THE ECMWF VARIATIONAL ANALYSIS: GENERAL FORMULATION AND USE OF BACKGROUND INFORMATION

W. A. Heckley, P. Courtier, J. Pailleux[1],
and E. Andersson
ECMWF
Reading, UK

Abstract

Formulation of the ECMWF variational analysis is described together with the detailed implementation of the background information term. Methods to obtain multivariate, balanced increments are introduced and illustrated. The importance of choice of control variable (pre-conditioning) is discussed. Results are illustrated through the use of single-observation experiments.

1. INTRODUCTION

In the 4D variational formulation of the data assimilation problem, where the model is assumed perfect, one has to find a model trajectory which fits the observations (y) available during an assimilation period (t_o, t_1), and which also fits the past information available as a background valid for t_o according to their respective statistical accuracy. This trajectory is entirely determined by the model state at time t_o, $x(t_o)$, through the integration of the model. $x(t_o)$, (or more generally $u(t_o) = F(x(t_o))$, where F is any invertible function) may be taken as control variable of the following minimization problem: minimize the cost function $J(u) = J_b + J_o$, where

J_b measures the distance between $x(t_o)$ and the background x_b;

J_o measures the distance between $x(t)$ and the observations.

It is not discussed here why the cost function is the sum of two terms. This can be shown to be related to hypotheses on the statistical distribution of the joint probability law of the couple (background information, observations) where one attempts to find the maximum likelihood estimator (*Lorenc*, 1986 and *Tarantola*, 1988). The model has to be integrated from t_o to the appropriate observation times in order to compare $x(t)$ with the observations. Large-scale minimization algorithms require the gradient of the cost function with respect to the control variable. The adjoint model has to be integrated back to t_o in order to obtain the gradient with respect to $x(t_o)$, the gradient with respect to u is then obtained by applying the adjoint of the change of control variable (*Le Dimet and Talagrand*, 1986).

---

[1] Present address: Météo-France, Toulouse

A three-dimensional variational analysis can be designed like the four-dimensional formulation described above in which the model integration is switched off: x is then compared directly with observations made at time t_o (or around t_o). It can be shown (*Lorenc*, 1986), that assuming a Gaussian probability law and quasi-linearity of the observation operator H, Optimal Interpolation is equivalent to the minimization of the cost function:

$$J = J_b + J_o \tag{1}$$

with $$J_b = \tfrac{1}{2}(x-x_b)^t B^{-1} (x-x_b) \tag{2}$$

$$J_o = \tfrac{1}{2}(H(x)-y)^t O^{-1}(H(x)-y) \tag{3}$$

where B is the covariance matrix of background error, H is the observation operator that allows computation of the model equivalent $H(x)$ of the observed quantity y, O is the covariance matrix of observation errors (which also contains the representativeness error). The superscript t denotes the transpose.

Minimization of $J(u)$ with respect to u requires the following steps:

a) Provide an initial estimate for u.

b) Computes $x = F^{-1}(u)$

c) Compute the cost function J and its gradient with respect to x.

d) Compute the gradient with respect to u (adjoint of step b)

e) Pass $\nabla_u J$ and $J(u)$ to a minimization scheme which computes a more accurate estimate of u.

f) Iterate on b) through e) until a desired convergence is achieved.

If other information is available, the cost function may have additional terms:

$$J = J_b + J_o + J_c + ... \tag{4}$$

where J_c might measure the distance to the slow manifold. The use of such a J_c is one technique to impose a mass-wind balance (*Courtier and Talagrand*, 1990).

As we are using minimization algorithms from libraries, this formulation of variational analysis, in practice, reduces to the computation of J_o, J_b, J_c, ... and their gradients. The gradient of J is given

$$\nabla_x J = \nabla_x(J_b + J_o + J_c + ...) = B^{-1} (x-x_b) + H'^t O^{-1} (Hx-y) + \nabla_x J_c + ..., \tag{5}$$

where H' is the tangent linear operator to H in the vicinity of x. The minimizing solution is characterised (if unique) by setting this expression to zero.

In methods of steepest descent the search direction is the direction of the local gradient J'. This is inefficient in a region of steep valleys in the cost function. If J is a strictly convex quadratic function then pre-conditioning with the inverse Hessian leads to the Newton algorithm which ensures the minimum is found in a single step. A limited storage quasi-Newton algorithm is still a great improvement on steepest descent without the necessity of the storage and computation of the Hessian. In this approach, using previous descent direction, one effectively takes a second order rather than first order local approximation to J. In the range of validity of the tangent linear operator H', J'' is given by

$$\nabla_x^2 J = B^{-1} + H'^t O^{-1} H'. \tag{6}$$

This will result in optimal pre-conditioning of the minimization. However the inverse of the Hessian is impossible to compute, or even to store, because of the size of the control variable u (10^7 in a model such as the ECMWF operational one). We shall use B^{-1} as a pre-conditioning, the change of variable u being such that $F B^{-1} F^{-1}$ is diagonal as detailed in section 2.2.

Minimization is performed using a limited storage quasi-Newton type algorithm M1QN3 provided by the Institut National de Recherche en Informatique et en Automatique (INRIA, France). A description of the algorithms and the performance of the code are given in *Gilbert and Lemaréchal* (1989). Essentially, the method uses the available in-core memory provided by the user to update an approximation of the inverse Hessian matrix of the cost function. Once the memory is used up, the quasi-Newton matrix (approximation of the inverse Hessian) keeps being modified during the minimization process by dropping information coming from the oldest gradient and inserting information coming from the more recently computed gradient.

2. BACKGROUND CONSTRAINT

2.1 Univariate formulation

The background term is given by equation (2):

$$J_b = \tfrac{1}{2}(x - x_b)^t B^{-1}(x - x_b)$$

The practical difficulty is the size of B which does not allow, in practice, its inversion. B, being symmetric, can be diagonalised:

$$B = \mathcal{L} \Lambda \mathcal{L}^{-1}$$

with $\mathcal{L}$ unitary $\mathcal{L}^{-1} = \mathcal{L}^t$

and J_b becomes

$$J_b = \tfrac{1}{2} [\Lambda^{-\frac{1}{2}} \mathcal{L}^{-1}(x-x_b)]^t [\Lambda^{-\frac{1}{2}} \mathcal{L}^{-1}(x-x_b)].$$

The idea of the practical implementation of 3D-Var is to approximate $\mathcal{L}^{-1}$ with a sequence of operators. For a univariate analysis we choose to have the following sequence of operations:

1)	Difference x and x_b in spectral space	
2)	Convert from vorticity, divergence to winds	W^{-1}
3)	Transform to grid-point space	S^{-1}
4)	Normalize with respect to background errors σ	N
5)	Transform to spectral space	S
6)	Convert from winds to vorticity, divergence	W
7)	Multiply by the square root of the inverse spectral horizontal background error covariance matrix	$h^{-\frac{1}{2}}$
8)	Project onto the eigenvectors of the vertical background error correlation matrices	P^{-1}

$$\chi = P^{-1} h^{-\frac{1}{2}} W S N S^{-1} W^{-1} (x - x_b) \tag{7}$$

and $$J_b = \tfrac{1}{2} \chi^t \Lambda^{-1} \chi \text{ with } \chi = \mathcal{L}^{-1}(x-x_b) \tag{8}$$

$$\nabla_\chi J_b = \Lambda^{-1} \chi. \tag{9}$$

Identifying $\mathcal{L}^{-1}$ as the sequence of operators 2) through 8), Λ is a diagonal matrix containing the eigenvalues of the vertical background error correlation matrices.

In order to obtain the gradient with respect to x the adjoints of the above operations have to be applied in reverse sequence.

$$\nabla_x J_b = (W^{-1})^*(S^{-1})^* N^* S^* W^* (h^{-\frac{1}{2}})^* (P^{-1})^* \nabla_\chi J_b \tag{10}$$

where * denotes an adjoint operator.

2.2 Choice of control variable

Let us assume that the minimization is performed in the space of the model variable x. Since the $\sigma's$ are spatially varying the B matrix will be far from diagonal, and the problem might be ill conditioned if a diagonal matrix is used for defining the scalar product. However, if the σ values are taken as constant over η levels and errors are assumed uncorrelated in the vertical, then B is diagonal, and exact minimization of J_b alone can be accomplished in 1 iteration using B^{-1} for defining the metric.

Alternatively, the control variable may be taken as χ then the Hessian is simply Λ^{-1}, which is diagonal. Such a change of control variable and the use of the matrix Λ^{-1} for defining the metric improves the pre-conditioning. The exact solution for J_b alone is found in a single step, even with full geographical variability of the forecast errors and vertical coupling. It is, of course, simpler to re-define the control variable as $\Lambda^{-½}\chi$, which reduces J_b to its simplest form - which is done in practice.

2.3 Multivariate analysis

In the variational analysis which became operational in June 1991 at the National Meteorological Center in Washington (*Parrish and Derber*, 1991), the model variables are constrained in order to stay close to the equilibrium of the linear balance equation applied on the model levels (slightly modified to account for divergence). This is achieved by a choice of control variable which takes into account the balance equation. The NMC analysis variables are:

- Departures from the 6 h forecast for vorticity and divergence;
- Departures from the balance equation solution of the temperature departures to the 6 h forecast.

By assigning appropriate statistics to the errors of these variables, a balance is achieved which the authors claim is good enough to obviate the need for normal mode initialization.

In the assimilation context, one is interested in finding an analyzed state which is close both to the observations and to the slow manifold. In the previous sections, the 3D-Var problem was expressed as

$$\min_{x \in E} J(x) = J_b + J_o \qquad (11)$$

find the minimum of J in the space E which is the phase space of the model.

Forecast evolution is confined to the attractor of the model, which is approximated by the slow manifold. A consequence is that the forecast errors lie, to first order of approximation, on the tangent plane of this slow manifold. In other words they do not span the whole phase space E but only a subspace, which one may denote E_R. In the current ECMWF operational implementation of OI, E_R is defined by geostrophic balance on the f-plane. In the NMC implementation of 3D-Var (*Parrish and Derber*, 1991), E_R is defined by the linear balance equation $\nabla^2\phi = \nabla.(f\nabla\psi)$ where ϕ is the geopotential and ψ is the stream function, with some enhancements to account for divergence.

A consequence of assuming that the errors are wholly within E_R is that B is singular: the kernel of B contains the orthogonal of E_R which will be denoted by E_G. B is no longer invertible, the formulation

defined by (11) and (2) is then no longer suitable for this problem. It can however easily be reformulated as

$$\min_{x \in E_R} J(x) = J_b + J_o \tag{12}$$

with

$$J_b = [S_R(x-x_b)]^t B_R^{-1} [S_R(x-x_b)] \tag{13}$$

where S_R is the projection onto the subspace E_R of Rossby modes and parallel to the subspace E_G of gravity modes. The last equation is equivalent to

$$J_b = (x-x_b)^t \, S_R^t \, \tilde{B}^{-1} \, S_R(x-x_b) \tag{14}$$

where $\tilde{B}^{-1}$ is any matrix identical to B_R^{-1} on E_R and which could take any value on E_G. Furthermore, one should notice that

$$\min_{x \in E_R} J(x) \Leftrightarrow S_R\left[\min_{x \in E} J(S_R(x))\right]$$

and as, in practice, a descent algorithm is used which computes descent directions as a linear combination of several gradients, and as the initial point of the minimization can be assumed to be on the slow manifold, one has the algorithmic equivalence

$$\min_{x \in E_R} J(x) \overset{alg}{\Leftrightarrow} \min_{x \in E} J(S_R(x)) \tag{15}$$

with

$$J_b = (x-x_b)^t \, S_R^t \, \tilde{B}^{-1} \, S_R(x-x_b) \tag{16}$$

Remark This is different from the problem

$$\min_{x \in E_R} J(x) = J_b + J_o \tag{17}$$

with

$$J_b = (x-x_b)^t \, \tilde{B}^{-1}(x-x_b) \tag{18}$$

which could be solved by resetting the gradient in the gravity part to zero. Since however, $S_G(x-x_b) \simeq 0$ at the beginning of the minimization, the two formulations lead to similar results.

2.3.1 *Shallow-water illustration*

From (16) it is clear that if a matrix $\tilde{B}^{-1}$ has been specified and if any kind of gravity wave control is used, the inverse of the effective matrix of covariance of first-guess error becomes $S_R^t \, \tilde{B}^{-1} \, S_R$.

Consider the implementation of J_b as it pertains to the shallow-water problem. The state variable of the model $x = (\zeta, D, \phi)$ consists of vorticity ζ, divergence D, and geopotential ϕ. One may define the intermediate variables

$$\tilde{\zeta} = \nabla \times \frac{\vec{v}-\vec{v}_g}{\sigma_{\vec{v}}} \tag{19}$$

$$\tilde{D} = \nabla \cdot \frac{\vec{v}-\vec{v}_g}{\sigma_{\vec{v}}} \tag{20}$$

$$\tilde{\phi} = \frac{\phi-\phi_g}{\sigma_\phi} \tag{21}$$

Isotropy is assumed for the autocorrelation function of $\tilde{\zeta}$, $\tilde{D}$ and $\tilde{\phi}$. The first-guess term J_b expressed in spectral space becomes

$$J_b = \sum_n \left[\frac{1}{a_\phi^n} \sum_m \tilde{\phi}_n^{m2} + \frac{1}{a_D^n} \sum_m \tilde{D}_n^{m2} + \frac{1}{a_\zeta^n} \sum_m \tilde{\zeta}_n^{m2} \right] \tag{22}$$

The a_ϕ^n, a_D^n and a_ζ^n are the expansion of the autocorellation functions of the fields $\tilde{\zeta}$, $\tilde{D}$, and $\tilde{\phi}$. They can be easily deduced from the grid point values of the autocorrelation function.

