A STRATEGY FOR OPERATIONAL IMPLEMENTATION OF 4D-VAR

P. Courtier, J.-N. Thépaut and A. Hollingsworth

1. INTRODUCTION

The development of variational four-dimensional assimilation (4D-Var) from a theoretical possibility to a practical reality is progressing at a rapid pace. The first results of four-dimensional variational assimilation using real observations were reported by *Thépaut et al.* (1992b) using an adiabatic primitive equation model at truncations T21 and T42. More recently *Andersson et al.* (1992) used 4D-Var with a T63 model to assimilate remotely sensed data such as infra-red and microwave TOVS radiance measurements, while *Thépaut et al.* (1992a) used the same model to assimilate normalised radar backscatter cross-section measurements from the ERS-1 scatterometer.

This paper discusses the scientific and practical problems to be solved before one can envisage an operational implementation of 4D-Var, and reviews several approaches to the problems. We begin in section 2 by discussing the cost of 4D-Var in its present formulation. We show that algorithmic improvements and increased computer power are needed for an operational implementation of 4D-Var within the next few years. We consider two approaches for reducing the cost of 4D-Var. Section 3 examines pre-conditioning as a way of speeding up the minimization. In section 4 we present an approximate and cost-effective formulation of the 4D-Var problem in terms of increments. This formulation should also help in solving some of the scientific problems. In section 5, we review the lines of research needed for further scientific enhancements of the current 4D-Var formulation. Our conclusions are summarised in section 6. Our main conclusion is that the formulation of the 4D-Var algorithm in terms of increments offers good prospects of success.

2. COMPUTER REQUIREMENTS AND THE NEED FOR ALGORITHMIC IMPROVEMENTS IN 4D-VAR

The main practical problem to be solved for an operational implementation of 4D-Var is to reduce to an affordable level the time needed to do the 4D-Var calculations at operational resolution. In current operational practice at ECMWF, the cost of 24 hours of data assimilation is equivalent to the cost of 4 days of integration of the model. With the introduction of 3D-Var, this could go up to an equivalent cost of 6 days. If 30 iterations of the minimization algorithm are necessary in 4D-Var, the CPU time of a 24 hour 4D-Var assimilation is equivalent to the CPU time of 100 days of integration of the model. To achieve this amount of computation within operational deadlines will require a significantly faster computer or a substantial algorithmic improvement, or both.

Assuming that the next generation computer will be 3 to 5 times faster than current machines, the CPU time for 100 days of model integration would be equivalent to between 20 and 33 times the CPU time taken today for a 1-day integration of the model. So, in order to be able to implement 4D-Var operationally, it is necessary to find algorithmic improvements which would cut the cost by a further factor between 3 and 5. In this argument the cost of the model is assumed to be constant. As discussed in the 1993-1996 ECMWF Four-Year Plan, it is expected that enhancements in physical parametrization will require extra prognostic variables and increased vertical resolution at the boundaries. A 50% increase in the cost of the model is likely. The factor 3 to 5 then becomes 5 to 8. Thus we need algorithmic developments to cut the cost of 4D-Var by a factor of 8 in order to implement 4D-Var on the next generation computer.

To summarise, taking as unit the CPU time of today's 24 hour forecast on the C90, and assuming a 50% increase in the model cost, and further assuming we can gain a factor of 8 on 4D-Var, the overall suite would cost: (100/8 + 10) x 1.5 = 33.75, whereas at the moment it is only: 10 + 4 = 14. This represents a cost increase by a factor of 2.4. If we were only able to reduce the cost of 4D-Var by a factor of 5, the cost of the operational suite would increase by a factor of 3.2. In other words, assuming there is an increase of computer speed in the range 2.4 to 3.2, we shall require an order of magnitude reduction in the cost of 4D-Var, arising from algorithmic improvements, before operational implementation becomes a practical possibility.

3. PRECONDITIONING OF VARIATIONAL ASSIMILATION

Variational data assimilation attempts to solve a minimization problem. The use of the adjoint technique allows an efficient computation of the gradient of the cost function (*Le Dimet et Talagrand*, 1986). However, the number of iterations of the minimization process can be large. *Courtier* (1987) and *Courtier and Talagrand* (1990) used 30 iterations in 2-dimensional problems. *Thépaut and Courtier* (1991), *Thépaut et al.* (1992), *Rabier and Courtier* (1992) and *Rabier et al.* (1992) also used about 30 iterations in a 3-dimensional problem for minimization problems of size 10^5; results obtained by doubling the number of iterations showed that convergence of the minimization was not saturated. In a similar problem to *Thépaut and Courtier* (1991), *Navon et al.* (1992) used 60 iterations. In their operational implementation of 3D-Var, *Derber and Parrish* (1992) use more than 100 iterations although most of the convergence is achieved in about 50 iterations. Even if cost is far less an issue for 3D-Var than for 4D-Var, efficient convergence of the minimization is important for 3D-Var and it is a key point for operational implementation of 4D-Var.

