Assimilation techniques (3): 3dVar
April 2001

By **Mike Fisher**

European Centre for Medium-Range Weather Forecasts.

Table of contents

1. INTRODUCTION

This aim of this lecture is to complement the more theoretical presentation of 3dVar given in the “Assimilation algorithms” module with a discussion of some of the techniques which are necessary for a practical implementation of the method. The ECMWF implementation of 3dVar is described. This was the operational ECMWF analysis between 30 January 1996 and 24 November 1997. The reader is referred to Andersson *et al.* 1998, Courtier *et al.* 1998 and Rabier *et al.* 1998 for a more detailed description.

The topics covered in this lecture are:

- The incremental method
- Initialization of analysis increments
- Formulation of the background cost function
- Methods for calculating background error statistics

2. THE INCREMENTAL METHOD

Ideally, the analysis cost function should be specified in terms of fields which have the same resolution as the forecast model. However, this makes the cost function computationally expensive to minimize (especially in 4dVar, where the much of the computational expense is in the tangent-linear and adjoint models.)

The incremental method reduces computational expense by minimizing a cost function which has a lower resolution than is used by the forecast model. This is reasonable because analysis increments are generally rather smooth, at least with current methods for specifying background error correlations.

The incremental method is an iterative procedure. For each iteration n, an approximate analysis x_n is generated from an initial approximation x_{n-1} by:

$$\mathrm{x}_n = \mathrm{x}_{n-1} + \mathrm{S}^{\dagger}\delta\mathrm{x}_n \tag{1}$$

Here, $\mathrm{S}^{\dagger}$ denotes the pseudo-inverse of an operator S which reduces the resolution of the model fields to the resolution used for the cost function. (For example, S might correspond to spectral truncation.) The pseudo-inverse of S is an operator which increases resolution (for example by padding the spectrum with zeros). In this case, $\mathrm{S}^{\dagger}$ increases the resolution of the increment $\delta\mathrm{x}_n$ to match that of the analysis.

The increment, $\delta\mathrm{x}_n$, is calculated by minimizing the cost function:

$$\begin{aligned} \mathrm{J}(\delta\mathrm{x}_n) = {} & \frac{1}{2}(\mathrm{Sx}_{n-1} + \delta\mathrm{x}_n - \mathrm{Sx}_b)^{\mathrm{T}}\mathrm{B}^{-1}(\mathrm{Sx}_{n-1} + \delta\mathrm{x}_n - \mathrm{Sx}_b) \\ & + \frac{1}{2}(H(\mathrm{x}_{n-1}) + \mathrm{H}'(\delta\mathrm{x}_n) - \mathrm{y})^{\mathrm{T}}\mathrm{R}^{-1}(H(\mathrm{x}_{n-1}) + \mathrm{H}'(\delta\mathrm{x}_n) - \mathrm{y}) \end{aligned} \tag{2}$$

Note in particular that the observation operator H acts on the high resolution fields x_{n-1} whereas the operator H' acts on the low resolution increment.

The incremental method is described above as an iterative procedure. However, there is generally *no guarantee that the iterations will converge*. For this reason, a typical implementation of the method performs only a few iterations.

2.1 Implementation of the Incremental method in the ECMWF 3dVar

In the case of the ECMWF 3dVar system, a single iteration of the incremental method is performed. The initial approximation to the analysis, x_0, is simply the background. The analysis consists of 3 steps:

2.1 (a) Comparison with the observations at high resolution. This step calculates $H(\mathrm{x}_b) - \mathrm{y}$. The background fields are transformed to gridpoint space and then interpolated to observation locations. 12-point bi-cubic interpolation is used for upper-air fields. Bi-linear interpolation is used for surface fields to avoid problems near surface discontinuities (e.g. coastlines) and steep orography.

Observation operators are applied to the interpolated model fields to calculate the model equivalents of the observations. The observations are compared with their model equivalents and the differences between the two (i.e. the departures) are written to the observation file, for use in the minimization. Obviously-bad observations are screened out at this stage by removing observations with very large departures.

2.1 (b) The minimization. This step calculates the low-resolution increment, $\delta\mathrm{x}$. The high-resolution upper-air background fields are truncated to T63. The gridpoint surface fields are transformed to spectral space, truncated to T63, and transformed back to gridpoint space. This ensures consistency between the spectral and gridpoint fields used in the minimization. In the case of the 31-level model, non-linear normal mode initialization (NNMI) is applied to the fields to adjust them to the low resolution orography. (NNMI has been found to give unsatisfactory results in the 50-level and 60-level versions of the model, so no initialization of the low-resolution background is performed for these vertical resolutions.)