2.3.2 *A simple example of the impact of imposing a balance*

Consider in the previous example a single wavenumber. For the (n,m) considered, the matrix $\tilde{B}$ is

$$\tilde{B} = \begin{pmatrix} a_\zeta^n & 0 & 0 \\ 0 & a_D^n & 0 \\ 0 & 0 & a_\phi^n \end{pmatrix} \tag{23}$$

Now assume that the balance condition which is imposed is

$$\begin{cases} \tilde{D}_n^m = 0 \\ \tilde{\phi}_m^n = \alpha \tilde{\zeta}_m^n \end{cases} \tag{24}$$

where α can be any number. The projection operator S_R is then such that the vector $(0,1,0)$ is in the kernel. The vector $(1,0,\alpha)$, being in balance, is kept unchanged and the vector $(-\alpha,0,1)$ is in the kernel. For the latter, it is implicitly assumed that the scalar product is the usual one defined by

$(\tilde{\zeta}_m^n)^2 + (\tilde{D}_m^n)^2 + (\tilde{\phi}_m^n)^2$. Actually there should be some scaling between momentum and mass but it does not change the point being made in this section. The matrix of the projection is then

$$S_R = \frac{1}{1+\alpha^2}\begin{pmatrix} 1 & 0 & \alpha \\ 0 & 0 & 0 \\ \alpha & 0 & \alpha^2 \end{pmatrix} \tag{25}$$

The effective covariance matrix of the background errors becomes

$$S_R \tilde{P} S_R^t = \frac{a_\zeta^n + \alpha^2 a_\phi^n}{(1+\alpha^2)^2}\begin{pmatrix} 1 & 0 & \alpha \\ 0 & 0 & 0 \\ \alpha & 0 & \alpha^2 \end{pmatrix} \tag{26}$$

This has a number of implications:

- Assuming that $a_\zeta^n = a_\phi^n$ and that $\alpha = 1$, which is reasonable for scales close to the Rossby radius of deformation, the effective matrix is half of that which has been specified. In other words, the variance of error which has been specified in grid point space is not that which is actually used, only half of it has an effective contribution.
- Assuming that the specification of the a_ϕ^n and a_ζ^n has been made consistently with the balance equation, that the effective matrix may well depend only on one of the a^n and not on the other for this particular scale.

One could have increments in geostrophic balance by using a univariate background term and controlling independently the gravity waves. The simple example described above shows that such an approach is not an acceptable solution in practice, since one would not be able to deduce from the specified σ_b the value effectively used. It is thus necessary to have the geostrophy embedded within the covariance matrices of the background errors $\underline{\underline{B}}$.

2.4 Implementation of a multivariate J_b

The basic idea is to split the J_b cost function into Rossby and gravity components and penalise the latter, thus ensuring the analysis increments lie close to the tangent plane of the slow manifold. For the higher vertical modes it makes little sense to treat the Rossby and gravity parts differently as the frequencies of the latter are no longer so large. These higher vertical modes may be conveniently treated as univariate. This may be achieved as follows:

1.1 Difference x and x_b in spectral space ΔX

1.2 Project onto Rossby modes for desired subset of vertical modes ∇x_R

1.3 Project onto Gravity modes for same subset of vertical modes Δx_G

1.4 Obtain Δx_U by differencing Δx with Δx_R and Δx_G

$$\Delta x_R = R(x-x_b) \tag{27}$$

$$\Delta x_G = G(x-x_b) \tag{28}$$

$$\Delta x_U = \Delta x - \Delta x_R - \Delta x_G \tag{29}$$

It is convenient to split J_b into "slow", "fast" and "univariate" components:

$$\begin{aligned} J_b = {} & \tfrac{1}{2} c_R \left(\frac{\Delta x_R}{\sigma_R}\right)^t (h_R^{-\frac{1}{2}})^t (V_R^{-\frac{1}{2}})^t V_R^{-\frac{1}{2}} h_R^{-\frac{1}{2}} \left(\frac{\Delta x_R}{\sigma_R}\right) \\ & + \tfrac{1}{2} c_G \left(\frac{\Delta x_G}{\sigma_G}\right)^t (h_G^{-\frac{1}{2}})^t (V_G^{-\frac{1}{2}})^t V_G^{-\frac{1}{2}} h_G^{-\frac{1}{2}} \left(\frac{\Delta x_G}{\sigma_G}\right) \\ & + \tfrac{1}{2} \left(\frac{\Delta x_U}{\sigma_U}\right)^t (h_U^{-\frac{1}{2}})^t (V_U^{-\frac{1}{2}})^t V_U^{-\frac{1}{2}} h_U^{-\frac{1}{2}} \left(\frac{\Delta x}{\sigma_U}\right) \end{aligned} \tag{30}$$

where $(\Delta x/\sigma)$ is a short hand notation for the sequence of operators $(WSNS^{-1}W^{-1})$ as described in section 2.1.

If one defines

$$\chi_R = P_R^{-1} h_R^{-\frac{1}{2}} WSN_R S^{-1}W^{-1}\Delta x_R, \tag{31}$$

$$\chi_G = P_G^{-1} h_G^{-\frac{1}{2}} WSN_G S^{-1}W^{-1}\Delta x_G, \tag{32}$$

$$\chi_U = P_U^{-1} h_U^{-\frac{1}{2}} WSN_U S^{-1}W^{-1}\Delta x_U \tag{33}$$

then, as shown in Appendix A, (30) reduces to

$$J_b = \tfrac{1}{2} c_R \chi_R^t \Lambda_R^{-1} \chi_R + \tfrac{1}{2} c_G \chi_G^t \Lambda_G^{-1} \chi_G + \tfrac{1}{2} \chi_U^t \Lambda_U^{-1} \chi_U$$

For the bulk of the spectrum c_G is set to $\frac{1}{2\varepsilon}$ and c_R to $\frac{1}{2(1-\varepsilon)}$.

ε controls the relative contributions of the first two terms. It may be thought of as the percentage error variance explained by the gravity wave part of the flow. ε is currently taken as 10% which mirrors the operational ECMWF OI assumption that 10% of the wind error variance is in the divergent part of the flow.

For large-scale gravity modes, for example those important in the description of tides c_G is set equal to c_R, thus these modes will be analysed with the same weight as Rossby modes. Note that with this formulation J_b does not exactly imply a univariate analysis.

The J_b gradient is now also split into three terms and the adjoint computations 2) → 8) of section 2.1 have to be carried out for each of the terms in turn:-

$$\nabla_{\Delta x_R} J_{b_R} = (W^{-1})^* (S^{-1})^* N_R^* S^* W^* (h_R^{-\frac{1}{2}})^* (P_R^{-1})^* c_R \chi_R \tag{34}$$

$$\nabla_{\Delta x_G} J_{b_G} = (W^{-1})^* (S^{-1})^* N_G^* S^* W^* (h_G^{-\frac{1}{2}})^* (P_G^{-1})^* c_G \chi_G \tag{35}$$

$$\nabla_{\Delta x_U} J_{b_U} = (W^{-1})^* (S^{-1})^* N_U^* S^* W^* (h_U^{-\frac{1}{2}})^* (P_U^{-1})^* \chi_U \tag{36}$$

the adjoints of 1.1) → 1.4) then gives $\nabla_x J_b$.

The above formulation is fairly general, allowing, in principle, different standard error fields for "fast", "slow" and univariate terms, different horizontal structure functions and different vertical structure functions. For the initial configuration it has been decided to opt for the simplest case of

$$\sigma_U = \sigma_R = \sigma_G, \quad h_U = h_R = h_G, \quad V_U = V_R = V_G$$

which implies $P_U = P_R = P_G$ and $N_U = N_R = N_G$. These constraints can be relaxed as experience dictates.

2.5 Vertical interpolation of fields and effective σ_b

In order to compute observation departures, the model field is interpolated both horizontally and vertically to the observation location. This interpolation can, depending on the location of the observation relative to the model grid and the degree of correlation of background errors, significantly reduce the effective σ_b.

For an observation lying between two model levels with temperature T_1 and T_2 the interpolated model value at the observation point (using linear interpolation) is given by

$$T = \alpha T_1 + (1-\alpha) T_2 \tag{37}$$

and the error in the interpolated value is given by:

$$\varepsilon = \alpha \varepsilon_{T_1} + (1-\alpha) \varepsilon_{T_2} + \varepsilon_p \tag{38}$$

where ε_p is the error in interpolation process.

Assuming ε_p is uncorrelated with ε_{T_1}, ε_{T_2} and $\sigma^2 = <\varepsilon_{T_1}, \varepsilon_{T_1}> = <\varepsilon_{T_2}, \varepsilon_{T_2}>$ and $\beta\sigma^2 = <\varepsilon_{T_1}, \varepsilon_{T_2}>$

and $\sigma_p^2 = <\varepsilon_p, \varepsilon_p>$, with $-1 \le \beta \le 1$,

then $\sigma_b^2 = (\alpha\ 1-\alpha)\begin{pmatrix}1 & \beta\\ \beta & 1\end{pmatrix}\begin{pmatrix}\alpha\\ 1-\alpha\end{pmatrix}\sigma^2 + \sigma_p^2$. (39)

More generally, for interpolation/extrapolation of one vertical profile $\underset{\sim}{x}_2$ from another $\underset{\sim}{x}_1$ using a linear operator $\underset{\sim}{D}$:

$$\underset{\sim}{x}_2 = \underset{\sim}{D} \cdot \underset{\sim}{x}_1 \tag{40}$$

the error covariance of $\underset{\sim}{x}_2$ is given by

$$\underset{\sim}{C}_2 = \underset{\sim}{D} \cdot \underset{\sim}{C}_1 \cdot \underset{\sim}{D}^T + \underset{\sim}{G} \tag{41}$$

where $\underset{\sim}{G}$ is the error covariance of the operation $\underset{\sim}{D}$. This problem is discussed by *Eyre* (1989) in the context of 1D-Var.

If $\underset{\sim}{G}$ is omitted, then one obtains the (usually false) result that the interpolated profile is <u>more</u> accurate than the profile from which it is interpolated. Examining (39), for $\beta = 0$ one finds that

$$\sigma_b^2 = \alpha^2\,\sigma^2 + (1-\alpha)^2\,\sigma^2 \tag{42}$$

which for $\alpha = 0$ or $\alpha = 1$ gives $\alpha_b^2 = \sigma^2$

but for $\alpha = 0.5$, gives $\sigma_b^2 = \frac{1}{2}\sigma^2$.

This indicates that the background is more accurate at intermediate levels. The analysis increments are controlled directly by the ratio $\frac{\sigma_b^2}{\sigma_o^2}$ where σ_o^2 is the observation variance. If σ_b^2 is halved, the observation will have less impact on the analysis. For realistic values of σ_o and σ_b, the analysis increments are smaller by 20%.

For $\beta = 1$ one obtains $\sigma_b^2 = \sigma^2$

and for $\beta = -1$ one obtains $\sigma_b^2 = \sigma^2(1-2\alpha)^2$.

This analysis indicates that an observation lying midway between model levels is given less and less weight as the structure functions become increasingly sharp. Ultimately, as the correlation between adjacent levels becomes negative the observation is ignored.

This is a genuine problem since situations exist where an observation is of no use. It is also a design feature of 3D-Var. It reveals a problem which becomes acute when the structure functions are too sharp compared to the model vertical discretization.

The feature does not show up in the current ECMWF implementation of OI. However, in the ECMWF OI, the background is interpolated at the observation point using (37) but the σ's which have to be explicitly interpolated are not interpolated using (39) but using (37). OI is thus not mathematically consistent (in its implementation!)

One could minimize the impact of the problem by using cubic interpolation in the vertical. One could also use tricks to mimic OI such as explicitly changing σ_o so as to compensate for the changed σ_b (P Undén, personal communication). However, it is not clear if one should do this. Is an observation located at a half level as informative as an observation located at a full level?

It is clear that one should ensure that the vertical structure functions are reasonably resolved by the vertical discretization of the model and, if at some levels they remain too sharp (e.g. at the tropopause), it is a strong argument for having more vertical resolution in the model at such levels. For multilevel observations, one should extract data such that their vertical resolution is consistent with the model vertical resolution. It is worth noting that the problem also occurs in the horizontal but is of smaller magnitude since i) we use bicubic interpolation, and ii) the horizontal correlation of errors are relatively broad.

3. THREE-DIMENSIONAL ANALYSIS EXPERIMENTS

3.1 Experiments with no J_o

The effect of the choice of control variable and balance constraints on speed of convergence is most easily seen in the absence of observations as the solution is known precisely, and one can expect to solve the minimization problem under certain conditions in a single step.

In the absence of observations and any balance constraints (univariate J_b) the solution is $x = x_b$, where $J_b = 0$ and $\nabla J_b = 0$.

The background error field has considerable horizontal and vertical variability - as is necessary for proper representation of the background error variances. When the model state vector in spectral space, x, is chosen as control variable, then even for J_b alone, convergence is achieved rather slowly typically reducing the cost function by ~20% in 30 iterations. This is because it is not possible to precondition the minimization sufficiently well simply by specifying the diagonal elements alone (see section 2.2). If, however, χ, the departure of the model state variables from the guess, normalized by the forecast error standard deviations and projected onto the eigenvectors of the vertical background correlation matrices, is chosen as control variable then (again, for J_b alone) exact convergence to machine precision is achieved in a single step of the minimization scheme as in this case the Hessian is diagonal and the correct pre-conditioning may be applied

The choice of χ as control variable in the multivariate formulation of J_b no longer ensures the Hessian is diagonal and so convergence is not achieved in a single iteration, even for J_b alone. However, it is still much better, in terms of conditioning, than using x itself. If the ε in J_b is chosen sufficiently small the optimal choice of control variable would become $P^{-1} h^{-\frac{1}{2}} \left(\frac{S_R(x-x_b)}{\sigma} \right)$, which of course, consistently, implies an analysis of the slow modes alone. Some experimentation will be necessary to determine an optimum choice for the more general formulation.

3.2 Single observation experiments

The multivariate balance imposed by J_b as described in section 2 provides a mass wind coupling over the whole globe. In the examples which follow all vertical modes are used in the separation between Rossby and gravity waves. Horizontal modes used are those corresponding to vertical modes 1 through 7. Beyond vertical mode 7 the same set of horizontal modes (those associated with vertical mode 7) are used for all higher modes. Note that there is no univariate component in this example (although such a component is allowed for in the general formulation as described in section 2.4). Fig. 1 shows the response to an isolated observation at 60° N of a) height, b) zonal wind, and c) meridional wind. In each case the analysis increment is near geostrophic. The horizontal scale is determined by the horizontal structure functions which are described in Appendix C and the vertical spread by the vertical structure functions described in Appendix D.