In an operational implementation of 4D-Var, it would probably be unnecessary to achieve a level of convergence such that the distance to the true minimum is a small fraction of the standard deviation of analysis error; a pragmatic and less stringent requirement will probably be adequate. However, the requirements on convergence criteria are more stringent in research work on 4D-Var because problems in

the scientific formulation generally show up in the later stages of the minimization. Good convergence is therefore necessary in the research phase to identify and remove weaknesses in the scientific formulation.

3.1 Earlier applications of pre-conditioning

Preconditioning has been widely used in applications of minimization. *Davidon* (1959) introduced the idea; *Navon and Legler* (1987) describe the most popular algorithms; while *Golub and O'Leary* (1989) provide a comprehensive review of the literature up to 1986. Even with the powerful algorithms now available, conditioning remains an issue for variational analysis. The ideal preconditioning for a quadratic problem is the matrix of the second derivatives of the cost function (the Hessian). The Hessian transforms an elliptic shaped cost function by an appropriate change of metric (or of variable) into a circular shaped one. The gradient then points toward the minimum, whereas it can be almost orthogonal to the direction of the minimum if the cost function is strongly elliptic.

Appropriate preconditioning speeds up the minimization in meteorological problems (*Navon and Legler*, 1987) and has been applied to the conjugate gradient algorithm since *Hestenes and Stiefel* (1952). In a satellite radiance inversion problem, *Thépaut and Moll* (1990) computed the Hessian, in a problem of dimension 30, and showed that the minimization became extremely efficient. They also showed that one got a good preconditioning by using only the diagonal of the Hessian. *Heckley et al.* (1992) show how an appropriate choice of control variable can significantly enhance the conditioning of the 3D-Var problem.

The minimization algorithms used by authors of meteorological applications usually belong to the quasi-Newton family. These methods improve the preconditioning during the course of the minimization; they are often called variable metric algorithms. We now discuss methods to estimate the Hessian in non-linear problems.

3.2 Relation between the Hessian matrix (the covariance matrix of analysis error) and the mathematical expectation of the covariance matrix of the gradient

Gauthier (1992) studied the behaviour of the covariance matrix of the gradient of a 4D-Var problem by considering the observations as random variables. He showed that this matrix is the inverse of the covariance matrix of analysis error. *Rabier and Courtier* (1992) used the same result in order to calculate error bars on the solution of their 4D-Var problem. Here we shall summarise the main results. Given a background field $\boldsymbol{x}_b$ whose error covariance is $\boldsymbol{B}$ and a set of observations $\boldsymbol{y}$ whose error covariance is $\boldsymbol{O}$ (including instrumental and representativeness errors as discussed by *Lorenc*, 1986), given the observation operator $\boldsymbol{H}$ which computes the model equivalent $\boldsymbol{Hx}$ of the observation $\boldsymbol{y}$, the variational analysis attempts to solve the minimization problem

$\wp$: minimise $\quad J(x) = \frac{1}{2}(x-x_b)^t B^{-1}(x-x_b) + \frac{1}{2}(Hx-y)^t O^{-1}(Hx-y)$ (1)

This formulation is also valid for 4D-Var if H contains a model integration from the validity time of x to the time t of the observation. In the following we assume that H is linear; the extension to the quasilinear case using the tangent linear operator H' of H is straightforward and does not bring anything of interest other than an enlargement of the validity of our discussion.

Result 1

The Hessian J'' of J at the minimum is given by:

$$J'' = B^{-1} + H^T O^{-1} H \tag{2}$$

Result 2

The analysis error covariance matrix is the inverse of the Hessian. Calling x_a the result of the minimization (the analysis) and x_t the truth

$$<(x_a-x_t)(x_a-x_t)^t> = J''^{-1} = (B^{-1}+H^t O^{-1} H)^{-1}$$

where $< >$ stands for the mathematical expectation. This result holds regardless of whether the error statistics are Gaussian or non-Gaussian. However if, in the Gaussian case, the analysis is the minimum variance estimate, then in the non Gaussian case it is the minimum variance only among the linear estimates.

Result 3

At the minimum x_a, the gradient ∇J is equal to 0. If we now introduce $\tilde{x}_b$ and $\tilde{y}$ as random variables whose expectations are respectively x_b and y and whose covariances are respectively B and O, for each realisation of $\tilde{x}_b$ and $\tilde{y}$ one can compute ∇J (at point x_a). ∇J is then a random variable and we have

$$<\nabla J \, \nabla J^t> = J'' \tag{3}$$

Thus the error covariance matrix of ∇J is equal to the Hessian.

