Forecasting Research

Forecasting Research Division
Technical Report No. 116

Development of an Operational Variational Assimilation Scheme

by

Andrew Lorenc

September 1994

Meteorological Office
London Road
Bracknell
Berkshire
RG12 2SZ
United Kingdom

Forecasting Research Division
Technical Report No. 116

Development of an Operational Variational Assimilation Scheme.

Andrew Lorenc

September 1994

Forecasting Research
Meteorological Office
London Road
Bracknell
RG12 2SZ

This paper was prepared for the Sixth BMRC Modelling Workshop "Data Assimilation in Meteorology and Oceanography" BMRC, Melbourne, 5-7 October 1994

ABSTRACT

The Met Office is developing a variational assimilation for its Unified Model forecast system, which contains a grid-point model, run operationally in global, regional, and mesoscale configurations. Key characteristics of the design are:

- development path from 3-dimensional to 4-dimensional scheme
- global and limited area configurations
- variational analysis of perturbations
- carefully designed, well conditioned "background" term

This paper describes the variational scheme, with some example results from a simple 2-dimensional variational analysis which has been developed as a prototype.

1 Introduction

Over the past few years it has become apparent that variational assimilation schemes could be made practicable, and that possibly they might make a significant improvement in forecast quality:

i in the extraction of useful information from satellite radiances, by three-dimensional retrieval, allowing for errors correlations.

ii in diagnosing dynamically consistent baroclinic structures, given observations that a system is developing.

iii in using observations affected by "physical" atmospheric processes which are represented in the forecast model.

Most of the benefit from (i) might be realised from a static three-dimensional variational (3DVAR) system, while (ii) and (iii) probably need a four-dimensional (4DVAR) system containing a forecast model and its adjoint.

The bulk of the effort in developing a practical assimilation scheme goes in careful design and testing, and attention to detail in the observation processing. Currently the Met Office has a project to do this work, building a practical variational assimilation facility for the Met Office's Unified Model system, which contains a grid-point model, run operationally in global, regional, and mesoscale configurations. The project's targets are to match the current operational system, and to make possible the developments outlined above. I hope that we will have finished development of the basic 3DVAR system by mid-1996. Implementation of this will facilitate the developments necessary to get the benefits mentioned above; we might have a feasible 4DVAR scheme by 1997.

As a prototype for critical aspects of this development, a simple 2-dimensional variational analysis has been developed (2DVAR). Aspects studied include:

- The use of a filter and its adjoint for calculating the background penalty and its gradient (digital and spectral filters have been tried).
- Preconditioning using the filter.
- non-Gaussian observational errors.

This paper describes the full variational scheme, using illustrations from the 2DVAR when appropriate.

2 Variational Analysis

The "standard" formulation of variational analysis (Lorenc 1986) is - find the model state $\boldsymbol{x}$ which minimises a penalty (J) made up from a background term (J_b) and an observational term (J_o):

$$J(\boldsymbol{x}) = \tfrac{1}{2}(\boldsymbol{x}_b-\boldsymbol{x})^T\mathbf{B}^{-1}(\boldsymbol{x}_b-\boldsymbol{x}) + \tfrac{1}{2}(\boldsymbol{y}_o-\boldsymbol{y})^T(\mathbf{O}+\mathbf{F})^{-1}(\boldsymbol{y}_o-\boldsymbol{y}) \qquad (1)$$

where $\boldsymbol{x}_b$ is a prior (background) estimate of $\boldsymbol{x}$, with error covariance $\mathbf{B}$, $\boldsymbol{y}_o$ is a vector of observed values, with error covariance $\mathbf{O}$, and $\boldsymbol{y}$ is a prediction of the observed values, given by:

$$y = K(x) \tag{2}$$

F is the error covariance in the "generalised interpolation" K, which in our 2DVAR examples is a simple interpolation, but which in 4DVAR includes an NWP forecast model.

For the practical solution of this problem we make two transformations; to increments, and to a preconditioned control variable.