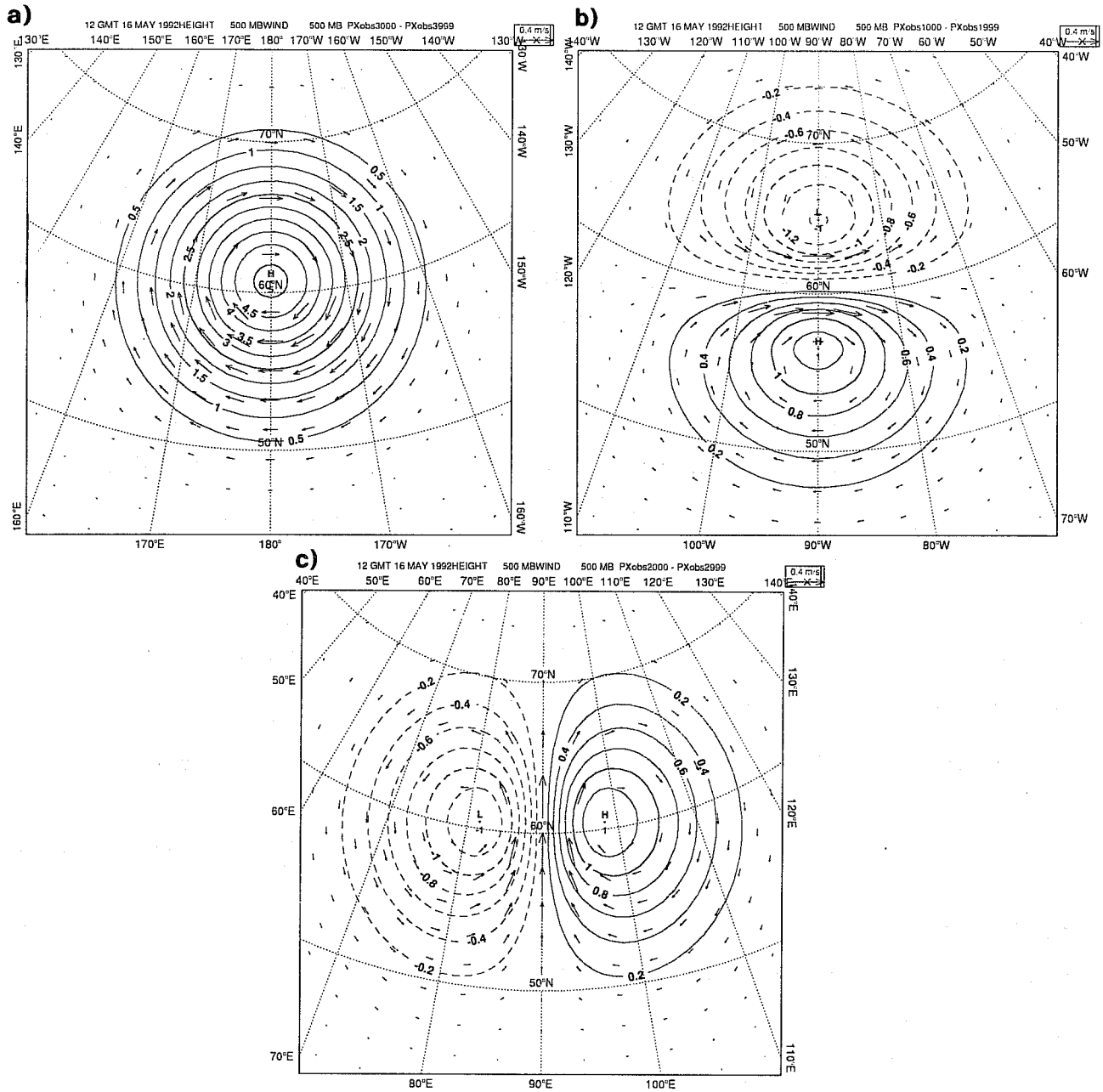

Fig.1 Response of the 3D-Var analysis at 500 hPa to an isolated observation at 60°N. Hough balance, $\varepsilon = 0.1$. a) height, b) zonal wind, c) meridional wind. Contours are of height and the arrows indicate vector wind. Polar stereographic projection centred on the observation location.

One of the limitations of the ECMWF OI analysis is the lack of mass/wind balance as one approaches the equator - the scheme becomes univariate at the equator. Fig. 2 shows the response to an isolated observation at the equator of a) positive zonal wind, b) negative zonal wind, and c) a southerly meridional wind. The variational analysis has a strong mass-wind balance even at 0°. Note, however, the absence of a Kelvin wave response, in the current formulation these are taken as "fast" modes and assigned relatively large errors. As *Parrish* (1988) and *Daley* (1993) point out, Rossby modes imply a negative u, ϕ correlation at the equator, whereas Kelvin modes imply a positive u, ϕ correlation at the equator. The addition of Kelvin modes in the "slow" term of J_b will considerably reduce the u, ϕ correlations at the equator (*Parrish*, 1988).

Fig. 3 shows the linear balance response to an isolated zonal wind observation at 0°N. Compare this with the Hough balance shown in Fig. 2a. The similarity with the linear solution is, of course, affected by the number of vertical modes used in the Rossby/gravity separation - in this case 7. Closest similarity with the linear solution is obtained when the external mode is used for the separation as Hough balance becomes linear in the limit of infinite equivalent depth. One can expect there to be sensitivity of the increments to the details of the Rossby/gravity/univariate separation, again this will have to be an area of further research.

Finally, it is worth noting the effect of varying the ε parameter in (30). In all the above $\varepsilon = 0.1$, which implies that 10% of the error variance lies in the gravity wave part of the fields. Fig. 4a shows the effect, for a single zonal wind observation at 0°N of $\varepsilon = 0.5$, and Fig. 4b of $\varepsilon = 0.9$, c.f. Fig. 2a which is the $\varepsilon = 0.1$ case. The balance changes from near geostrophic with $\varepsilon = 0.1$, to almost entirely a geostrophic with $\varepsilon = 0.9$. As discussed in section 2.4, the ε parameter is analogous to the OI formulation in which one assumes that a certain percentage of the variance is described by the divergent component of the wind. *Daley's* (1983) experiments indicated that 10% was a reasonable figure for this, and following further experimentation by *Undén* (1989) this is the value used operationally by the ECMWF OI. Its role is slightly different with the variational analysis and this is another area which may benefit from a closer study.

4. FUTURE DEVELOPMENT

4.1 Other approaches to dynamical balance

Within the variational approach it is possible to include balance constraints in a number of different ways, one method in which the background constraint is explicitly formulated in terms of fast and slow modes has already been introduced in section 2.4.

Two other approaches have also been implemented as options:

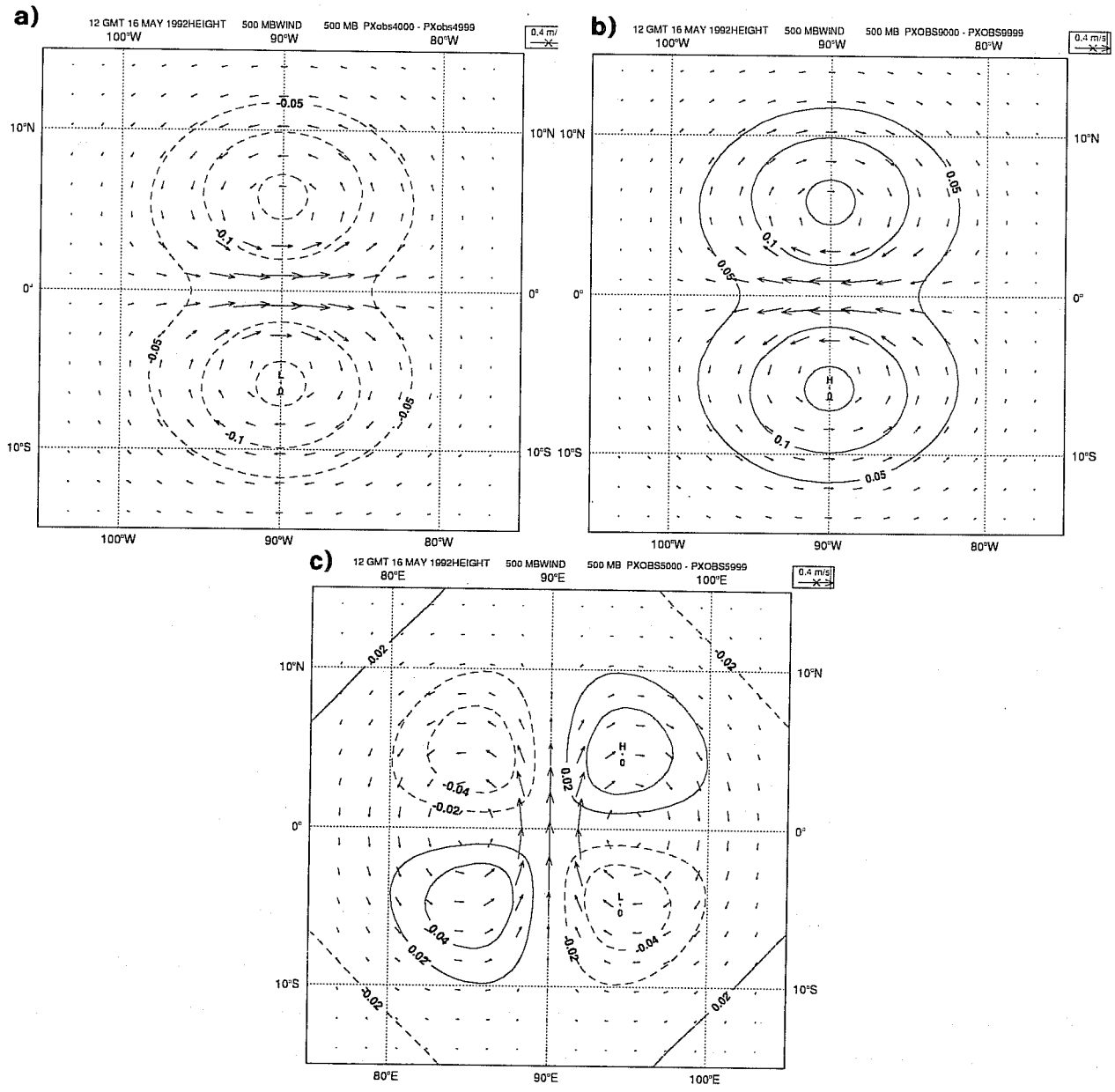

Fig.2 Response of the 3D-Var analysis at 500 hPa to an isolated observation at 0°N. Hough balance ε = 0.1. a) positive zonal wind, b) negative zonal wind, c) southerly meridional wind. Contours are of height and the arrows indicate vector wind. Regular latitude/longitude projection.

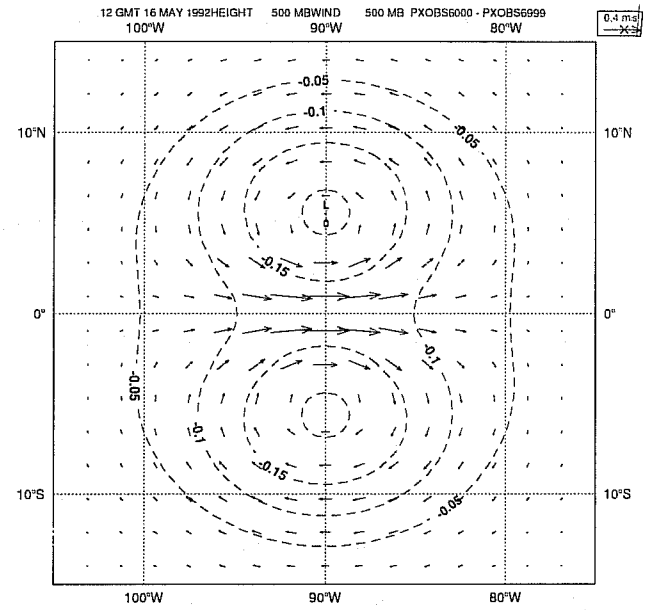

Fig.3 As Fig.2(a) but for linear balance.

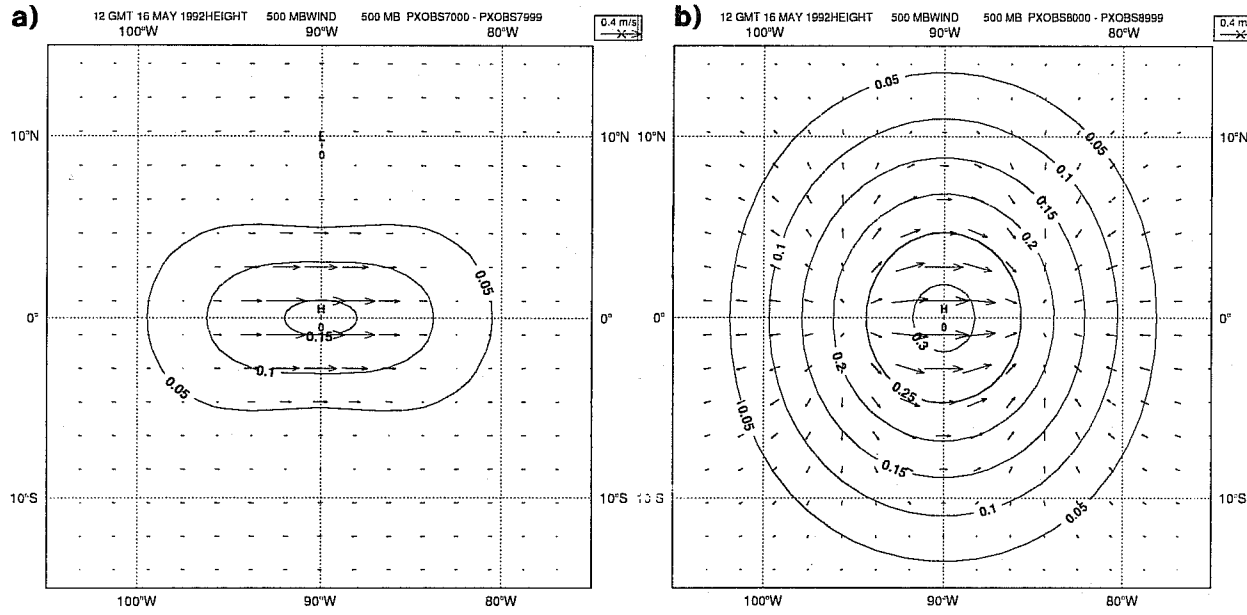

Fig.4 As Fig.2(a) but for a) $\varepsilon = 0.5$ and b) $\varepsilon = 0.9$

4.1.1 *NNMI in the cost function*

The cost functions may be formulated in terms of initialized fields, $NMI(x)$, so that

$$J_b = (NMI(x) - x_b)^t B^{-1}(NMI(x) - x_b)$$

$$J_o = (H(NMI(x)) - y)^t O^{-1}(H(NMI(x)) - y)$$

This involves a change of variable by performing a non-linear mode initialization (*NNMI*) on x, the adjoint of which is needed in calculating the gradient of the cost function with respect to x. Implementation is straightforward as the *NNMI* and its adjoint are simply operators which are applied at the appropriate points in the chain.