The low resolution cost function for the increments is minimized using a quasi-Newton method (Gilbert and Lemaréchal, 1989). Variational quality control of observation departures is applied during the minimization (Andersson and Järvinen, 1998).

2.1 (c) Updating at high resolution. This step calculates the high-resolution analysis, $\mathrm{x}_a = \mathrm{x}_b + \mathrm{S}^{\dagger}\delta\mathrm{x}$.

Since the low resolution analysis is balanced with respect to the low resolution orography, it is desirable to initialize the analysis increment to adjust it to the high-resolution orography. However, NNMI is a non-linear procedure. It must be applied to whole fields rather than to increments. For the 31-level model, the high resolution analysis is defined as:

$$x_a = x_b + \mathrm{NNMI}(x_b + S^\dagger \delta x) - \mathrm{NNMI}(x_b) \tag{3}$$

No initialization is performed for the 50-level and 60-level versions of the model.

3. THE BACKGROUND COST FUNCTION

The background error term of the cost function is crucial to the performance of the analysis system. A simple example suffices to show why.

3.1 Example

Suppose we have a single observation of the value of a model field (e.g. temperature) at one gridpoint, corresponding to the k^{th} element of the state vector. The observation operator is very simple in this case, and is represented by the $1 \times N$ matrix whose k^{th} element is equal to one, and whose other elements are all zero:

$$\mathrm{H} = (0,0,\ldots,0,1,0,\ldots,0) \tag{4}$$

The gradient of the cost function is zero at the minimum, and (in the non-incremental formulation) is given by:

$$\nabla J(x_a) = B^{-1}(x_a - x_b) + H^{\mathrm{T}}R^{-1}(Hx_a - y) = 0 \tag{5}$$

Multiplying through by B, and rearranging gives

$$x_a - x_b = BH^{\mathrm{T}}R^{-1}(y - Hx_a) \tag{6}$$

But, for this example, BH^{T} is simply equal to the k^{th} column of B. Also, since we have just a single observation, $R^{-1}(y - Hx_a)$ is simply the scalar value $(y - (x_a)_k)/\sigma_o^2$, where $(x_a)_k$ is the analysed gridpoint value corresponding to the observation, and where σ_o^2 is the variance of observation error. Thus:

$$x_a - x_b = \left(\frac{y - (x_a)_k}{\sigma_o^2}\right)\begin{bmatrix} B_{1k} \\ B_{2k} \\ \ldots \\ B_{Nk} \end{bmatrix} \tag{7}$$

That is, the analysis increment is proportional to a column of the background error covariance matrix, B. In other words, the background covariance matrix controls how information is spread out from the single observation, to provide statistically consistent increments at the neighbouring gridpoints and levels of the model, and to ensure that observations of one model variable (e.g. temperature) produce dynamically consistent increments in the other model variables (e.g. vorticity and divergence).

3.2 Formulation of the background cost function in the ECMWF 3dVar

The background error covariance matrix is enormous, typically $10^6 \times 10^6$. This is much too large to fit into computer memory. Moreover, even if this was possible, we don't have enough statistical information to determine all its elements. We are forced to simplify things.

One way to approach the construction of the background error covariance matrix is to build a matrix L for which the variable $\chi = \mathrm{L}(\mathrm{x} - \mathrm{x}_b)$ has covariance matrix equal to the identity matrix. The background cost function may then be written as

$$\mathrm{J_b} = \frac{1}{2}\chi^{\mathrm{T}}\chi \tag{8}$$

The background error covariance matrix is defined implicitly by L as $\mathrm{B} = (\mathrm{L}^{\mathrm{T}}\mathrm{L})^{-1}$.

The effect of multplying background error by the matrix L is to remove the correlations between its elements. (Remember that the correlation matrix for χ is the identity matrix. That is, the elements of χ are uncorrelated.) A natural way to build L is as a sequence of steps, each of which removes some correlation from the background error.