3 Analysis of Increments

Following Courtier et al. (1994), we solve instead for a model perturbation w', which may be at lower resolution than x. That is, we find the perturbation model state w' which minimises:

$$J(w') = \tfrac{1}{2}(w'_b - w')^T \mathbf{B}_w^{-1}(w'_b - w') + \tfrac{1}{2}(y_o - y)^T(\mathbf{O}+\mathbf{F})^{-1}(y_o - y) \tag{3}$$

where we use interpolation G to transform the background x_b, and the outer-loop estimate x_{ol}, to the lower resolution of w':

$$w'_b = G(x_b) - G(x_{ol}) \tag{4}$$

y, the prediction of the observed values, is now given by the sum of a contribution calculated in an outer-loop, and a perturbation calculated each iteration in the variational inner-loop:

$$y = K(x_{ol}) + \tilde{K}(w') \tag{5}$$

This transformation to a variational problem in w' is based on the belief that $x_{ol}+\mathbf{H}w'$ (where the interpolation **H** transforms from the low resolution of w' to that of x), will be a good approximation to the x which minimises (1). It is possible to iterate this correction process for x, outside of the minimisation iteration which finds w'. We use the suffix $_{ol}$ to denote the current outer-loop estimate for x.

Other groups working on 4DVAR have started from a full-fields approach, which needs the adjoint of the linearisation of the full model about its four-dimensional trajectory - usually called the tangent-linear model. A tangent-linear model is derived by differentiating the equations used in the full model. For a model with full physical parametrisations this is difficult to do. Our approach is different. The first-guess estimate of the atmosphere's four-dimensional trajectory is going to differ from the truth by a finite amount, with a spread governed by the background error variance. So we are designing a perturbation forecast model which gives an approximation to the evolution of finite perturbations. For instance if, in the trajectory of the full model, it is not raining, but nearly saturated, then some finite perturbations will be such as to make it rain. Thus the perturbation forecast model, designed

for the best average of all perturbations, should allow for some latent heating. Threshold processes, which can lead to difficulties in differentiating in the tangent-linear approach, should be smoothed in the perturbation forecast approach. The perturbation forecast model is designed from physical principles, and can have different resolution and algorithms from the full model. We use it, and its adjoint, to find the increment which most reduces the misfit to observations (show as y in figure 1) and the background. This is done using an iterative descent algorithm; the process is shown using solid arrows in figure 1. Adding it, we make a new full resolution four-dimensional trajectory (show using dotted arrows in figure 1), and can then repeat the inner incremental variational step.

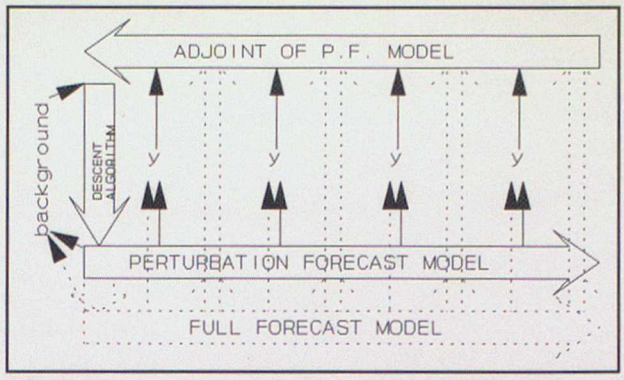

Figure 1 Incremental four-dimensional variational assimilation.

4 Pre-conditioned control variable

Secondly, we transform to a variable $\boldsymbol{v}$ designed to improve the conditioning of the Hessian matrix in the minimisation process. The Hessian is a matrix of second order partial derivatives with respect to the control variables. e.g. for (1) the Hessian is defined as:

$$\left(\frac{\partial^2 J}{\partial \boldsymbol{x}^2}\right) = \begin{pmatrix} \frac{\partial^2 J}{\partial x_1 \partial x_1} & \frac{\partial^2 J}{\partial x_1 \partial x_2} & \cdots & \frac{\partial^2 J}{\partial x_1 \partial x_{n_x}} \\ \frac{\partial^2 J}{\partial x_2 \partial x_1} & \frac{\partial^2 J}{\partial x_2 \partial x_2} & \cdots & \frac{\partial^2 J}{\partial x_2 \partial x_{n_x}} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{\partial^2 J}{\partial x_{n_x} \partial x_1} & \frac{\partial^2 J}{\partial x_{n_x} \partial x_2} & \cdots & \frac{\partial^2 J}{\partial x_{n_x} \partial x_{n_x}} \end{pmatrix} \qquad \textbf{(6)}$$