It has been found in *Courtier and Talagrand* (1990) and *Thépaut and Courtier* (1991) that this approach acts to speed the convergence. However, if the number of iterations increases, the *NNMI* process is inverted by the minimization and the minimizing solution contains gravity components.

Also, this formulation leaves a certain amount of ambiguity in the control variable itself since both x and χ in general contain gravity wave components. The minimization will attempt to fit the slow component of x to x_b through modification of the control variable. In practice x seems to reach the slow component of x_b rather quickly, within 5 or 6 iterations. Thereafter the minimization has difficulty in reconstructing the 'fast' components in x_b, and in the case of a 'noisy' x_b may never succeed.

The balance condition which is imposed through *NNMI* imposes a geostrophic coupling only for the vertical modes included in the *NNMI* (at ECMWF this is five). As the amplitude of these modes is large mainly in the stratosphere and for the largest vertical scales in the troposphere, the analysis (even if it is free of gravity waves) becomes univariate close to the boundary layer: increments on temperature in the low troposphere leads to small increments of wind.

Use of $NMI(x)$ in the cost function will not automatically provide a balanced final state, but it will usually speed up the initial rate of convergence since the minimization acts in a subspace of x (the slow modes). Balance considerations may be addressed through a further, weak constraint J_c described below.

4.1.2 J_c

The second approach consists of introducing a cost function J_c which contains a penalty term on the tendency of gravity modes G

$$J_c = \alpha_c |dG/dt|^2$$

This is carried out by computing the tendency of the gravity modes of the analyzed state through one time-step of the model; and the adjoint through one timestep of the corresponding adjoint model.

Consider an example where the background, x_b, and the starting point for the minimization (first-guess), x_b, are two initialized analyses 24 hours apart. These have been truncated to T21 from T106. Both fields have been operationally diabatically initialized, whereas the variational analysis currently only uses adiabatic initialization, therefore the fields are not fully 'balanced' for this model. The control variable used for this example is the model state vector x (with horizontally constant forecast error variances). It is found that, with only one iteration of *NNMI*, the amount of gravity wave activity increases rapidly. Two iterations of *NNMI* is sufficient to control it to a reasonable level. Increasing the number of *NNMI* iterations beyond this point has relatively little impact upon the amount of gravity wave activity.

Fig. 5 shows the cost function gradient as a function of the number of simulations. Most rapid convergence occurs when *NNMI* is used alone, in this case x moves towards the slow components of x_b relatively quickly. Slowest convergence is found when using J_c alone, where the minimization is trying to consider the whole phase space of the control variable. Combining *NNMI* with J_c seems to give the benefits of both - enhanced initial convergence plus an explicit control of gravity waves.

J_c is a constraint on the gravity mode tendencies and can be thought of as a progressive *NNMI* applied to x. J_c may complement the use of *NMI(x)* in the computation of the J_b and J_o cost functions by introducing a constraint on the fast modes which is absent from these terms.

The use of *NNMI* in the J_o and J_b terms of the cost function and together with a J_c term attempts to produce analysis increments tangent to the slow manifold as defined by *NNMI*. As the curvature of the slow manifold is small in midlatitudes, the impact expected in a purely geostrophic analysis is small. One could, of course, have 10 cheap iterations with no *NNMI* or J_c followed by 10 expensive iterations with *NNMI* instead of having the 20 iterations all performed with *NNMI*. As 3D-Var is significantly less expensive without *NNMI* and J_c than with them, their use should be avoided unless the benefit is significant.

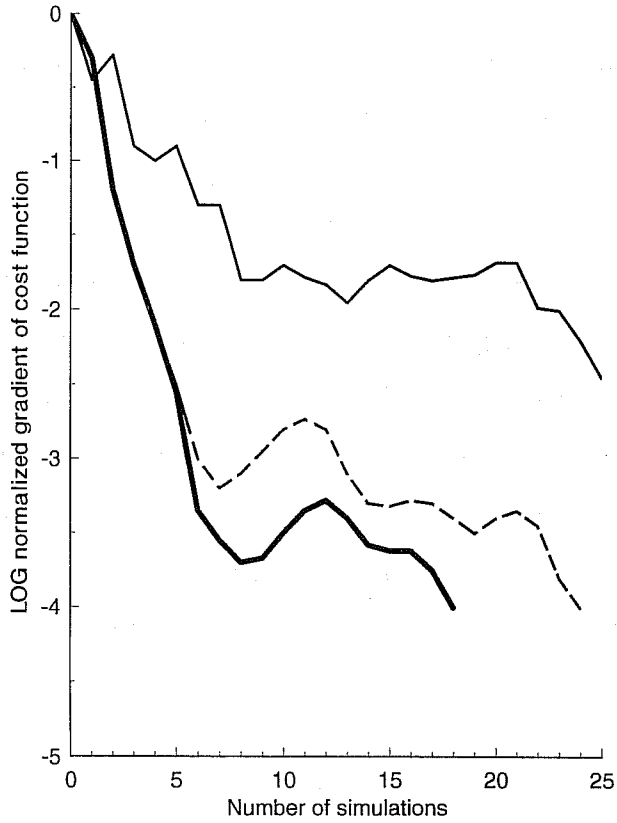

Fig.5 Decimal logarithm of the square of the gradient norm (normalized by the initial value) for different experiments: NMI alone, (thick solid line), J_c alone (thin solid line), NMI+J_c (dashed). In each case $x \neq x_b$ and J_o

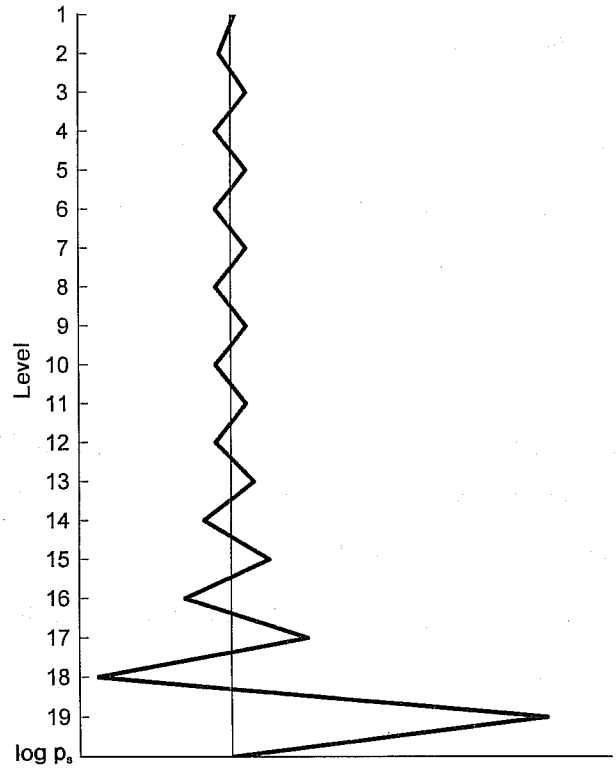

Fig.6 Structure of the kernel of the transform $v' = \underline{\underline{E}}\,\underline{v}$.

The addition of constraints also complicates the minimization issue. For the most efficient minimization one should include the Hessian of the ***NNMI*** (and of J_c, which will be similar) in the pre-conditioning. No attempt to do this has so far been made.

4.2 Analysis of the kernel

In view of the problems experienced in diagnosing the kernel of the $P \rightarrow T$, lnp_s transform (Appendix B), it is clear that a more satisfactory approach would be to analyze the kernel directly, i.e. make it a part of the control variable. This would have the advantage of extracting more information from the observations. Preliminary attempts show that the technique is about as successful in generating a known profile as is the diagnosed kernel approach, but that there is a strong sensitivity to the formulation of the scalar product. This requires some thought as one would not want to upset the J_b pre-conditioning.

5. CONCLUSIONS

The general formulation of the ECMWF variational analysis system has been described. It is shown how the cost function J is split into a number of components J_b, J_o and J_c. This paper has concentrated on general aspects of the formulation and the detailed implementation of the background term J_b.

It is emphasized that computational efficiency is of paramount importance. The background constraint is formulated in such a way as to ensure that the Hessian is as close to diagonal as can reasonably be achieved, thus enabling suitable pre-conditioning for the minimization. Appropriate choice of control variable for the minimization is found to have a substantial impact upon the rate of convergence. It is demonstrated how by a suitable choice of control variable one may account for geographical variability of the background errors, and vertical correlation of the errors without in any way deteriorating the convergence properties of the scheme. A conscious decision has been taken not to work on minimization algorithms 'in-house'. Instead general minimization packages are being used as developed by INRIA (*Gilbert and Lemaréchal*, 1989), the algorithm currently being used is the M1QN3 package.

Mass wind balance may be treated through a number of different mechanisms - for example use of normal mode initialization and its adjoint, or through a penalty term which keeps the tendency of gravity waves small. These two approaches impose balance independent of the form of the background constraint J_b. It is shown that this is not an acceptable solution in practice since it is not then possible to deduce from the specified background errors the value that has effectively been used. It is shown that it is necessary for the balance constraints to be imbedded (at least to a large extent) within the covariance matrices of the

background errors $\underline{\underline{B}}$. The implementation of this constraint is described in detail and the resulting multivariate increments illustrated through the use of the single-observation experiments.

Other aspects of the use of *NNMI* are discussed. Use of *NNMI* in the cost function acts to speed convergence since the minimization acts in a subspace (the slow modes). With the penalty term on gravity modes initial convergence is generally slower than when *NNMI* is used in the cost function (because minimization is no longer primarily acting on the slow modes) but, when used in combination with *NNMI* in the cost function, convergence is not unduly affected. It does, however, act to progressively dampen the gravity wave tendencies and introduce a degree of balance into the fields. This result is consistent with the findings of *Courtier and Talagrand* (1990) using a shallow water model, and *Thépaut and Courtier* (1991) using a 3D primitive equation model.

Specification of the background error statistics is described in detail. A simple, yet flexible, parametric form has been introduced for the horizontal background error spectrum. The vertical covariances are based on those currently used in the ECMWF OI analysis.

The scheme as currently implemented offers a suitable and convenient basis for future development. Obvious avenues to be explored including tuning of the horizontal length scales perhaps including a vertical variation. Wave number dependence of the vertical correlations. Choice of number of vertical modes used for the separation of Rossby/gravity components. Inclusion of Kelvin modes in the "slow" terms. Direct analysis of the kernel. Differing statistics for Rossby/gravity terms. Experience will dictate which of these are important, and indeed, other issues will undoubtedly emerge.

Acknowledgements

Many people have helped at one time or another directly or indirectly with this work. We would like to thank in particular Jean-Noël Thépaut who coded the J_c formulation. Tony Hollingsworth for stimulating discussions on the J_b formulation. Florence Rabier who found some bugs. Jan Haseler and Per Undén for help with the ECMWF OI system. P. Marquet and D. Giard from Météo-France who coded most of the NMI routines. Last, but not least, Draško Vasiljević who coded most of the J_o components.

APPENDIX A: FORMULATION OF THE BACKGROUND ERROR COVARIANCE MATRIX $\boldsymbol{B}$

A.1 Decomposition of $\boldsymbol{B}$

The model state vector $\underline{x}$ is given by:

$$\underline{x} = \begin{pmatrix} \underline{\xi} \\ \underline{D} \\ \underline{P} \\ \underline{q} \end{pmatrix}$$

where

$$\underline{\xi} = \begin{bmatrix} \begin{pmatrix} \xi_1 \\ \cdot \\ \cdot \\ \cdot \\ \xi_L \end{pmatrix}_{\textit{spectral coefficient 1}} \\ \cdot \\ \cdot \\ \begin{pmatrix} \xi_1 \\ \cdot \\ \cdot \\ \cdot \\ \xi_L \end{pmatrix}_{\textit{spectral coefficient N}} \end{bmatrix}$$

L model η levels

N spectral coefficients (A1)

4 variables

Dimension of $\underline{x} = 4 \times N \times L$

Let $x'_R = R\Delta x/\sigma_R$, $x'_G = G\Delta x/\sigma_G$ and $x'_U = (\Delta x - R\Delta x - G\Delta x)$, where $\Delta x = x - x_b$ and $\boldsymbol{R}$ is a projection operator onto the "slow" modes, $\boldsymbol{G}$ is a projection operator onto "fast" modes. This is a shorthand notation for the sequence of operators $(WSNS^{-1}W^{-1})$ as described in section 2.1 and 2.4. x'_R, x'_G and x'_U are then spectral space variables describing normalised departures from the background field.