The most obvious correlation in the background errors is the balance between mass errors and wind errors in the extra-tropics. We therefore define our L as $\mathrm{L} = \mathrm{L}_u\mathrm{K}^{-1}$, where the matrix K^{-1} takes the model variables (vorticity, divergence, temperature, log(surface pressure), specific humidity, and ozone mixing ratio) and subtracts from the temperature and log(surface pressure) components which are in balance with the vorticity. A component is also subtracted from the divergence to remove the correlation between divergence and vorticity due to Ekman pumping near the surface, and compensating upper-tropospheric outflow. Similarly, a balanced component is subtracted from the ozone to account for the correlation between vorticity and ozone background errors.

The balanced components of temperature, log(surface pressure), divergence and ozone are defined as:

$$\begin{aligned}(T, \ln p_s)_b &= \mathrm{M}\Psi\zeta \\ D_b &= \mathrm{N}\Psi\zeta \\ (O_3)_b &= \mathrm{O}\Psi\zeta\end{aligned} \tag{9}$$

where the matrix Ψ has the same form as the linear balance operator relating vorticity to height, but has coefficients which are statistically determined. The matrices M, N and O account for the vertical correlation between "height" (as defined by Ψ) and the balanced temperature, log(surface pressure), divergence and ozone. These matrices are block diagonal, with one full vertical matrix for each spectral component, whose coefficients depend only on the total wavenumber n.

The balance operator, K^{-1}, is intended to remove all correlations between different variables. The remaining part of the change of variable, L_u, is therefore block diagonal, with one block for each of the variables (vorticity, temperature-and-log(surface pressure), divergence, and ozone). Each block has the same form, for example the block corresponding to vorticity is:

$$(\mathrm{L}_u)_\zeta = \mathrm{C}_{v\zeta}^{-1/2}\mathrm{C}_{h\zeta}^{-1/2}\Sigma_\zeta^{-1} \tag{10}$$

Here, Σ_ζ^{-1} divides the vorticity by its standard deviation of background error; $\mathrm{C}_{h\zeta}^{-1/2}$ removes horizontal correlation; and $\mathrm{C}_{v\zeta}^{-1/2}$ removes vertical correlation.

Division by background error standard deviation is performed in gridpoint space, to allow a spatial variation of background error. However, the horizontal correlations are assumed to be spatially homogeneous and isotropic. This makes $C_{h\zeta}^{-1/2}$ diagonal, with coefficients which depend only on the total wavenumber n .

The matrix $C_{v\zeta}^{-1/2}$ is block diagonal, with one vertical matrix for each spectral component, whose coefficients depend only on the total wavenumber. This structure allows the vertical correlations to vary with horiziontal scale, so that large horizontal scales have deeper vertical correlations than small horizontal scales. It does not allow spatial variation of the vertical correlations. Note, however, that the vertical correlations of temperature *do* vary spatially. This is because in the tropics they are determined by the correlations defined for the unbalanced temperature, whereas in middle latitudes they are largely determined implicitly by the action of the balance operator on the vorticity correlation matrix.

Fig. 1 (a) shows the vertical correlation of temperature error with model level 18 of the ECMWF 31-level model (level 18 is at approximately 500hPa). The statistics were estimated using the NMC method (see below). Note that the vertical extent of the correlation is smaller in the tropics than in middle latitudes. Figure 1b shows the vertical correlations for temperature implied by the J_b formulation. Note that the main latitudinal variation in the vertical correlations is retained.

The effect of the balance operator in accounting for correlation between geopotential height and wind is demonstrated by Fig. 2 , which shows the wind increments generated by a single geopotential height observation at 60N, 30W. In Fig. 2 (a), the observation is placed at 1000hPa, and wind increments are shown for the nearest model level to 1000hPa. Note that the wind increment includes a convergent component. This is generated by the inclusion in the balance operator of a correlation between divergence and vorticity. Fig. 2 (b) shows the wind increment near 300hPa for a height observation at 300hPa. In this case, the increment is slightly divergent.