For (3), if $\bar{K}$ is linear, the Hessian is given by

$$\left(\frac{\partial^2 J}{\partial \boldsymbol{w}'^2}\right) = \mathbf{B}_w^{-1} + \bar{K}^T(\mathbf{O}+\mathbf{F})^{-1}\bar{K} \qquad \textbf{(7)}$$

The generalised interpolation $\bar{K}$ in the second term in (7) depends on the positions of the observations being used. It is hard to analyse its conditioning in a general way, so we concentrate on the first term, which depends on the background error covariance. It has been observed that the errors in $\boldsymbol{x}_b$ are usually balanced, and smooth. We assume that $\boldsymbol{x}_{ol}$ is similarly balanced and smooth. This means that balanced and smooth modes will correspond to small eigenvalues of $\mathbf{B}_w^{-1}$, while imbalanced, or rough modes will correspond to large eigenvalues. This large range of eigenvalues means that $\mathbf{B}_w^{-1}$ is ill-conditioned.

To alleviate this ill-conditioning, we use a filter $\mathbf{U}$ designed to reduce the power in

unbalanced or rough modes, and its inverse $\mathbf{T}$. We design these such that, approximately:

$$\begin{aligned} \mathbf{B}_w^{-1} &\approx \mathbf{T}^T\mathbf{T} \\ \mathbf{B}_w &\approx \mathbf{U}\mathbf{U}^T \end{aligned} \tag{8}$$

Then, defining a new control variable $\boldsymbol{v}$ such that

$$w' = \mathbf{U}\boldsymbol{v} \tag{9}$$

our transformed variational problem is to find the $\boldsymbol{v}$ which minimises

$$J(\boldsymbol{v}) = \tfrac{1}{2}(\boldsymbol{v}_b-\boldsymbol{v})^T\mathbf{B}_v^{-1}(\boldsymbol{v}_b-\boldsymbol{v}) + \tfrac{1}{2}(\boldsymbol{y}_o-\boldsymbol{y})^T(\mathbf{O}+\mathbf{F})^{-1}(\boldsymbol{y}_o-\boldsymbol{y}) \tag{10}$$

where

$$\boldsymbol{v}_b = \mathbf{T}\, w'_b \tag{11}$$

and the estimates of the observations are now given by:

$$\boldsymbol{y} = K(\boldsymbol{x}_{ol}) + \tilde{K}(\mathbf{U}\boldsymbol{v}) \tag{12}$$

The Hessian of (10) is given by:

$$\left(\frac{\partial^2 J}{\partial \boldsymbol{v}^2}\right) = \mathbf{B}_v^{-1} + \mathbf{U}^T\tilde{K}^T(\mathbf{O}+\mathbf{F})^{-1}\tilde{K}\mathbf{U} \tag{13}$$

Because of (8)

$$\mathbf{B}_v^{-1} \approx \mathbf{I} \tag{14}$$

so the first term in (13) is much better conditioned than in (7).

5 Variable transforms for the full model

As well as the above conditioning consideration, we need to be able to evaluate J_b and its gradient, for which multiplication by $\mathbf{B}_v^{-1}$ is needed, so again a simple diagonal form is desired. Note however that there is no requirement that (8) should be exactly obeyed. Any discrepancy will lead to a discrepancy in (14), and the background penalty term in (10) allows for this.

The transforms are to be constructed in stages: using simple physical ideas to transform parameters ($\mathbf{T}_p$), and zonal and seasonal-average statistics to transform into empirical modes in the vertical ($\mathbf{T}_v$), and to allow for different scales in the horizontal ($\mathbf{T}_h$). Statistics on the residual covariances $\mathbf{B}_v$ will be collected, and modelled in a simple way for use in the background penalty term.

We know physical relationships between variables, such as the closeness to balance, and non-divergence, which imply that elements of w' which are different physical parameters, e.g. temperature and wind, are correlated[1]. We use these relationships to design parameter transform $\mathbf{T}_p$ so as to separate w' into three-dimensional fields of variables which are uncorrelated with each other. In the first version these will be: velocity potential, stream function, the geostrophically unbalanced part of the hydrostatic pressure, and relative humidity.