In terms of x', the background constraint given by (36) may be expressed as:-

$$\begin{aligned} J_b = {} & \tfrac{1}{2}\, c_R\, \underline{x}'^{t}_R \left(\underline{h}_R^{-\frac{1}{2}}\right)^t \left(\underline{V}_R^{-\frac{1}{2}}\right)^t \underline{V}_R^{-\frac{1}{2}}\, \underline{h}_R^{-\frac{1}{2}}\, \underline{x}'_R \\ & + \tfrac{1}{2}\, c_G\, \underline{x}'^{t}_G \left(\underline{h}_G^{-\frac{1}{2}}\right)^t \left(\underline{V}_G^{-\frac{1}{2}}\right)^t \underline{V}_G^{-\frac{1}{2}}\, \underline{h}_G^{-\frac{1}{2}}\, \underline{x}'_G \\ & + \tfrac{1}{2}\, \underline{x}'^{t}_U \left(\underline{h}_U^{-\frac{1}{2}}\right)^t \left(\underline{V}_U^{-\frac{1}{2}}\right)^t \left(\underline{V}_U^{-\frac{1}{2}}\right)^t \left(\underline{h}_U^{-\frac{1}{2}}\right) \underline{x}'_U \end{aligned} \qquad \text{(A2)}$$

which assumes separability between the horizontal and "vertical". Because of the normalisation by the σ's,

$\underline{\underline{h}}$ and $\underline{\underline{V}}$ contain only correlations. The 'c' terms are there to provide a relative weighting between the first two terms (see section 2.4).

Since J_b is split into three components of identical form it is only necessary to consider one of them in detail. In the following x', $\underline{\underline{h}}$ and $\underline{\underline{V}}$ can be taken as referring to either the 'R', 'G' or 'U' terms.

$\underline{\underline{h}}$ takes the form

$$\underline{\underline{h}} = \begin{bmatrix} \underline{\underline{h}}_{\xi} & 0 & 0 & 0 \\ 0 & \underline{\underline{h}}_{D} & 0 & 0 \\ 0 & 0 & \underline{\underline{h}}_{P} & 0 \\ 0 & 0 & 0 & \underline{\underline{h}}_{q} \end{bmatrix} \tag{A3}$$

where $\underline{\underline{h}}_{\xi} = \begin{bmatrix} \underline{\underline{h}}_{\xi 1} & 0 & & \\ 0 & \underline{\underline{h}}_{\xi 2} & & \\ & & \ddots & \\ & & & \underline{\underline{h}}_{\xi N} \end{bmatrix}$ (A4)

and $\underline{\underline{h}}_{\xi n} = \begin{bmatrix} h_1 & 0 & & \\ 0 & h_2 & & \\ & & \ddots & \\ & & & h_L \end{bmatrix}$ (A5)

$\underline{\underline{h}}_{\xi n}$ is of dimension L, $\underline{\underline{h}}_{\xi}$ of dimension $N \times L$ and $\underline{\underline{h}}$ of dimension $4 \times N \times L$

This, in principle, allows the horizontal structure functions to vary in the vertical and also for each model variable.

The vertical term $\underline{\underline{V}}$ has the form

$$\underline{\underline{V}} = \begin{bmatrix} \underline{\underline{V}}^{\xi\xi} & \underline{\underline{V}}^{\xi D} & \underline{\underline{V}}^{\xi P} & \underline{\underline{V}}^{\xi q} \\ \underline{\underline{V}}^{D\xi} & \underline{\underline{V}}^{DD} & \underline{\underline{V}}^{DP} & \underline{\underline{V}}^{Dq} \\ \underline{\underline{V}}^{P\xi} & \underline{\underline{V}}^{PD} & \underline{\underline{V}}^{PP} & \underline{\underline{V}}^{Pq} \\ \underline{\underline{V}}^{q\xi} & \underline{\underline{V}}^{qD} & \underline{\underline{V}}^{qP} & \underline{\underline{V}}^{qq} \end{bmatrix} \tag{A6}$$

where, for the moment, it is assumed that $\underline{\underline{V}}^{ij} = [0]$ for $i \neq j$, thus $\underline{\underline{V}}$ is block diagonal, with elements $\underline{\underline{V}}^{\chi\chi}$ where $\chi = \xi, D, P, q$.

$$\underline{\underline{V}}^{\chi\chi} = \begin{bmatrix} [V_{ij}^{\chi}]_1 & & & \\ & [V_{ij}^{\chi}]_2 & & \\ & & \ddots & \\ & & & [V_{ij}^{\chi}]_N \end{bmatrix} \tag{A7}$$

where it has been assumed $V_{ii} = 1$ and $V_{ij} = V_{ji}$.

$[V_{ij}^{\chi}]$ is the vertical background error correlation matrix for variable χ, and is of dimension L.

$\underline{\underline{V}}^{\chi\chi}$ is of dimension $L \times N$, and $\underline{\underline{V}}$ of dimension $4 \times L \times N$.

In principle, this form allows $[V_{ij}^{\chi}]$ to vary with horizontal wavenumber (n,m). Note, however, that this would imply non-separable structure functions when used in combination with vertical variations in length scale.

The background operator B has the form

$$B = \underline{\underline{h}}^{1/2}\ \underline{\underline{V}}^{1/2}\ (\underline{\underline{V}}^{1/2})^t\ (\underline{\underline{h}}^{1/2})^t \tag{A8}$$

from which one obtains the expression for B^{-1}

$$B^{-1} = (\underline{\underline{h}}^{-1/2})^t\ (\underline{\underline{V}}^{-1/2})^t\ \underline{\underline{V}}^{-1/2}\ \underline{\underline{h}}^{-1/2} \tag{A9}$$

as used in (14).

$\underline{\underline{h}}$ is a diagonal matrix of dimension $4 \times L \times N$. The inverse of a diagonal matrix $[h_{ii}]$ is simply a diagonal matrix $\left[\frac{1}{h_{ii}}\right]$

Thus $\underline{h}^{-\frac{1}{2}} = \begin{bmatrix} \frac{1}{h^{\frac{1}{2}}} & & \\ & . & \\ & & . \end{bmatrix} = \begin{bmatrix} k_1 & & \\ & . & \\ & & k_N \end{bmatrix} = \underline{K}$ where $k_{ij} = \frac{1}{h_{ij}^{\frac{1}{2}}}$ (A10)

$$\underline{K} = \begin{bmatrix} \underline{k}_\xi & 0 & 0 & 0 \\ 0 & \underline{k}_D & 0 & 0 \\ 0 & 0 & \underline{k}_P & 0 \\ 0 & 0 & 0 & \underline{k}_q \end{bmatrix} \text{ where } \underline{k}_\xi = \begin{bmatrix} \underline{k}_{\xi 1} & 0 & . & . \\ 0 & \underline{k}_{\xi 2} & & \\ . & & . & \\ . & & & \underline{k}_{\xi N} \end{bmatrix} \text{ and } k_{\xi n} = \begin{bmatrix} k_1 & 0 & . & . \\ 0 & k_2 & . & . \\ . & . & . & \\ . & . & . & k_L \end{bmatrix} \quad \text{(A11)}$$

$k_{\chi n}$ is of dimension L, $\underline{k}_\xi$ of dimension $N \times L$ and $\underline{K}$ of dimension $4 \times N \times L$.

$$K = \underline{h}^{-\frac{1}{2}}, \text{ and } k_1, \ldots k_N = \frac{1}{h_1^{\frac{1}{2}}} \ldots \frac{1}{h_N^{\frac{1}{2}}}$$

In a similar way, $\underline{V}$ is a block diagonal matrix

$$\underline{V} = \begin{bmatrix} \underline{V}^{\xi\xi} & 0 & 0 & 0 \\ 0 & \underline{V}^{DD} & 0 & 0 \\ 0 & 0 & \underline{V}^{PP} & 0 \\ 0 & 0 & 0 & \underline{V}^{qq} \end{bmatrix} \quad \text{(A12)}$$

The inverse $\underline{V}^{-1}$ is given by

$$\underline{U} = \underline{V}^{-1} = \begin{bmatrix} \underline{V}^{\xi\xi^{-1}} & 0 & 0 & 0 \\ 0 & \underline{V}^{DD^{-1}} & 0 & 0 \\ 0 & 0 & \underline{V}^{PP^{-1}} & 0 \\ 0 & 0 & 0 & \underline{V}^{qq^{-1}} \end{bmatrix} \quad \text{(A13)}$$

But $\underline{V}^{\chi\chi}$ is as given by A7.

If $[V_{ij}^{\chi}]$ is the vertical prediction error correlation matrix for variable χ of dimension L and its inverse is $[U_{ij}]$ then

$$\underline{U}^{xx} = \underline{V}^{xx^{-1}} = \begin{bmatrix} [U_{ij}] & 0 & 0 \\ 0 & [U_{ij}] & \cdot \\ 0 & 0 & \cdot \\ \cdot & \cdot & \end{bmatrix} \tag{A14}$$

$$\underline{B}^{-1} = (\underline{h}^{-\frac{1}{2}})^t \, (\underline{V}^{-\frac{1}{2}})^t \, \underline{V}^{-\frac{1}{2}} \, \underline{h}^{-\frac{1}{2}} \tag{A15}$$

which becomes

$$\underline{B}^{-1} = \underline{K}^t \, (\underline{U}^{\frac{1}{2}})^t \, \underline{U}^{\frac{1}{2}} \, \underline{K} \tag{A16}$$

and the cost function (A2) becomes

$$J = \tfrac{1}{2} \, \underline{x}'^t \, \underline{K}^t \, (\underline{U}^{\frac{1}{2}})^t \, \underline{U}^{\frac{1}{2}} \, \underline{K} \, \underline{x}' \tag{A17}$$

or $$J = \tfrac{1}{2} \, (\underline{U}^{\frac{1}{2}} \underline{K} \underline{x}')^t \, (\underline{U}^{\frac{1}{2}} \underline{K} \underline{x}') \tag{A18}$$

The equation may be simplified by using the eigenvectors of the matrix U.

$\underline{U}$ is real and symmetric from which it follows that the eigenvalues λ_j are real and the eigenvectors are orthogonal.

$$\underline{R}^{-1} \, \underline{U} \, \underline{R} = \underline{\Delta} = \begin{pmatrix} \lambda_1 & 0 & . & . \\ 0 & \lambda_2 & . & . \\ . & . & . & . \\ . & . & \lambda_L & \end{pmatrix} \tag{A19}$$

where $\boldsymbol{R}$ contains as its columns the L, linearly independent (and in this case) orthogonal eigenvectors of $\underline{U}$.

If the eigenvectors are normalised then $\boldsymbol{R}$ becomes an orthonormal matrix and

$$\underline{R}^{-1} = \underline{R}^t$$

In terms of the eigenvectors $\underline{R}$, $\underline{U}$ is given by

$$\underline{U} = \underline{R} \, \underline{\Delta} \, \underline{R}^t \tag{A20}$$

and $$\underline{U}^{\frac{1}{2}} = \underline{\Delta}^{\frac{1}{2}} \, \underline{R}^t \tag{A21}$$

Substituting this expression into (A18) gives

$$J = \tfrac{1}{2}(\underline{\Delta}^{\frac{1}{2}} \, \underline{R}^t \, \underline{K} \, \underline{x}')^t \, (\underline{\Delta}^{\frac{1}{2}} \, \underline{R}^t \, \underline{K} \, \underline{x}') \tag{A22}$$

Let $$\chi = \underline{\Delta}^{\frac{1}{2}} \, \underline{R}^t \, \underline{K} \, \underline{x}' \tag{A23}$$

Then $$J = \tfrac{1}{2} \, \chi^t \, \chi \tag{A24}$$

χ is simply the projection of $\underline{\underline{K}}\, x'$ onto the eigenvectors of the vertical background error correlation matrix multiplied by the square root of the eigenvalues of the latter.

If this procedure is carried out for the "slow", "fast" and univariate terms of J_b in turn, one obtains

$$J_b = \tfrac{1}{2}\, c_R\, \chi_R^t\, \chi_R + \tfrac{1}{2}\, c_G\, \chi_G^t\, \chi_G + \tfrac{1}{2}\, \chi_U^t\, \chi_U \tag{A25}$$

A.2 Computational form of J_b

The univariate form is computed as:

$$J = \tfrac{1}{2} \sum_{\chi}^{\xi,D,P,q} \sum_{l=1}^{L} \sum_{n=1}^{N} \chi_{n,l}^2 \tag{A26}$$

$$\nabla_\chi J = \begin{bmatrix} \begin{pmatrix} \xi_1 \\ \cdot \\ \cdot \\ \cdot \\ \xi_L \end{pmatrix}_{\text{spectral coefficient 1}} \\ \vdots \\ \begin{pmatrix} \xi_1 \\ \cdot \\ \cdot \\ \cdot \\ \xi_L \end{pmatrix}_{\text{spectral coefficient } N} \end{bmatrix} \tag{A27}$$

The multivariate form is computed as:

$$\begin{aligned} J = {} & \tfrac{1}{2}\, c_R \left(\sum_{l=1}^{L} \sum_{n=1}^{2} (\chi_R^2)_{n,l} \right)_{\chi=p} + \tfrac{1}{2}\, c_R \left(\sum_{l=1}^{L} \sum_{n=1}^{N} (\chi_R^2)_{n,l} \right)_{\chi=\xi} \\ & + \tfrac{1}{2}\, c_G \sum_{\chi}^{D,P} \left(\sum_{l=1}^{L} \sum_{n=1}^{N} (\chi_G^2)_{n,l} \right) \\ & + \tfrac{1}{2} \sum^{G,D,P,q} \sum_{l=1}^{L} \sum_{n=1}^{N} (\chi_U^2)_{n,l} \end{aligned} \tag{A28}$$

q is, of course, included in the univariate part χ_U.

It is convenient to analyse the Rossby part in terms of vorticity as one is then sure of the value of the background error being applied, if all terms were present the effective background error would become scale-

dependent since at some scales the wind error term would dominate and at others the mass term would. Similar remarks apply for the gravity part. It is, however, necessary to include the mass term for $n = 0,1,2$ as the vorticity alone is insufficient for these wavenumbers.