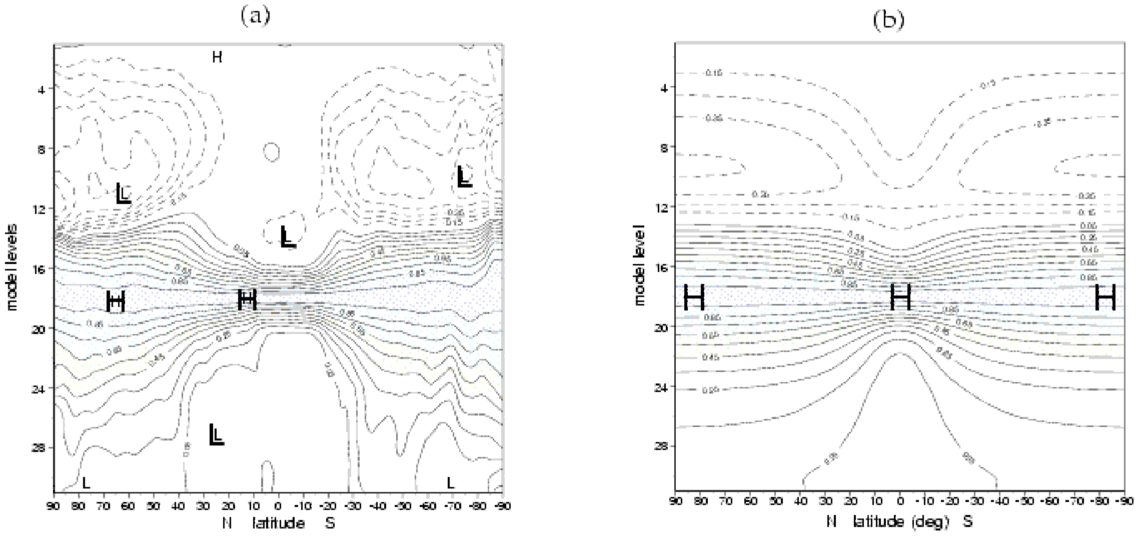

Figure 1. (a) Vertical correlation of temperature for 48h-24h forecast differences. (b) Vertical correlations of temperature implied by the J_b formulation. The vertical axis is model level for the 31-level model.

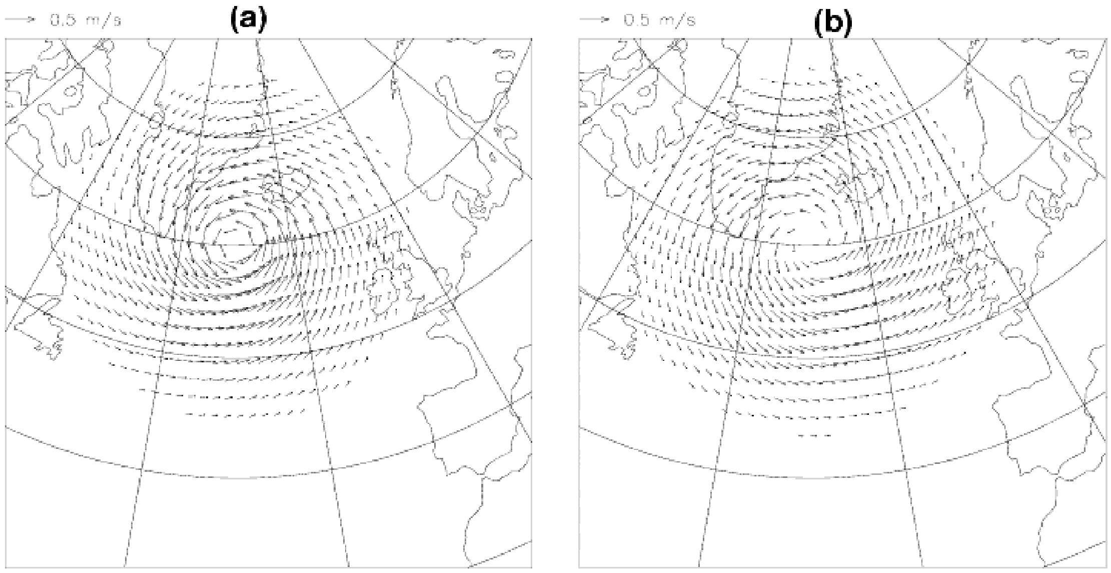

Figure 2. Wind increment generated by a single observation of geopotential. (a) Increments at model level 30 (near 1000hPa) from an observation at 1000hPa. (b) Increments at model level 13 (near 300hPa) from an observation at 300hPa. In both cases, the observed height is 10m lower than the background.

Fig. 3 demonstrates the way in which information from an observation is spread in the vertical. It shows a cross section of the increment generated by a single temperature observation at 200hPa.

