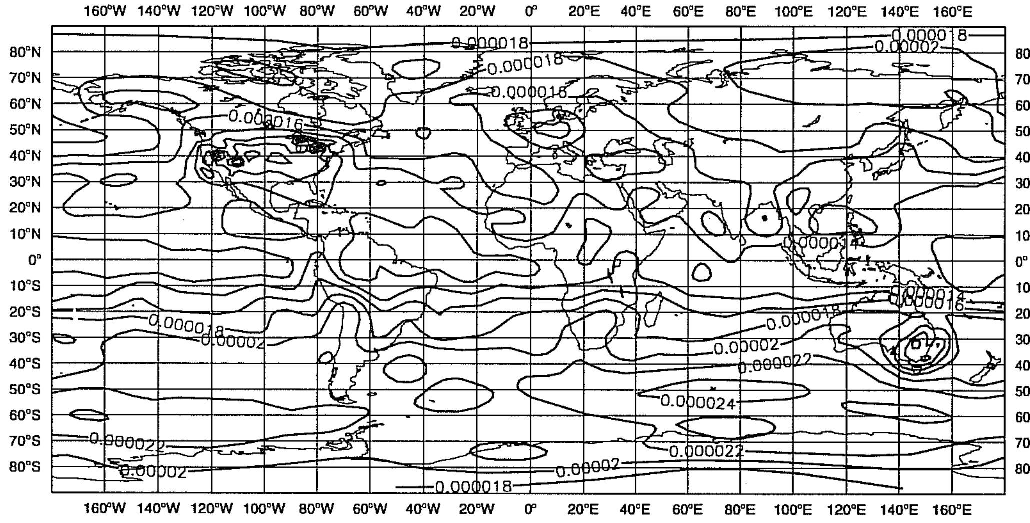

Figure 4: Estimated 6 hour forecast error for vorticity on model level 18 at 12UTC on 1 July 1997.

Summary and Future Plans

The specification of analysis and background error variances in the ECMWF variational analysis system is based on a diagnostic estimation of the Hessian matrix of the cost function. Although the calculation underestimates the reduction in error variance due to the observations, it is felt that a diagnostic method is preferable to a purely prescriptive approach. The algorithm adapts automatically to changes in the observing system, to the background error statistics, and to the analysis system. In 4d-Var, the algorithm will provide a flow-dependent estimate of analysis error variances.

A large part of the observing network remains unchanged with time. It is likely that a significant improvement in the estimation of analysis error variance could result from incorporating this information into the calculation. One way to do this is to precondition the calculation using the estimate of the Hessian produced for an earlier analysis.

The error growth model is currently rather simple. It is likely (Thompson, 1986) that advection plays an important role in the short-term evolution of forecast error variance. It is planned to incorporate advection of error variance into the error growth model. It may also be desirable to parameterize the growth rate of errors as a function of local vorticity gradients.

References

Thépaut, J-N., and P. Moll, 1990, Variational inversion of simulated TOVS radiances using the adjoint technique. Q. J. Roy. Meteor. Soc., Vol. 116, pp1425-1448.

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.

Bouttier, F., J. Derber and M. Fisher, 1997, The 1997 revision of the J_b term in 3D/4D-Var, ECMWF Res. Dept. Tech. Memo. no. 238.

Gilbert. J. CH., and C. Lemaréchal, 1989, Some numerical experiments with variable storage quasi-Newton algorithms, Mathematical Programming, Vol. 45, pp407-435.

Nocedal, J., 1980, Updating quasi-Newton matrices with limited storage, Mathematics of Computation, Vol. 35, no. 151, pp773-782.

Savijarvi, H. 1995, Error growth in a large numerical forecast system, Mon.Wea.Rev., Vol. 123, pp212-221.

Thompson, P.D., 1986, A Simple Method of Stochastic-Dynamic Prediction for Small Initial Errors and Short Range. MWR 114, pp1709-1715