If one wishes to analyse certain gravity modes, e.g. tides, then it is straightforward to transfer such components to the univariate term, in which case they will automatically gain equal weight with the Rossby components.

APPENDIX B: TRANSFORMATIONS BETWEEN $\underline{P}$ AND $\underline{T}$, $\underline{\ln p_s}$

The model uses a linearised mass variable $\underline{P}$ where $\underline{P}=\underline{\Phi} + R_d\overline{T}\,\underline{\ln p_s}$, where $\underline{\Phi}$ is linearized geopotential height and $\overline{T}$ is a (constant) reference temperature. As the intention is to use the model's Hough modes to distinguish between balanced and unbalanced components of the fields it is convenient to work in terms of $\underline{P}$ rather than temperature and log surface pressure.

However, the advantage of the vertical coupling is not gained without some loss. There are $(L+1)$ degrees of freedom in the $\underline{T}$, $\underline{\ln p_s}$ combination, but only L in $\underline{P}$, where L is the number of model levels. Clearly there is no problem in defining $\underline{P}$ from $\underline{T}$ and $\underline{\ln p_s}$, but the transformation from $\underline{P}$ to $\underline{T}$ and $\underline{\ln p_s}$ involves a degree of arbitrariness. It is interesting to study this latter transform in more detail. The transform from $\underline{T}$ and $\underline{\ln p_s}$ to $\underline{P}$ may be expressed as $\underline{P} = \underline{\underline{G}}\,\underline{V}$ where $\underline{\underline{G}}$ is a L by $(L+1)$ matrix and $\underline{V}$ is a vector of dimension $(L+1)$ containing L temperature values plus log surface pressure, $\underline{P}$ is a vector of dimension L. $\underline{\underline{G}}$ basically contains the linearized hydrostatic integral of temperature to obtain geopotential height. In a similar way, the inverse transform may be expressed as $\underline{V} = \underline{\underline{H}}\,\underline{P}$, where $\underline{\underline{H}}$ is a $(L+1)$ by L matrix. The transform from $\underline{T}$, $\underline{\ln p_s}$ to $\underline{P}$ and back to $\underline{T}$, $\underline{\ln p_s}$ may be written $\underline{V}' = \underline{\underline{E}}\,\underline{V}$ where $\underline{\underline{E}} = \underline{\underline{H}}\,\underline{\underline{G}}$, and $\underline{\underline{E}}$ is of dimension $(L+1)$ by $(L+1)$. If one calculates the eigenvalues and eigenvectors of the matrix $\underline{\underline{E}}$ one finds L eigenvalues equal to 1 and one eigenvalue equal to zero. The latter eigenvalue is associated with the kernel, or "nullspace", of the matrix $\underline{\underline{E}}$. The structure of the kernel, shown in Fig. 6, describes the information which is lost in the transformation from $\underline{T}$, $\underline{\ln p_s}$ to $\underline{P}$. This information cannot be reinstated by the transform from $\underline{P}$ to $\underline{T}$, $\underline{\ln p_s}$ and the final field is characterised by a zero projection onto the kernel. As an example, if one takes the following temperature profile, as shown by the solid line in Fig. 7, and applies the operator $\underline{\underline{E}}$, one obtains the result shown by the dotted line in Fig. 7. The temperature difference between the two profiles at model level 19 is 159° C!

Clearly, this is not a very good description of the original temperature profile. The information lost is vital to a correct description of the profile.

The problem is how to deduce the correct amplitude of the kernel. In the above example there is a clear solution: one calculates the amplitude of the projection of the original field onto the kernel, apply the

operator E, then simply add back the kernel with its original amplitude. This technique reproduces the **original profile to within machine accuracy.**

If one does not know the original T, $\ln(p_s)$ field (which is the case with variational analysis since the control variable is a function of P) then the amplitude of the kernel has to be diagnosed in some way. In order to do so it is necessary to close the problem by applying an additional constraint. An obvious choice is one which minimizes the second derivative of temperature in the vertical. Define a matrix S such that

$$\underline{S} = \begin{pmatrix} 1+a_2 & -2 & 1-a_2 & 0 & . & . \\ 0 & 1+a_3 & -2 & 1-a_3 & . & . \\ 0 & 0 & 1+a_4 & -2 & . & . \\ 0 & 0 & 0 & 1+a_5 & . & . \\ . & . & . & . & . & . \\ . & . & . & . & . & . \\ 0 & 0 & 0 & 1+a_{n-1} & -2 & 1-a_{n-1} \end{pmatrix}$$

where $\underline{S}$ is of dimension L by $(L-2)$. The 'a' terms are introduced by the irregular spacing of the eta-levels. A measure of the 'noise' in the profile is then given by $J = T'^t \underline{S}^t \underline{S} T'$, where T' is the temperature profile as derived from P. What is the amplitude of the kernel such as to minimize J? Define a vector L containing the first L elements of the kernel K (the one dropped is that operating on surface pressure), then the problem may then be expressed:-

find the value c such that $\frac{\partial}{\partial c} (T' - cL)^t \underline{S}^t \underline{S} (T' - cL) = 0$. This is a linear equation in c, the root of which is easily determined. A little algebra leads to the result that $c = \frac{T'^t \underline{S}^t\underline{S} L}{L^t \underline{S}^t\underline{S} L}$; note that the denominator is a constant. One may precalculate a vector $Z = (\underline{S}^t\underline{S} L) / (L^t \underline{S}^t\underline{S} L)$ then $c = T'^t Z$.

Carrying this out (i.e. adding cK to the T, $\ln p_s$ field derived from P) gives the new profile shown by the dashed line in Fig. 7 (almost distinguishable from the original profile - full line).

At most levels the difference with the original profile is about tenth of a degree, the largest difference, of about a half degree, occurs at the lowest level. There is no sign of 'noise'.

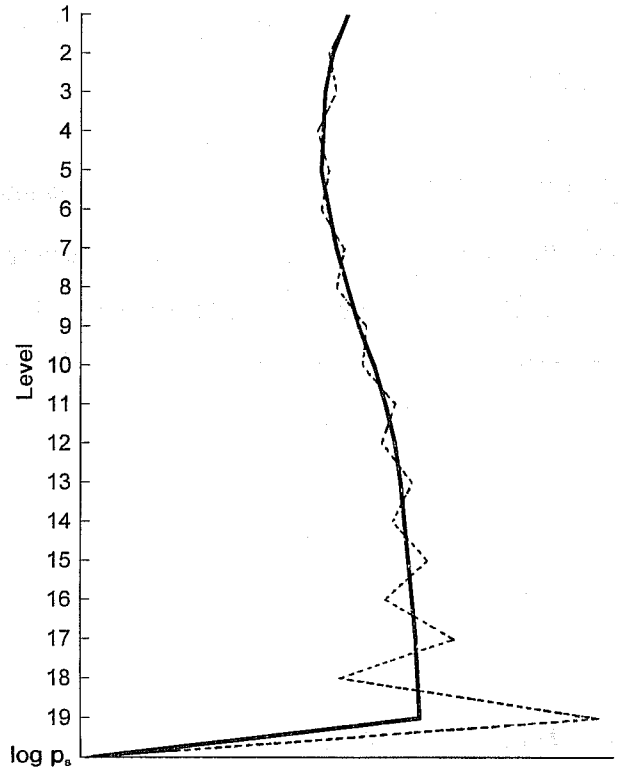

Fig.7 Solid line shows a typical temperature profile. The dotted line shows the effect of transforming *(T, lnp$_s$)* to *(P)* and back to *(T, lnp$_s$)*. It is effectively the original profile with zero amplitude for the kernel. The temperature difference at model level 19 is 159°C. The dashed line (barely distinguishable from the solid) is the temperature profile recovered from $\underline{P}$ after diagnosing the amplitude of the kernel.

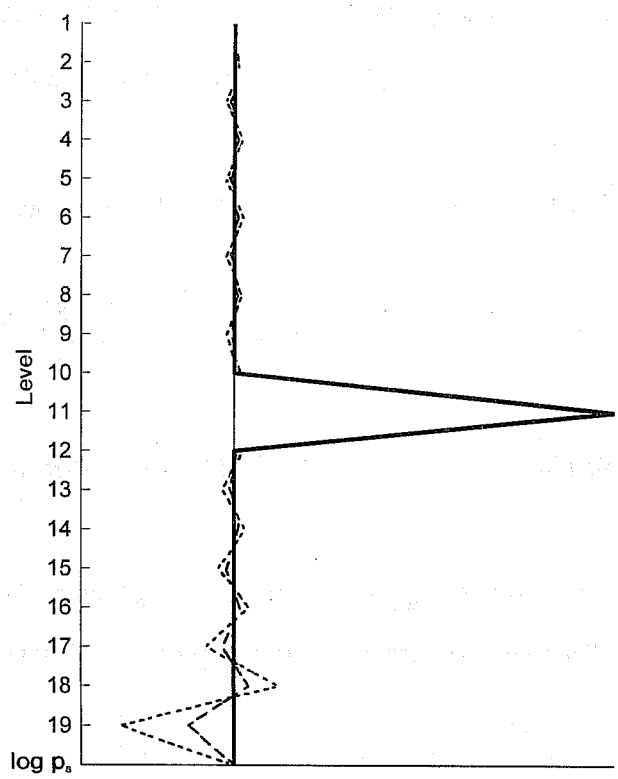

Fig.8 The solid line represents a delta function in temperature centred at model level 11. The dotted line shows the profile after transforming *(T, lnp$_s$)* to *(P)* and back to *(T, lnp$_s$)*. The dashed line is the transformed profile plus diagnosed kernel.

It is important to note that this approach only changes the amplitude of the kernel. There is no change to the corresponding $\underline{P}$ field. Information on $\underline{P}$ supplied through the minimization is passed on intact.

Effect on very sharp profiles

Fig. 8 below shows the effect of the transformation on a delta function at model level 11. In this case the field as recovered from P shows rather a lot of noise, the spurious signal at level 19 is 30% of the value at level 11.

It should be noted that the criterion for diagnosing the amplitude of the kernel (minimum second derivative of T in the vertical) is clearly inappropriate when trying to recover a delta function. Even so the technique reduces this noise by over a factor of two.

NMC approach

According to the *Parrish and Derber* (1992) paper the NMC approach is to explicitly minimize a cost function for temperature $J = \underline{T}^t \underline{S}^t \underline{S} \underline{T}$ where the matrix $\underline{S}$ is an $(L-2) \times L$ matrix with all zeros except for the three diagonals $S_{jj} = 1$, $S_{jj+1} = -2$, $S_{j,j_2} = 1$, $1 \le j \le L - 2$. Which, for regularly spaced levels, applies 2nd derivatives to the temperature in the vertical. First, the equation for J is expressed in terms of $\underline{P}$ and $\ln p_s$ using the definition of $\underline{P}$. The resulting equation is then minimized, at constant $\underline{P}$, to solve for $\ln p_s$. Finally $\underline{P}$ and this value of $\ln p_s$ is used to solve for $\underline{T}$ using the definition of $\underline{P}$. The result should be very similar to that obtained using our approach.

Other solutions

Hoskins and Simmons (1975) also choose $\ln p_s$ so as to minimize the two-grid wave in $\underline{T}$. In their case they do this by demanding that a binomial filter (1,-4,6,-4,1) on $\underline{T}$ produce zero result (they were using a five-layer model). Thus $\ln p_s$ is chosen such that a cubic polynomial can be fitted through the five values of $\underline{T}$.

A remark on NNMI

The 2 delta eta noise causing such a problem is not apparent in initialized fields. A likely explanation follows from the fact that the kernel has zero frequency and therefore is described by the Rossby component of the flow. As NNMI only affects the gravity part of the flow, the kernel can not be changed by the initialization process.

APPENDIX C: HORIZONTAL STRUCTURE FUNCTIONS

C.1 Observational evidence

Published literature concerning the shape of horizontal structure functions of short-range forecast errors is somewhat scarce. The information at our disposal consists of the two papers *Hollingsworth and Lönnberg* (1986) (hereafter HL86) and *Lönnberg and Hollingsworth* (1986) (hereafter LH86) which were the basis for the structure functions of the "new" ECMWF OI described by *Shaw et al.* (1985). *Lönnberg* (1988) (hereafter L88) describes some "revised structure functions" as used at ECMWF. From elsewhere, we have *Bartello and Mitchell* (1992) (hereafter BM92) and *Mitchell et al.* (1990) the latter has not been used since no spectra were presented in it.

These authors have been using the northern American radiosonde network (the only homogeneous network available) and thus have sampled scales from 300 km to 3000 km which corresponds to total wave numbers 6 to 66. Comparing Fig. 2 of HL86 with Fig. 11 of BM92, one sees that the spectra agree pretty well in the range 15 to 60 with, in particular, a maximum around wave number 20. The maximum at wave number 9 in HL86 is suspect and this can be explained by an accumulation of larger scale energy which was not properly sampled.

In terms of slope of the wind spectrum, BM92 are very careful, saying that it is negative in the range 3-6. However, for geopotential they claim a range of minus 3-4. These figures are contradictory since, under the geostrophic assumption, a slope $-p$ for wind leads to a slope $-(p+2)$ for geopotential. LH86 came to the conclusion of a negative slope for wind of between ½ and 1 which is again contradictory with Fig. 2 of HL86 where the wind slope is in the range minus 3 to 4 and closer to 3 than to 4. The explanation of this apparent paradox is in LH86 Fig. 8 where we can see that the end of the spectrum is noisy to the point that it is difficult to infer any sensible geopotential slope in the inertial range. This has been recognised by L88 since the revised structure functions have been obtained using a slope 4 for geopotential.

Fig. 2 of HL86 is believable for the inertial range since:

- it is stable with respect to number of Bessel functions retained
- it is confirmed by Fig. 14 of HL86
- it is confirmed for wave number 15 to 40 by Fig. 8 of LH86.

It has been chosen to use a slope of -2 in the inertial range for the wind power spectrum (which corresponds to -3 in terms of modal spectrum) since it is consistent with *Charney* (1971) theoretical analysis of 2D quasigeostrophic flow.