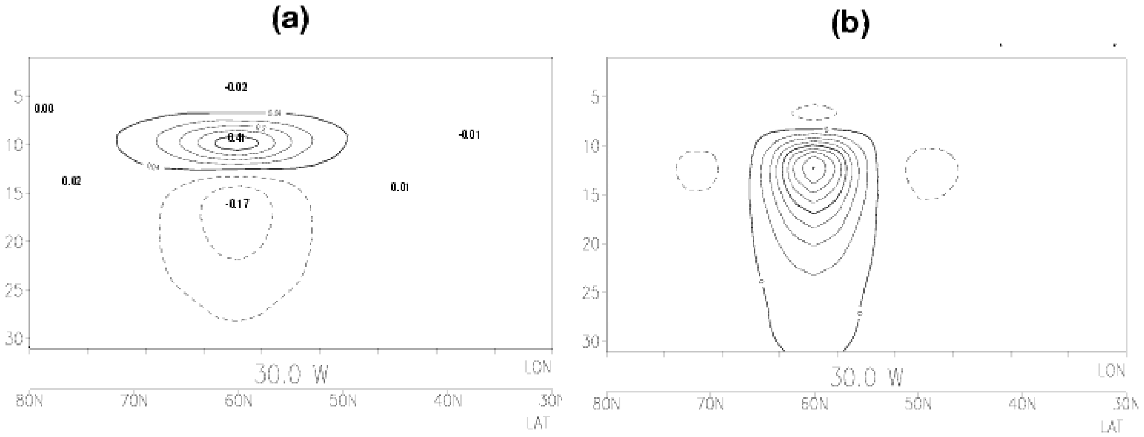

Figure 3. Cross section of the increment generated by a temperature observation at 200hPa, 60N, 30W. The observed value is 0.5K warmer than the background. (a) Temperature increment. (b) Vorticity increment (contour interval is 4×10^{-7}). The vertical axis for both plots is model level for the 31-level model.

3.3 Calculation of the background-error correlations

The previous section showed that, by making some simplifying assumptions, the number of non-zero elements of the background error covariance matrix may be drastically reduced. The largest matrices are of of order the number of levels of the model, and may be estimated statistically from a sample of background error of size a few times the number of levels.

Unfortunately, we cannot calculate background error, since this would require knowledge of the true state of the atmosphere. There are three main approaches to get round the problem:

3.3 (a) The Hollingsworth and Lönnberg (1986) method. This method looks at the spatial covariance of differences between observations and the background. These differences are a combination of background and observation error. We can partition the error into background errors and observation errors by assuming that the observation errors are spatially uncorrelated. If we bin observation-minus-background as a function of distance from each observation, only the zero-distance bin will contain a contribution from the observation error.

The Hollingsworth and Lönnberg (1986) method has the advantage that it is a direct diagnosis of background error covariance. However, it requires a uniform set of unbiased observations with spatially uncorrelated error. This makes it unsuitable for calculating the global statistics required by 3dVar. In addition, the method produces statistics for observable quantities such as wind and temperature, whereas the background cost function requires statistics for vorticity and unbalanced components of temperature, etc.. Nevertheless, the method remains a valuable tool. In particular, it can be used to verify that the standard deviations of background and observation error are correctly specified.

3.3 (b) The NMC method (Parrish and Derber, 1992). This method was used until recently in the ECMWF 3dVar and 4dVar systems. The method assumes that the statistical structure of forecast errors varies little over 48 hours. Under this assumption, the spatial correlations of backgound error should be similar to the correlations of differences between 48h and 24h forecasts verifying at the same time.

The advantage of the method is that it is straightforward to calculate the required global statistics. The disadvantage is that the underlying assumption that the statistical structure of 48h forecast error is similar to that of background error is difficult to justify.

3.3 (c) The analysis–ensemble method. The method currently used at ECMWF to estimate background error statistics is to run an ensemble of independent analysis experiments. For each experiment, the observations are perturbed by adding random noise drawn from the assumed distribution of observation error. The effect of the perturbations is to generate differences in the analyses for each experiment. These are propagated to the next analysis cycle as differences in backgrounds. After a few days of assimilation, the statistics of differences between background fields for pairs of members of the ensemble equilibrate, and in principle become representative of the true statistics of background error. As a further refinement to the method, the effect of model error may be represented by introducing random perturbations to the physical parameterizations used in the assimilating model.