Within each three-dimensional field there are still correlations between points close in space. We can accumulate average vertical covariances within each three-dimensional field, for instance by comparing forecasts valid at the same time. Making some assumptions we can design $\mathbf{T}_v$ so as to separate each three-dimensional field into two-dimensional fields of EOF coefficient.

Finally we design $\mathbf{T}_h$ to act on each two-dimensional field, allowing for horizontal correlations. This is described further below.

6 Implementation in 2DVAR

2DVAR is designed to test the above for a single two-dimensional field on a sphere, either globally, or for a rectangular limited area. For a single field there is no concept of "balance"; the only prior knowledge about background errors is that they are likely to be smooth. Thus the transformation $\mathbf{U}$ is implemented using a horizontal filter. It is assumed that (8) and (14) are exactly true. The v which minimises (10) is found using a descent algorithm.

6.1 Horizontal transform $\mathbf{U}$

Horizontal correlations between grid-points are normally defined using a continuous correlation function of the grid-point positions. Usually they are taken to be (locally) homogeneous and isotropic, so that the correlation is a function of the distance between the points only. In this case it is a standard result that the fourier transform of the covariance function is the power spectrum.

Our equations use matrix notation: $\mathbf{B}_h$ is a symmetric matrix and w' is a column vector. Multiplication by the matrix $\mathbf{B}_h$ does not represent a simple physical operation on the spatial field represented by w'; the physical interpretation of the resulting vector depends on the grid. In contrast a filter does represent a physical process; for resolved scales the result should be independent of the grid. So $\mathbf{B}_h$ cannot be represented solely by a filter. In order to relate $\mathbf{B}_h$ to a filter, we need to define a (symmetric) inner product matrix $\mathbf{P}_h$; this is a diagonal matrix of grid box areas. We can use the "square root" of this to transform

[1] strictly, the expected errors are correlated

variables. As explained in Lorenc et al. (1994) section 5, this gives:

$$\mathbf{B}_h = \mathbf{S}_h \mathbf{P}_h^{-1}$$
$$\mathbf{B}_h^{-1} = \mathbf{P}_h \mathbf{S}_h^{-1} \qquad (15)$$

We have two options for implementing $\mathbf{S}_h$; either as a grid-point or spectral filter:

Grid-point filter

Some meteorological variables have error correlations which can be approximated using a Second Order Auto-Regressive (SOAR) function. Lorenc (1992) showed that this was approximately equivalent to two passes of a recursive filter. We assume that $\mathbf{S}_h$ can be expressed as two applications of a filter $\mathbf{R}$, a one-pass recursive filter, followed by a scaling proportional to the filter scale, designed to make the filter equivalent to a correlation (i.e. filtering a unit delta function should give a correlation field, with maximum value one at the position of the delta function). We construct the two-pass filter from a one-pass filter and its adjoint in such a way that the result is exactly self-adjoint:

$$\mathbf{S}_h \equiv \mathbf{R}\,\mathbf{R}^* \qquad (16)$$

Using the recursive filter method allows us to vary the horizontal correlation scale smoothly in the horizontal, for example we can have different values in northern hemisphere, southern hemisphere, and tropics.

Spectral filter

An alternative way to perform a scale selective filter is via a spectral transform.

$$\mathbf{S}_h \equiv \mathbf{F}\,\mathbf{E}^2\,\mathbf{F}^{-1} \qquad (17)$$

Here $\mathbf{F}$ is the spectral transform from wave-space to grid-point-space, and $\mathbf{E}^2$ is a diagonal matrix defining the damping required for each wave. For a complete spectral transform, the inverse is also the adjoint:

$$\mathbf{F}^{-1} \equiv \mathbf{F}^* \qquad (18)$$

So, remembering that $\mathbf{E}$ is diagonal, we can define:

$$\mathbf{R} \equiv \mathbf{F}\,\mathbf{E}$$
$$\mathbf{S}_h \equiv \mathbf{R}\,\mathbf{R}^* \qquad (19)$$

Then for either the grid-point or spectral filter we get:

$$\mathbf{U}_h = \mathbf{R}\,\mathbf{P}_h^{-\frac{1}{2}} \qquad (20)$$

$$\mathbf{T}_h = \mathbf{P}_h^{\frac{1}{2}} \mathbf{R}^{-1} \tag{21}$$

These two approaches were tested in the 2DVAR program for a limited area grid. The spectral approach, using a double-sine FFT, proved to be more flexible; it could be tuned to match the grid-point filter nearly exactly, or some other power spectrum could be modelled.