For the larger scales, one can hardly believe Fig. 2 of HL86 (wind pairs lead to less information on the large scales than height pairs). Furthermore, it is not in agreement with either Fig. 8 of LH86 or Fig. 11 of BM92. The fact that the slope of height is negative in Fig. 7 of LH86 or Fig. 5 of BM92 and that it is positive in wind (Fig. 8 of LH86 or Fig. 11 of BM92) shows that the wind slope is between 0 and 2. At 0 the wind spectrum is flat and at 2 it is the height spectrum which would be flat. Using the 2 points available in LH86 lead to a negative slope 1 for height. In BM92 the 2 points would lead to a negative slope 0.3 for height. For wind this transforms to a positive slope* in the range 1 to 1.7 (assuming geostrophic balance).

C.2 A parametric formula for the spectrum

Consider the following expression for the wind power spectrum

$$f(n) = \frac{\varepsilon + \left(\frac{n}{n_1}\right)^{p_1}}{1 + \left(\frac{n}{n_o}\right)^{p_o+p_1}}$$

with $p_o = 2$

$p_1 = 3$

$n_o = 15$

$n_1 = 2$

$\varepsilon = 0.1$

For large n, $f(n) \sim n^{-p_o}$ which justifies the choice $p_o = 2$.

For small n (and ε) $f(n) \sim n^{p_1}$. Thus p_o gives the (negative) slope in the inertial range and p_1 the (positive) slope for the large scales.

The maximum of $f(n)$ is for $n \geq n_o$ but close to n_o. For given slopes p_o and p_1 the correlation length scale is quite sensitive to the choice of n_o, for the slopes chosen$n_o = 15$ gives a length scale of about 500 km for geopotential.

- ε has been made negligible $\varepsilon = 0.1$, it controls the shape (and not slope) of the spectrum for very large scales $n = 0$ to n_1. Once ε is negligible, n_1 is just a scaling factor, it plays then no role in the shape of f

- the value chosen for is based on forecast error studies using satellite radiances (*Rabier*, personal communication).

The spectrum is rescaled so that the correlation for zero separation is equal to 1. The scaling factor is easily computed as

$$\left(\sum_n f(n)\, p_n^o(o)\right)^{-1} - \left(\sum_n f(n)\times\sqrt{2n+1}\right)^{-1}$$

Fig. 9 presents the spectrum obtained and Fig. 10 presents the corresponding $<\phi,\phi>$ grid point correlation function superimposed with what is used in ECMWF OI.

No information was available on the planetary scales. ε and n_1 are the free parameters of the formulation which will have to be evaluated. Studies currently being undertaken (*Rabier*, personal communication) should shed some light in this.

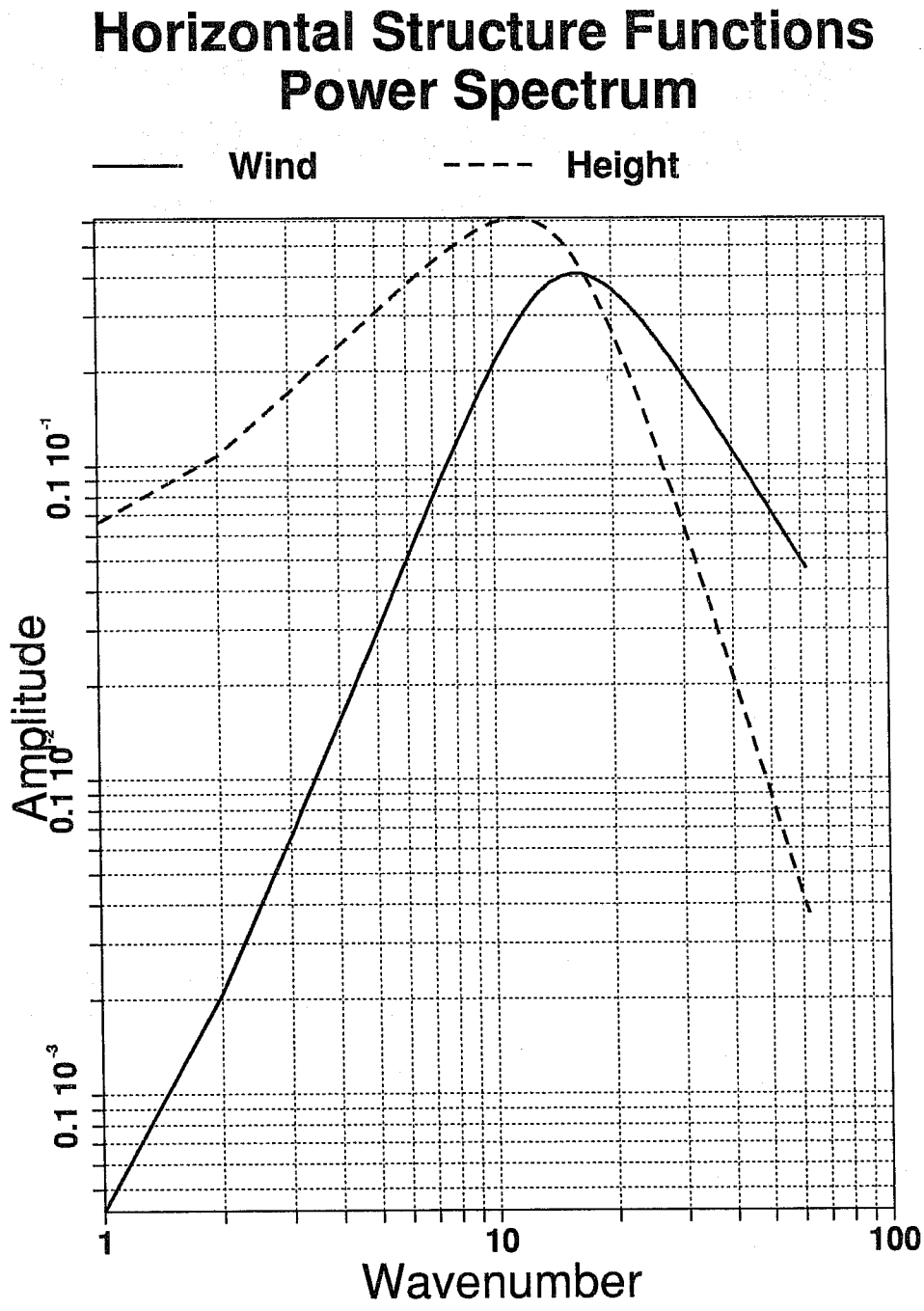

Fig.9 log/log plot of the power spectrum describing the horizontal background error structure for wind and height.

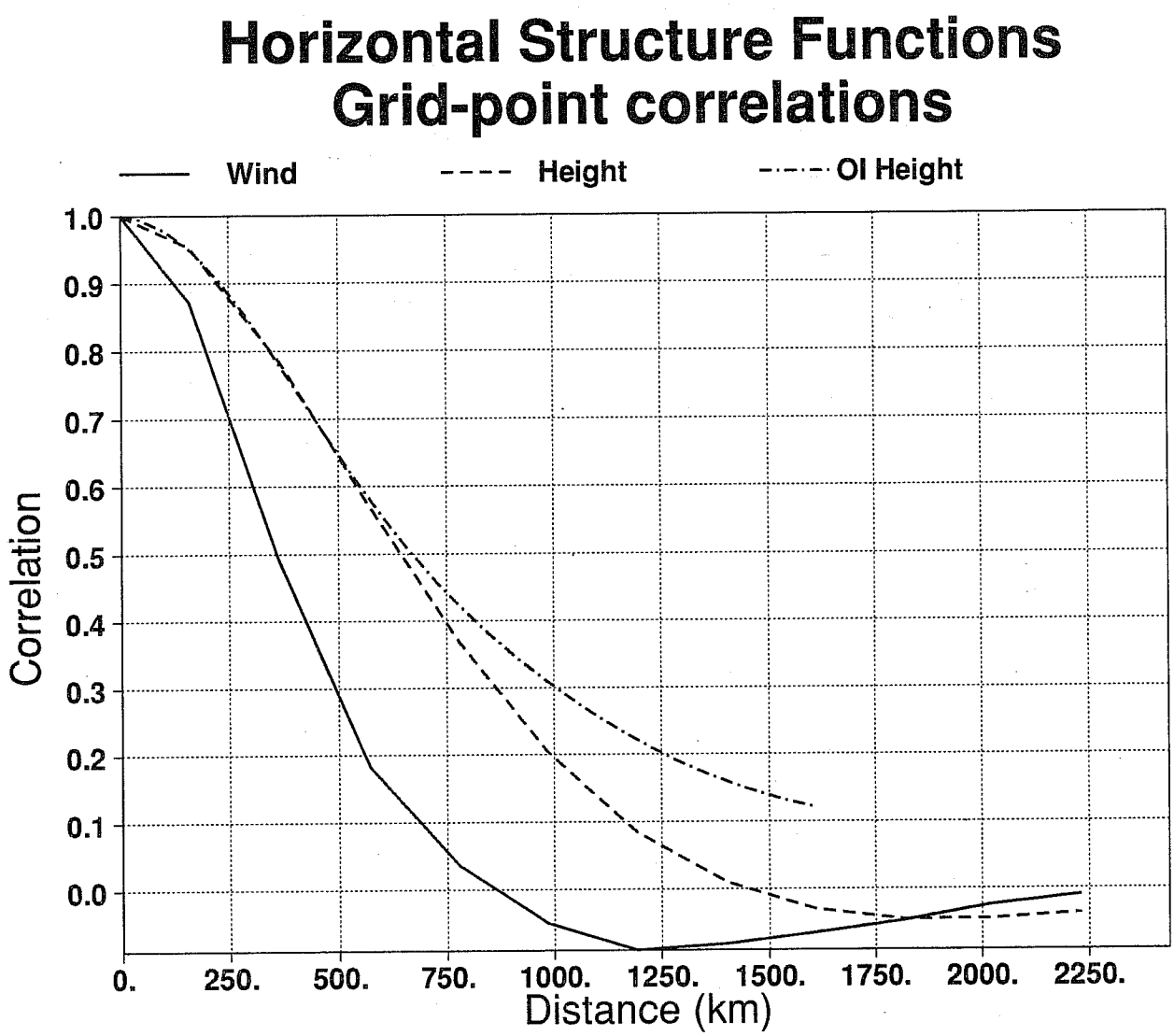

Fig.10 Grid point correlation function corresponding to the spectra shown in Fig.9. The current <Φ,Φ> correlation function currently being used by the ECMWF OI in the southern hemisphere is also shown.

$\underline{V}$ is a block diagonal matrix with block elements square matrices of dimension , one block for each of the 3-D state variables.

For consistency with the mass-wind balance, as imposed through , it is advisable that the same vertical correlations are used for both mass and wind. The covariance is not explicitly specified, but is implied through the covariance, although because of the limited accuracy to which one can construct the kernel (Appendix B) it is not uniquely determined. As it will become clear, the implied covariance is very sensitive to the $\langle P,P \rangle$ covariance, for this reason it is chosen to use the latter to specify the $\langle \xi,\xi \rangle$ and $\langle D,D \rangle$ rather than the reverse. $\langle q,q \rangle$ covariances may be independently specified - these have been modelled on the current ECMWF OI form for relative humidity correlations.

In the absence of information specific to P, a natural starting point is to generate $\langle P,P \rangle$ covariances from those of height. One may do this, for example using the form for $\langle \Phi,\Phi \rangle$ used by the operational ECMWF OI scheme. Fig. 11a shows the resultant P correlation of all levels with model level 11 about 500 hPa. One may now use the model code to explicitly compute the corresponding $\langle T,T \rangle$ covariances. Fig. 11b shows the resulting T correlation between level 11 and all levels and with p_s. The structure is very noisy. If one uses these $\langle P,P \rangle$ covariances in an assimilation with a single temperature observation at 500 hPa, one obtains temperature analysis increments as shown in Fig. 11c - again very noisy. This problem occurs because the $\langle T,T \rangle$ covariance structure implied by the $\langle P,P \rangle$ covariances is too sharp in the vertical to be properly resolved by the model resolution.

A safer approach is to take as a starting point a reasonable $\langle T,T \rangle$ covariance structure and use the model code to generate the corresponding $\langle P,P \rangle$ covariance. As an example, Fig. 12a shows the temperature correlation between model level 11 and all levels; and Fig. 12b the $\langle T,\ln p_s \rangle$ correlations. These, together, now imply a $\langle P,P \rangle$ correlation for model level 11 as shown in Fig. 12c. Comparing with Fig. 11a we see that these are somewhat broader. One can repeat the exercise of using the model code to reconstruct the implied $\langle T,T \rangle$ correlations for model level 11. The result is shown in Fig. 12d. Comparing with 12a one sees that the original correlation structure has been maintained through the transforms. Repeating the single observation analysis experiment using the new $\langle P,P \rangle$ covariances one obtains the increments shown in Fig. 12e. This is more like what one would hope to see - it has none of the noise evident in the earlier case (c.f. Fig. 11c), however it is not identical to the $\langle T,T \rangle$ correlations shown in 12a, and this requires further investigation.