Figs. 4 and 5 show some statistics of background error calculated using the analysis-ensemble method (figure 4) and the NMC method (Fig. 5). The analysis-ensemble method gives background correlations which are smaller in horizontal and vertical scale than those produced by the NMC method. The background differences are also less balanced than the forecast differences used by the NMC method.

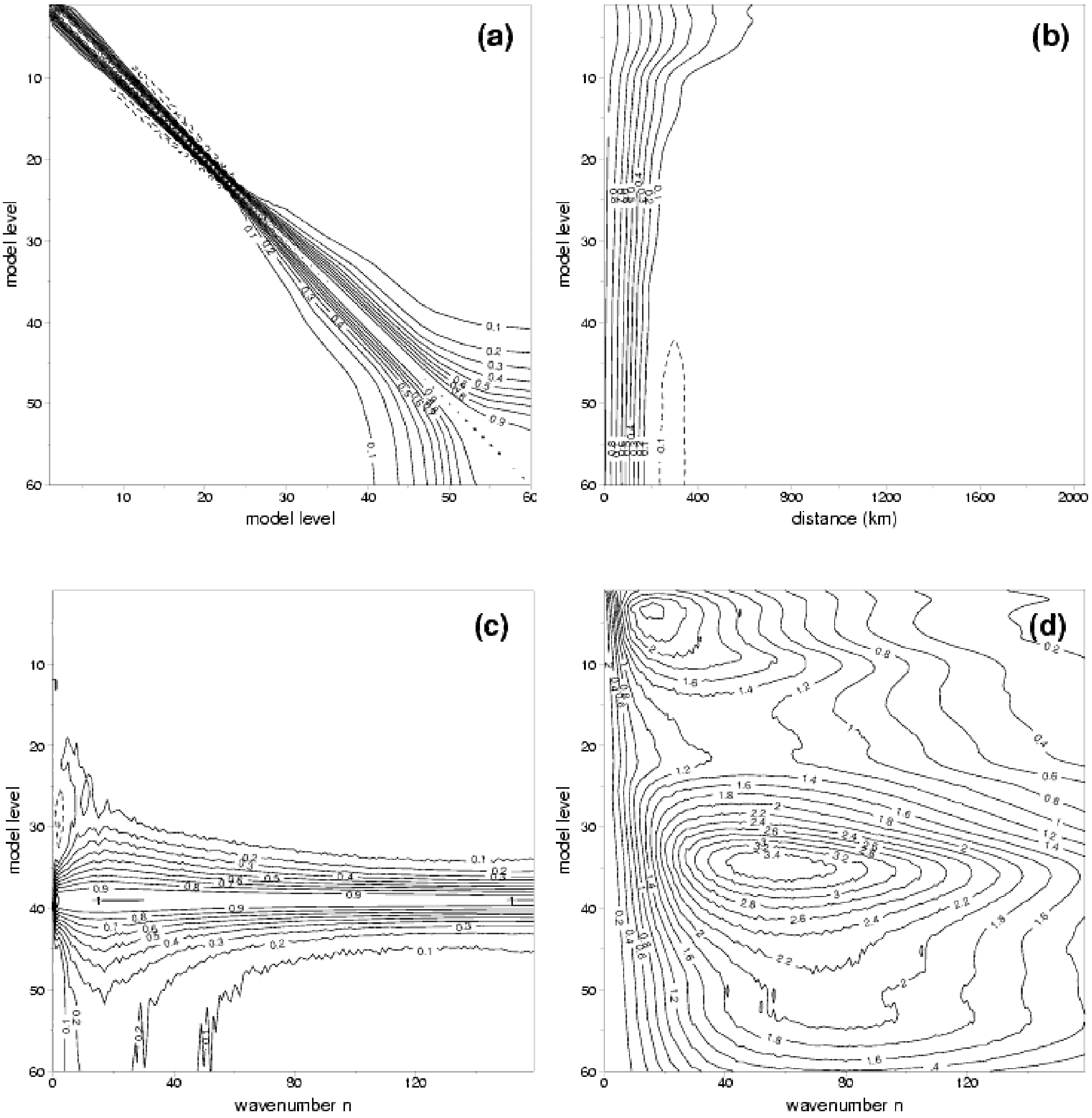

Figure 4. Statistics of background error for vorticity calculated using the analysis-ensemble method. (a) Wavenumber-averaged vertical correlation matrix. (b) Horizontal correlation as a function of model level and great-circle distance. (c) Vertical correlation with model level 39 (approx. 500hPa) as a function of wavenumber. (d) Standard deviation of vorticity error as a function of model level and wavenumber.