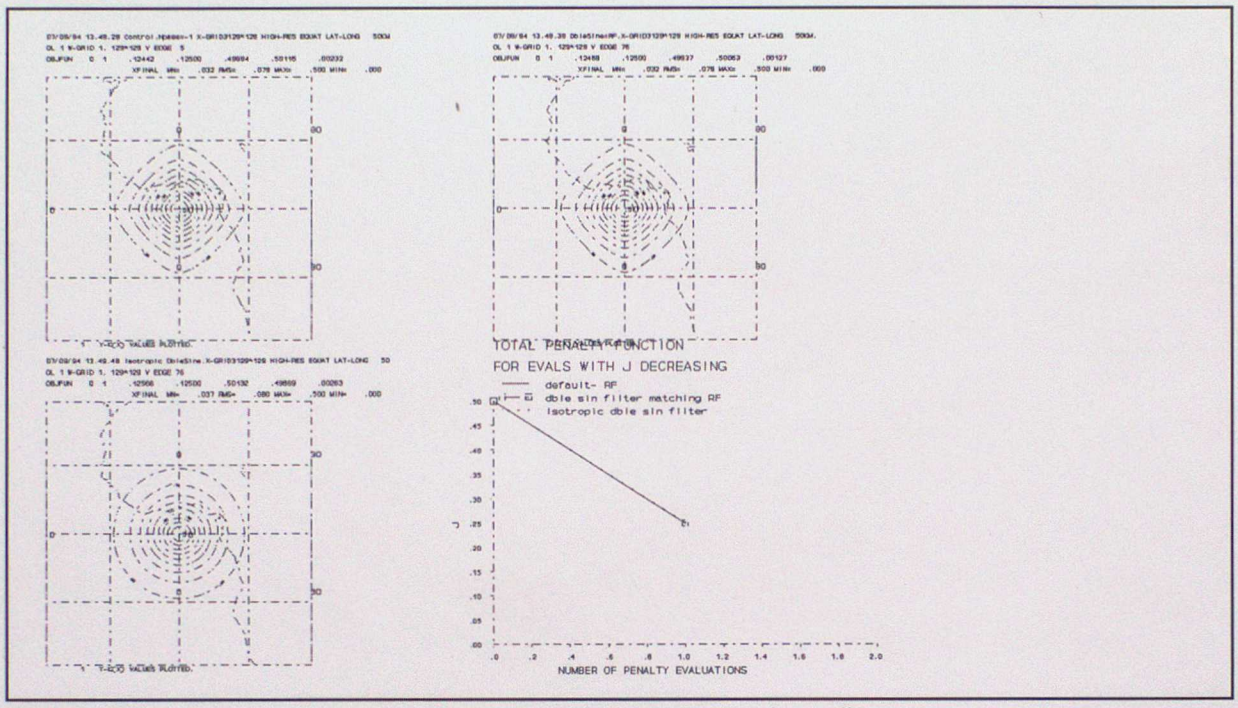

Figure 2 Increments due to a single observation, for various filters.

This is illustrated in fig.2, which shows the analysis increments due to a single observation, using the recursive filter designed to match a SOAR (top left), and a spectral filter designed to match the recursive filter (top right). The recursive filter, as it acts along the grid rows and columns, does not have an isotropic response; it is easy to make the spectral filter isotropic (bottom left). For this example all methods converged in one iteration. In a similar test using 1000 randomly placed observations, 18 penalty evaluations were needed. Computational costs for the two types of filter were similar.

6.2 The effect of preconditioning

The system is designed to find the minimum of (10). We also did experiments to instead find the minimum of (3). This is referred to below as the method without preconditioning.