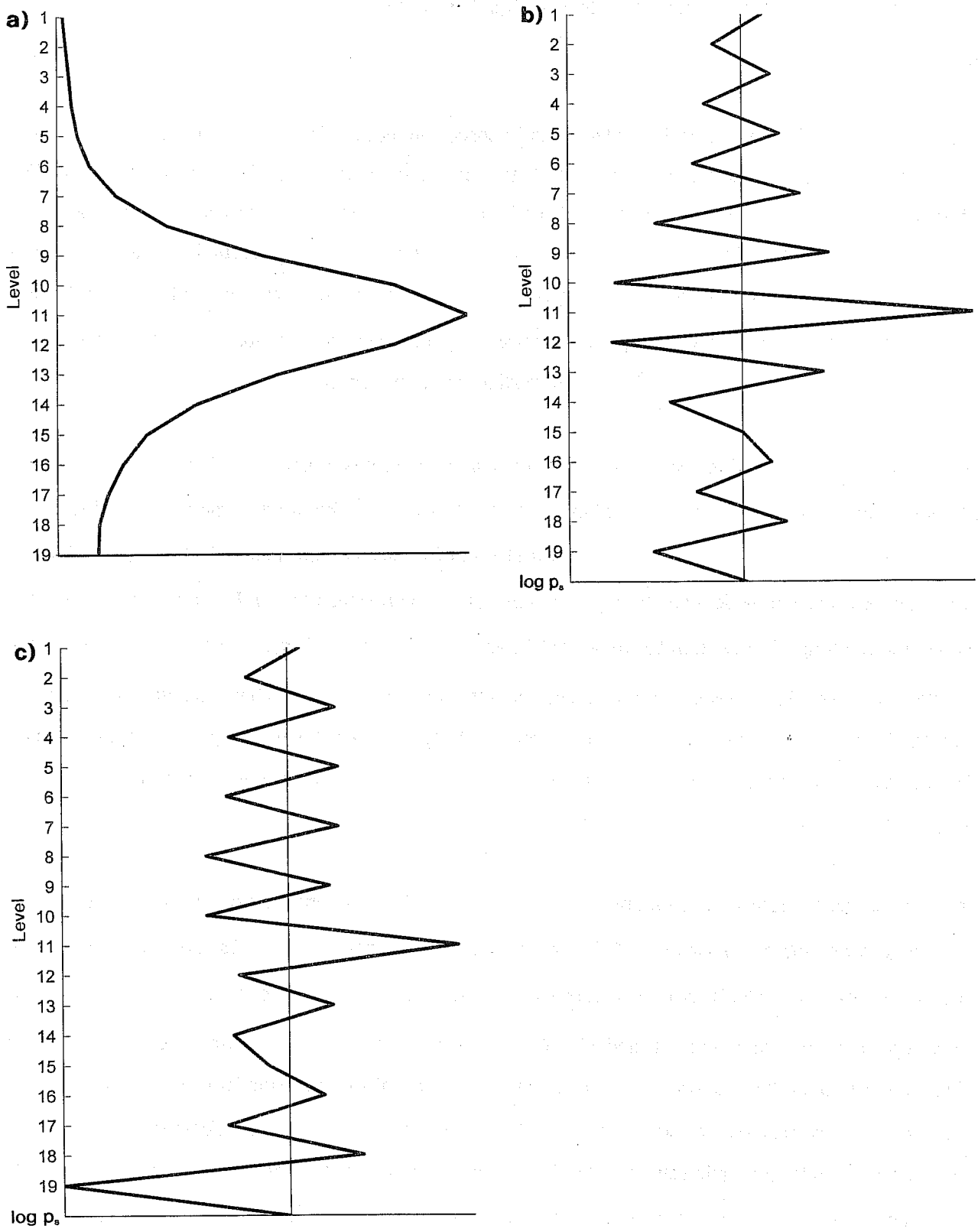

Fig.11 a) <*P,P*> correlations of all model levels with level 11 as calculated from ECMWF OI height covariance.
b) <*T,T*> correlations of all model levels with level 11 as implied by the <*P,P*> correlations shown in Fig.11a.
c) 3D-Var analysis increments resulting from a single temperature observation at 500 hPa (approx model level 11).

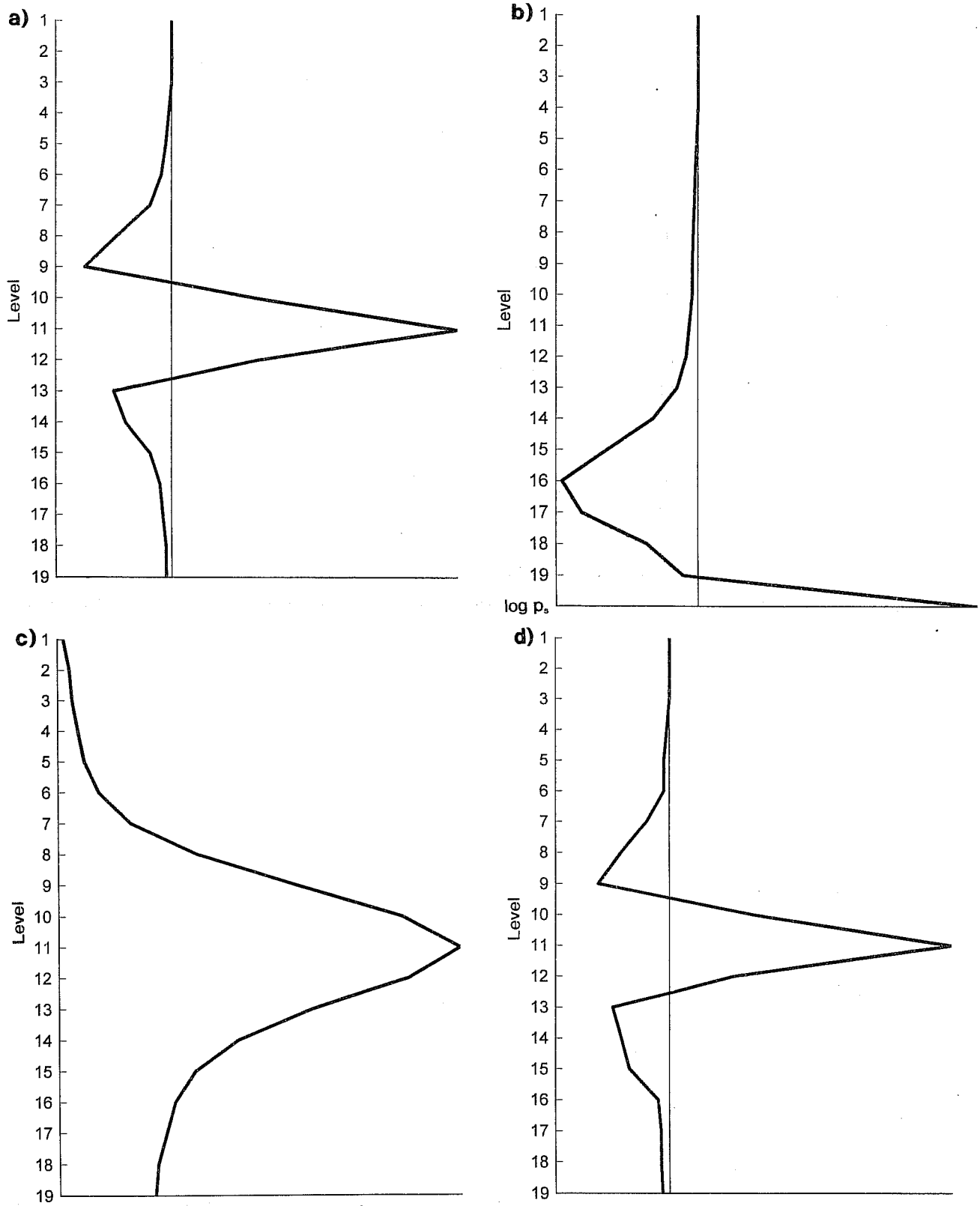

Fig.12 a) <*T,T*> correlations of all levels with model level 11 as used by OI,
b) Form of <*T*,*lnp*$_s$> correlation used,
c) <*P,P*> correlations of all model levels with level 11 as derived from 12a), b).
d) <*T,T*> correlations of all model levels with level 11 as implied by the <*P,P*> correlations shown in Fig.12c.

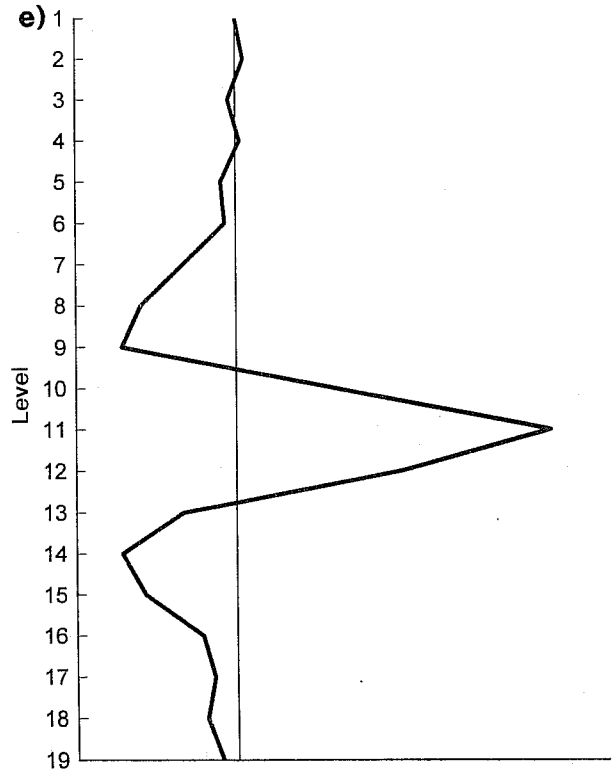

Fig. 12 e) 3D-Var analysis increment resulting from a single temperature observation at 500 hPa (approx model level 11).

APPENDIX E: SPECIFICATION OF THE STANDARD DEVIATIONS OF BACKGROUND ERRORS

An estimate of the rms forecast errors of u, v, T and h ($E_{u,v,h}$) is available from the operational OI analysis on a 6° x 6° lat/long grid at 7 pressure levels (1000, 500, 300, 200, 100, 50 and 10 hPa).

i) Horizontally interpolate $E_{u,v,h}$ to the model gaussian grid.

ii) Vertically interpolate the $E_{u,v,h}$ to the model η-levels, using the background surface pressure to define the levels.

iii) Generate error fields for the variables used: contra- and co-variant components of the wind, $\ln p_s$, and q.

Contra and co-variant wind.

$$\sigma_{u*,v*} = E_{u,v} cosine(lat)$$

Surface pressure.

$$\sigma_{p_s} = \rho g E_{h_s}$$

$$\sigma_{\ln(p_s)} = \frac{\rho g E_{h_s}}{p_s}$$

Specific humidity.

$$\sigma_q = \frac{(1 - q)^2 e_s R_d}{(p - e_s) R_v} \left(\frac{\Delta r}{100} \right)$$

where currently Δr varies quadratically with p from 10% at 1000 hPa to 50% at 300 hPa passing through 30% at 500 hPa. Above 300 hPa it remains 50%, and below 1000 hPa at 10%.

Mass variable P

The current specification of the P standard errors is in terms of the OI height errors:

$$\sigma_p = g E_h$$

This, approximate, form for the P errors follows from the fact that P represents a linearized computation of geopotential height.

The approach adopted is:-

1) Specify the $<T,T>$ and $<T,\ln p_s>$ correlations, and climatological T standard errors.

2) Compute the implied $\langle P,P \rangle$ correlations using the model code, including the diagnosis of the kernel.

3) Use the diagnosed kernel in all P to T, lnp_s transforms.

4) P standard errors used are derived from the OI height errors of the first guess forecast.

Options exist for further processing of the rms errors, such as setting σ's to constant values over η-levels, and spectral smoothing of $(1/\sigma)$. The latter is necessary to avoid too much aliasing in the calculation of the cost function.

References

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

Courtier, P. and O. Talagrand, 1990: Variational assimilation of meteorological observations with the direct and adjoint shallow-water equations. Tellus, 42A, 531-549.

Daley, R., 1983: Spectral characteristics of the ECMWF objective analysis system. ECMWF Techn.Report No.40, 117pp. ECMWF, Reading.

Daley, R., 1993: Atmospheric data assimilation on the equatorial beta plane. (Submitted for publication).

Eyre, J.R., 1989: Inversion of cloudy satellite sounding radiances by nonlinear optimal estimation. 1: Theory and simulation for TOVS. Q.J.R.Meteor.Soc., 115, 1001-1026.

Gilbert, J.Ch., and C. Lemaréchal, 1989: Some numerical experiments with variable storage quasi-Newton algorithms. Mathematical Programming, B25, 407-435.

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.

Hoskins, B.J., and A.J. Simmons, 1975: A multi-layer spectral model and the semi-implicit method. Q.J.R.Meteor.Soc., 101, 637-655.

Le Dimet, F-X. and O. Talagrand, 1986: Variational algorithms for analysis and assimilation of meteorological observations: Theoretical aspects. Tellus, 38A, 97-110.

Lönnberg, P., 1988: High resolution analysis experiments at ECMWF. Preprints Eighth Conference on Numerical Weather Prediction, Baltimore, Md. Boston, Americ.Meteor.Soc., 165-171.

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.C., 1986: Analysis methods for numerical weather prediction. Q.J.R.Meteor.Soc., 112, 1177-1194.

Lorenc, A.C., 1988: Optimal non-linear objective analysis. Q.J.R.Meteor.Soc., 114, 205-240.

Mitchell, H.L., C. Charette, C. Chouinard and B. Brasnett, 1990: Revised interpolation statistics for the Canadian data assimilation procedure: their derivation and application. Mon.Wea.Rev., 118, 1591-1614.

Parrish, D., 1988: The introduction of Hough functions into optimal interpolation. Proceedings of the Eighth Conference on Numerical Weather Predictionm Baltimore, Md., Feb.22-26. Americ.Meteor.Soc., Boston, MA, 191-196.

Parrish, D.F. and J.C. Derber, 1991: The National Meteorological Centre's spectral statistical interpolation analysis system. Mon.Wea.Rev., 120, 1747-1763.

Shaw, D., P. Lönnberg, A. Hollingsworth and P. Undén, 1987: Data assimilation: The 1984/85 revisions of the ECMWF mass and wind analysis. Q.J.R.Meteor.Soc., 113, 533-566.

Tarantola, A., 1988: Inverse problem theory: methods for data fitting and model parameter estimation. Publ.Elsevier, Amsterdam.

Thépaut, J.N. and P. Courtier, 1991: Four-dimensional variational data assimilation using the adjoint of a multilevel primitive equation model. To appear in Q.J.R.Met.Soc. Also available from ECMWF as Technical Memorandum 178.

Undén, P., 1989: Tropical data assimilation and analysis of divergence. Mon.Wea.Rev., 117, 2485-2517.