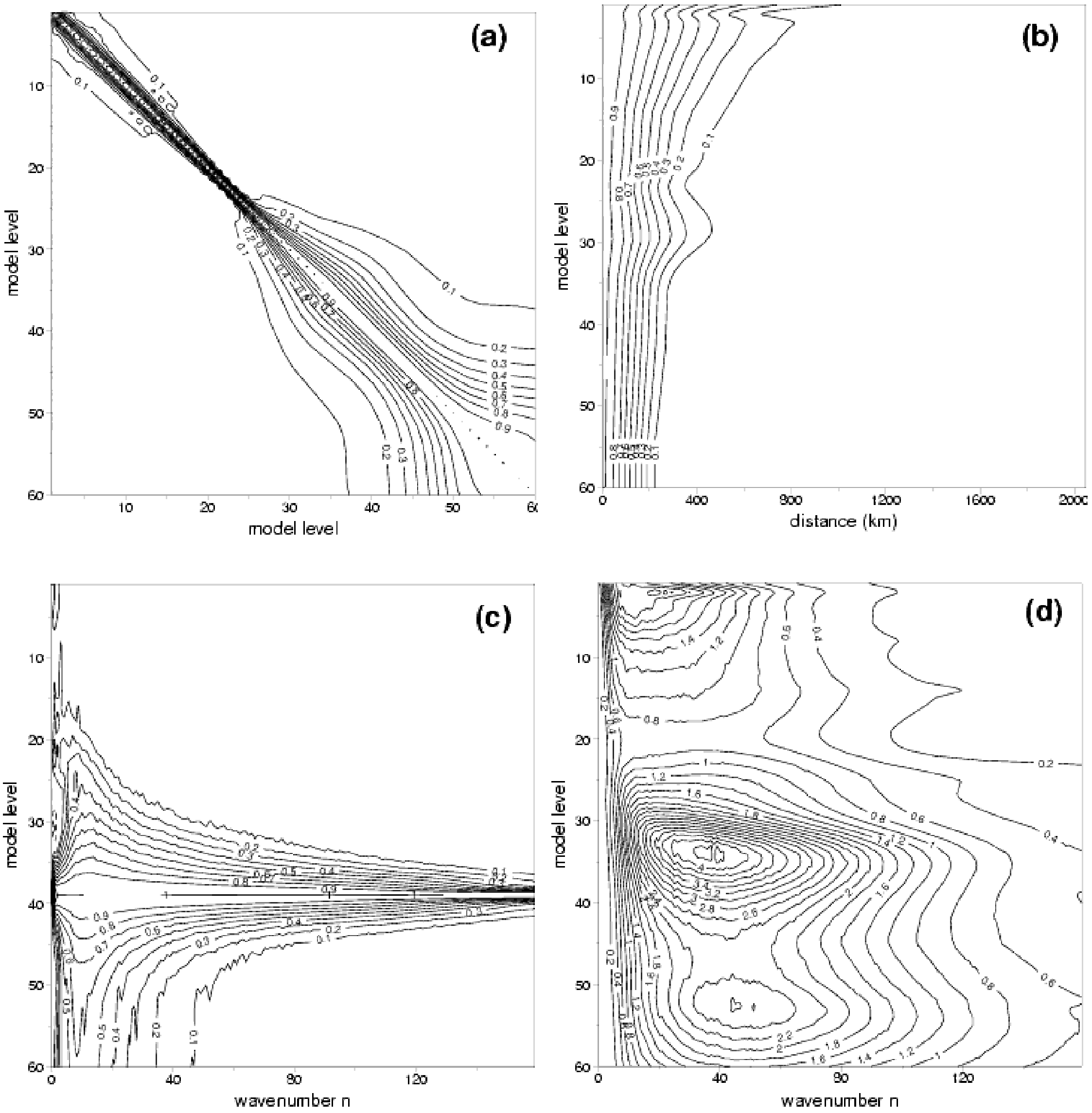

Figure 5. Statistics of vorticity background error calculated using the NMC method. (a) Wavenumber-averaged vertical correlation matrix. (b) Horizontal correlation as a function of model level and great-circle distance. (c) Vertical correlation with model level 39 (approx. 500hPa) as a function of wavenumber. (d)Standard deviation of vorticity error as a function of model level and wavenumber.