In the recursive filter which matches a SOAR (Lorenc 1992), a one-dimensional wave of length $2\pi / k$ is damped by each pass by a factor given by:

$$S(k)=\frac{1}{1+\frac{\alpha}{(1-\alpha)^2}\left[2\sin\left[\frac{k\ \delta x}{2}\right]\right]^2} \qquad (22)$$

where the filter coefficient α is given by:

$$\begin{aligned}\alpha &= 1+E-\sqrt{E(E+2)} \\ E &= 2\ N_{pass}\ \delta x^2/4s^2\end{aligned} \qquad (23)$$

where s is the horizontal correlation scale, δx is the grid-length, and $N_{pass}=1$.

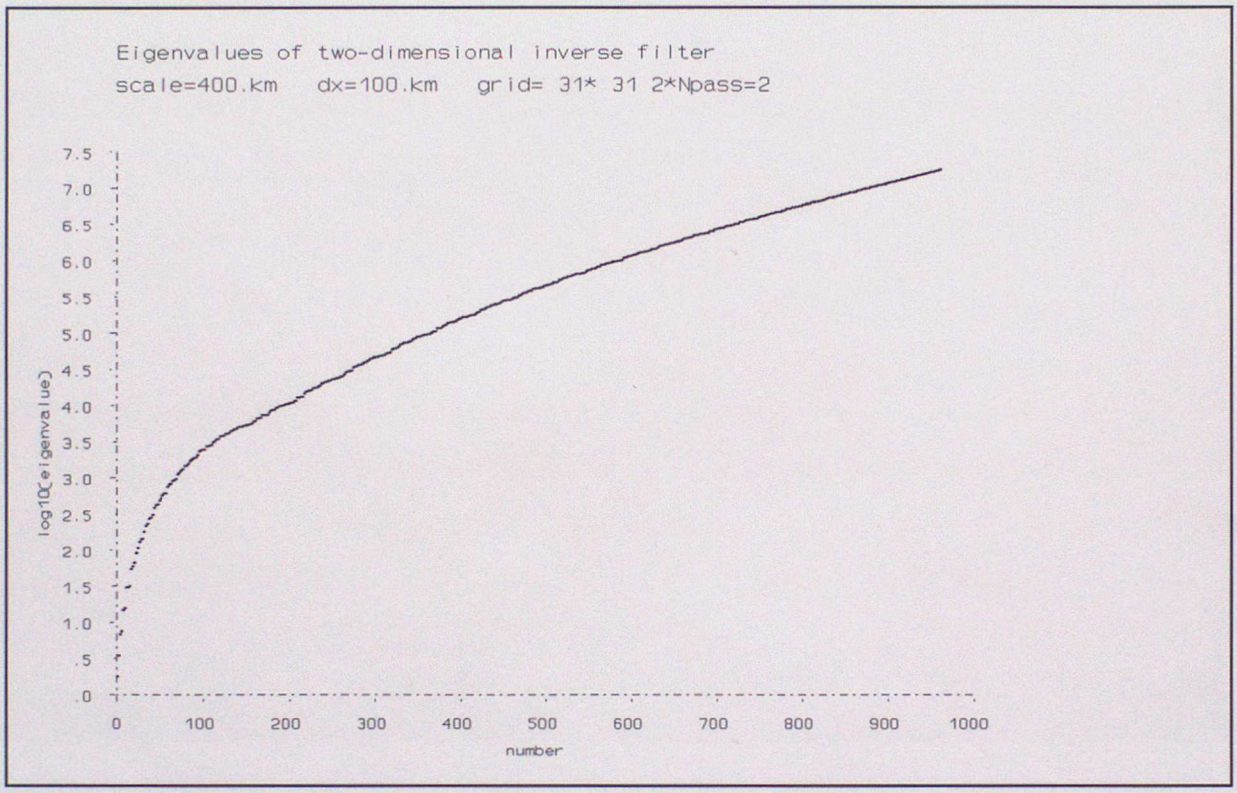

Figure 3 Eigenvalues of an inverse filter proportional to the inverse error covariance.

Using (15) we can calculate the eigenvalues of $\mathbf{B}_w^{-1}$. Figure 3 shows a plot of the relative[2] eigenvalues for the configuration used in most of the tests described below, for a 31by31 grid with a gridlength of 100km, and a horizontal correlation scale of 400km. From the ratio of the largest and smallest eigenvalues we can see the contribution of $\mathbf{B}_w^{-1}$ to the condition number of the Hessian. For our example this is larger than 10^7, and it increases rapidly as the grid-length is reduced below 100km. So if J_b dominates, the problem is badly ill-conditioned, and we can expect convergence of minimisation routines without preconditioning to be slow.