3.4 Calculation of background-error variances

The variance of background error is specified in gridpoint space. This allows the spatial variability of background error to be taken into account. (For example, background error variance is likely to be relatively smaller in areas of dense observational coverage, than in areas where there are few observations.) In practice, only the vorticity and specific humidity variances have a horizontal variation of background variance in the ECMWF 3dVar. (Note, however that the balance operator implies a horizontal variation of the balanced components of the other analysed variables.)

For specific humidity, background error variances are given by an empirical formula which expresses the relative humidity background error as a function of temperature and relative humidity. Additionally, background errors for humidity are reduced to very small values in the stratosphere, and are reduced at low levels over sea.

For vorticity, three-dimensional fields of background error standard deviation are calculated using a cycling algorithm (Fisher and Courtier, 1995) which applies an empirical error-growth model to a diagnostic estimation of the standard deviations of analysis error.

REFERENCES

Andersson, E. and H. Järvinen, 1998. Variational Quality Control. Quarterly Journal Royal Met. Society, Vol. 125, No. 554, pp 697.

Andersson, E., J. Haseler, P. Undén, P. Courtier, G. Kelly, D. Vasiljevic, C. Brancovic, C. Cardinali, C. Gaffard, A. Hollingsworth, C. Jakob, P. Janssen, E. Klinker, A. Lanzinger, M. Miller, F. Rabier, A. Simmons, B. Strauss, J-N. Thepaut and P. Viterbo, 1998, The ECMWF implementation of three dimensional variational assimilation (3D-Var). Part III: Experimental results. Quarterly Journal Royal Met. Society. Vol. 124, No. 550, pp1831-1860.

Andersson, E., M. Fisher, R. Munro, A. McNally, 2000, Diagnosis of background errors for radiances and other observable quantities in a variational data assimilation scheme, and the explanation of a case of poor convergence. Quarterly Journal Royal Met. Society. Vol. 126, pp1455-1472.

Courtier P., E. Andersson, W. Heckley, J. Pailleux, D. Vasiljevic, M. Hamrud, A. Hollingsworth, F. Rabier and M. Fisher, 1998, The ECMWF implementation of three-dimensional variational assimilation (3D-Var). I: Formulation. Quarterly Journal Royal Met. Society, Vol 124, No. 550, pp 1783.

Dennis J.E. Jr. and J.J. Moré, 1977, Quasi-Newton Methods, Motivation and Theory. SIAM Review, Vol. 19, No. 1, pp 46-89.

Fisher M. and P. Courtier, 1995, Estimating the covariance matrix of analysis and forecast error in variational data assimilation. ECMWF Technical Memorandum No. 220.

Gilbert J.C. and C. Lemaréchal, 1989, Some numerical experiments with variable storage quasi-Newton algorithms, Math. Prog 45, pp 407-436.

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, pp 111-136.

Liu D.C. and J. Nocedal, 1989, On the limited memory BFGS method for large scale optimization, Mathematical Programming, 45, pp 503-528.

Nocedal J., 1980, Updating quasi-Newton matrices with limited storage, Mathematics of Computation, 35, No. 151, pp 773-782.

Parrish, D. and J.C. Derber, 1992, The National Meteorological Center's spectral statistical interpolation analysis system. Monthly Weather Review, 120, pp1747-1763.

Rabier F., A. McNally, E. Andersson, P. Courtier, P. Undén, J. Eyre, A. Hollingsworth and F. Bouttier, 1998, The ECMWF implementation of three-dimensional variational assimilation (3D-Var). II: Structure function. Quarterly Journal Royal Met. Society, Vol. 124, No. 550, pp 1809.

Shewchuk J.R. 1994, An introduction to the conjugate gradient method without the agonizing pain. Edition $1\frac{1}{4}$. http://www.cs.cmu.edu/~jrs/

Wang Z., I. Navon, F.X. Le Dimet and X. Zou, 1992, The second order adjoint analysis: theory and applications.

Meteorology and Atmospheric Physics, 1992.

Wang Z., I. Navon and X. Zou and F.X. LeDimet, 1995, A truncated Newton optimization algorithm in meteorology applications with analytic Hessian/vector products, Computational Optimization and Applications, 4, pp 241-262.