---

[2] The full inverse covariance also contains a scaling dependent on the background error variance, and the horizontal scale.

This is demonstrated in a simple experiment with a single observation, with value $y_o=1.0$, at the centre of the square grid area. The background field, and first-guess, is zero everywhere. For this case the preconditioning is perfect, so the descent algorithm converges in one iteration (which requires two penalty evaluations). The method without preconditioning however needs several thousand penalty evaluations before nearing convergence, (Convergence is judged by the approach of the norm of the gradient of J to zero), as shown in figure 4.

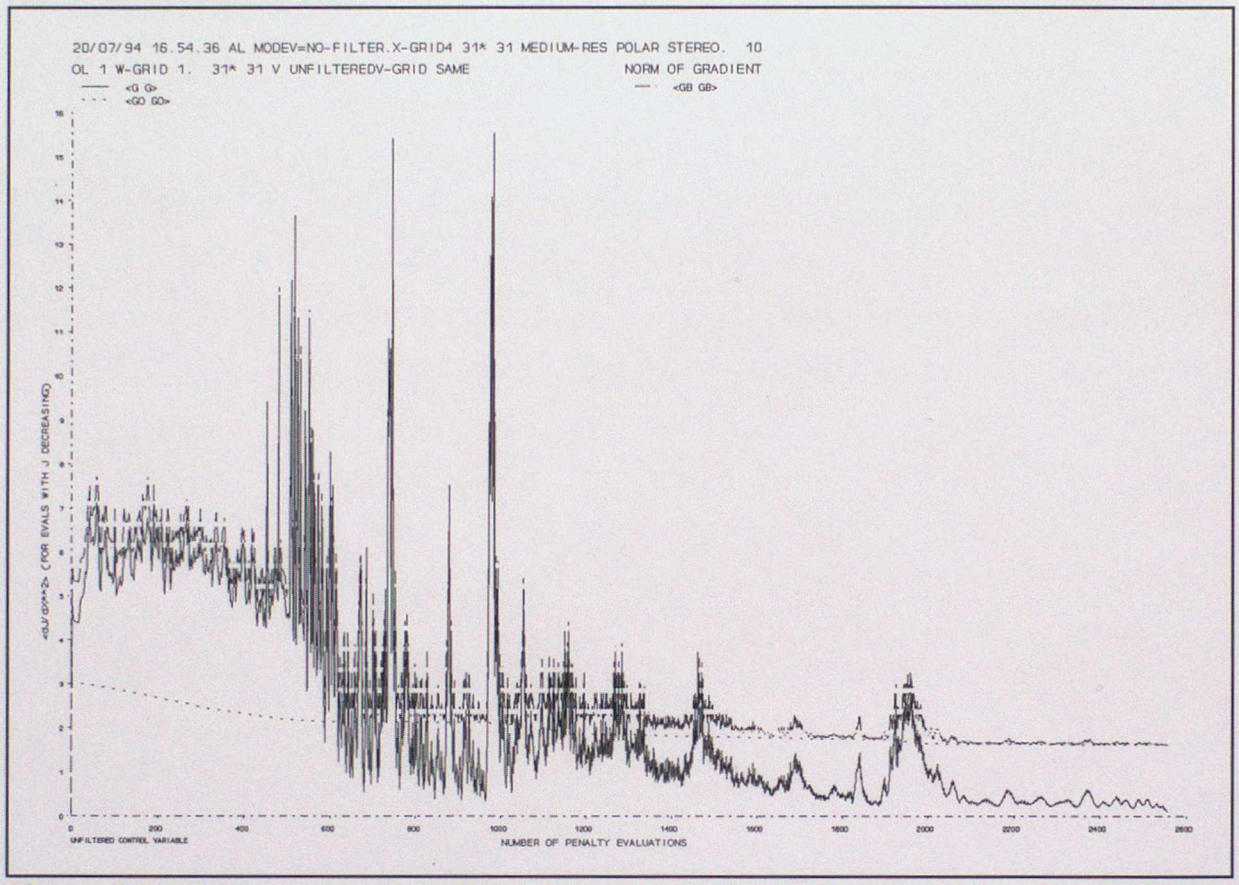

Figure 4 Solid line - norm of gradient of penalty function evaluated each iteration during the iterative minimisation (at convergence it should be zero). Dashed line & dotted line - norm of contribution to gradient from Jb and Jo terms.

Despite the many iterations, the method without preconditioning has still not converged to the exact solution, as can be seen in figure 5. We verified that the method with preconditioning was indeed giving the exact solution by using it as first-guess for another descent iteration. Neither method could improve on it.

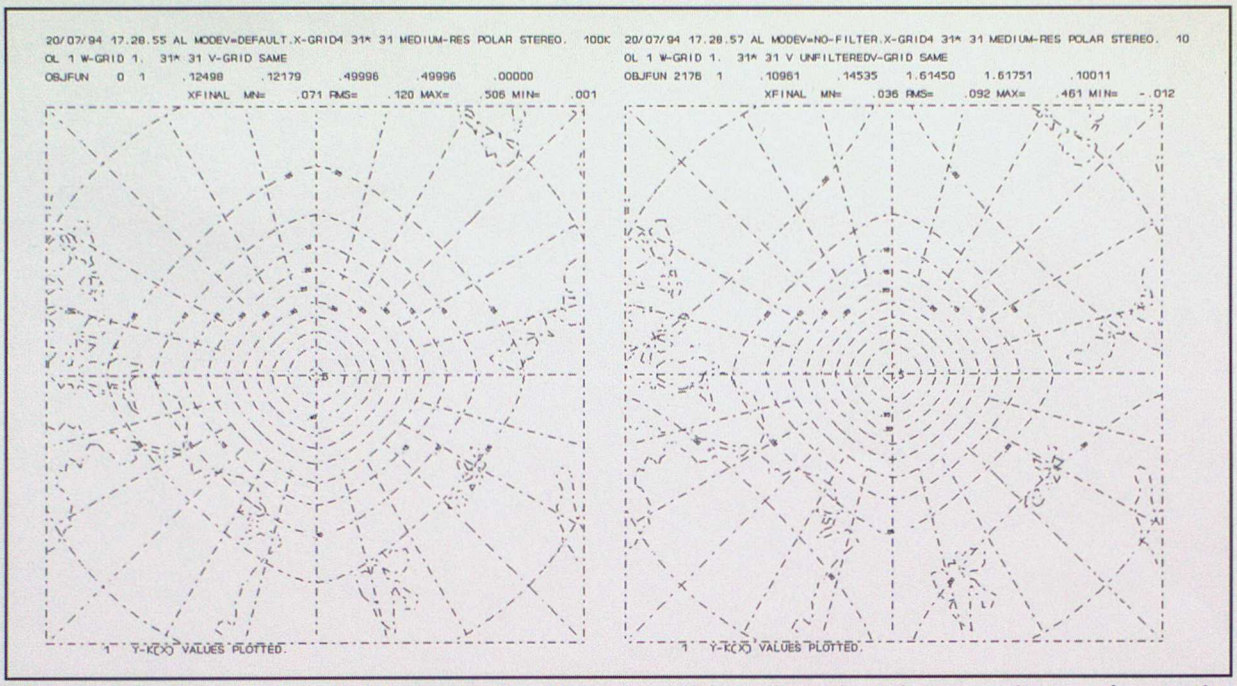

Figure 5 Left - solution field obtained with preconditioned method for one observation at the pole. Right - solution field obtained with the method without preconditioning.

7 References

Courtier, P., J-N. Thepaut, and A. Hollingsworth, 1994
"A strategy for operational implementation of 4D-Var, using an incremental approach". *Quart. J. Roy. Met. Soc.*, **519**, 1389-1408

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

Lorenc, Andrew 1992
"Iterative analysis using covariance functions and filters" *Quart. J. Roy. Met. Soc.*, **118**, 569-591

Lorenc, Andrew, et al. 1994
"Variational Data Assimilation - Control Variable Transforms" Met O(FR) UM Documentation Paper 101