3.3 Application to the preconditioning

One of the difficulties in using Eq. 3 in practice is that one does not know a priori the result x_a of the minimization. However given that J'' is independent of the values of the observations we can define a minimization problem $\wp_1$ for which we know the solution and which has the same Hessian as the original problem $\wp$

$\wp_1$: minimise $\quad J_1(x) = \frac{1}{2}(x-x_b) B^{-1}(x-x_b)^t + \frac{1}{2}(Hx-y_b)O^{-1}(Hx-y_b)^t$ (4)

with $y_b = Hx_b$. Clearly, J_1 and J have the same Hessian given by Eq.2. In addition, the minimum of J_1 occurs when $x = x_b$. Defining the random variables $\tilde{x}_b = N(x_b,B)$ and $\tilde{y}_b = N(y_b,O)$ as Gaussian variables with expectations respectively x_b and y_b and covariances B and O, we have

$$J'' = J_1'' = <\nabla J_1 \nabla J_1^t>$$

with $\nabla J_1 = B^{-1}N(0,B) + H^tO^{-1}N(0,O)$, since the centred variables $\delta\tilde{x}_b = \tilde{x}_b - x_b$ and $\delta\tilde{y}_b = \tilde{y}_b - y_b$ have the same covariance as $\tilde{x}_b$ and $\tilde{y}_b$. If we now consider p realisations δx_b^i and δy_b^i of the above random variables, we have

$$J'' = J_p'' = \frac{1}{p}\sum_{i=1}^{p} \nabla J_1^i \, \nabla J_1^{i\,t} \qquad (5)$$

with $\nabla J_1^i = B^{-1}\,\delta x_b^i + H^t\,O^{-1}\,\delta y_b^i$

J_p'' converges toward J'' when p becomes large. It follows a Wishart law which is the generalisation to the multi-dimensional case of the χ^2 law. In practice, p will have to remain small. One immediately sees that the rank of J_p'' is at most p, which implies that we cannot directly use J_p'' as a preconditioning. Nevertheless we can extract useful information from this matrix.

As in *Thépaut and Moll* (1990) we shall, in the following, concentrate on the diagonal of the Hessian in order to improve the relative scaling of the different variables. Indeed, *Forsythe and Strauss* (1955) have shown that using the diagonal of the Hessian is optimal among all diagonal preconditioning. In the experiment presented below, we use $p = 60$ in order to evaluate the potential of the approach.

3.4 Numerical results

3.4.1 *Effectiveness of the diagonal preconditioning!*

As in *Thépaut and Courtier* (1991) we use a 19 level T21 primitive equation spectral model. The scientific difference between the version used here (cycle 9 of the IFS: the Integrated Forecasting System developed in collaboration with Météo France, where it is called ARPEGE) and the version they used (cycle 3 of the IFS) is in the evaluation of the pressure gradient term. Here a (u,v) formulation of the primitive equations is used instead of the vorticity-divergence formulation. In a shallow-water model both formulations are equivalent but in a primitive equation model the aliasing of the cubic (and higher order) terms is different. The differences in the forecast remain, however, meteorologically small. Extra attention had to be paid to

the validation of the adjoint since the curl of the pressure gradient term is no longer equal to 0 and this led to roundoff errors in the test of the gradient. The technical details are discussed in *Courtier* (1991).

All the experiments are designed along the same lines. A reference forecast from $t_o = 0$ to $t = 24\ h$ is performed. The results of the integration at $t = 24\ h$ are used as "observations" and one tries to recover the state at time t_o. The initial condition of the reference forecast is the operational ECMWF initialised analysis valid for 30 December 1991, 1200 UTC truncated at total wave number 21 and the initial point of the minimization is a 6 hour forecast valid for the same time. Again as in *Thépaut and Courtier* (1991), the cost function is defined as the energy of the difference between the actual trajectory at time $t = 24\ h$ and the reference. In this paper all the experiments have been performed using the M1QN3 minimization algorithm described in *Gilbert and Lemaréchal* (1989). In the experiments using pre-conditioning, Eq 5 is used with an ensemble of 60 randomly chosen state vectors to estimate the diagonal of the Hessian.

Fig. 1 presents the variation of the cost function during the course of the minimization with (dot) and without (solid) preconditioning. No horizontal diffusion is used. The preconditioning clearly has a positive impact: the same decrease of the cost function is obtained in 24 iterations with the preconditioning as in 30 iterations without the preconditioning.

The square norm of the gradient (Fig. 2) decreased by one order of magnitude more with the preconditioning. However the metric for which the norm of the gradient is defined is not the same in the two cases since the preconditioning changes this metric. This is the reason why only relative variations are considered here and not the absolute values.

If the Hessian were diagonal, one should expect the present approach to be even more successful. In order to demonstrate this we use a model with only horizontal diffusion, suppressing all the primitive equation dynamics. The e-folding time of the smallest wave resolved is 4 hours in the troposphere and decreases in the stratosphere roughly as the square root of the density of a standard atmosphere. The resolvent of this dynamic is diagonal in spectral space and so the Hessian is also diagonal. Fig. 3 presents the decrease of the cost function with (dot) and without (solid) preconditioning. The impact of the preconditioning is to lead to a superlinear convergence even in the early stages of the minimization.

The difference between these two sets of results indicates that the Hessian is far from diagonal in the case of the inversion of a full model. This is not surprising since otherwise the error growth of the model would be homogeneous, and it is known from the use of Kalman filter applied to barotropic models that this is certainly not the case, as the spatial variation of estimation error is substantial (*Gauthier et al.*, 1992).

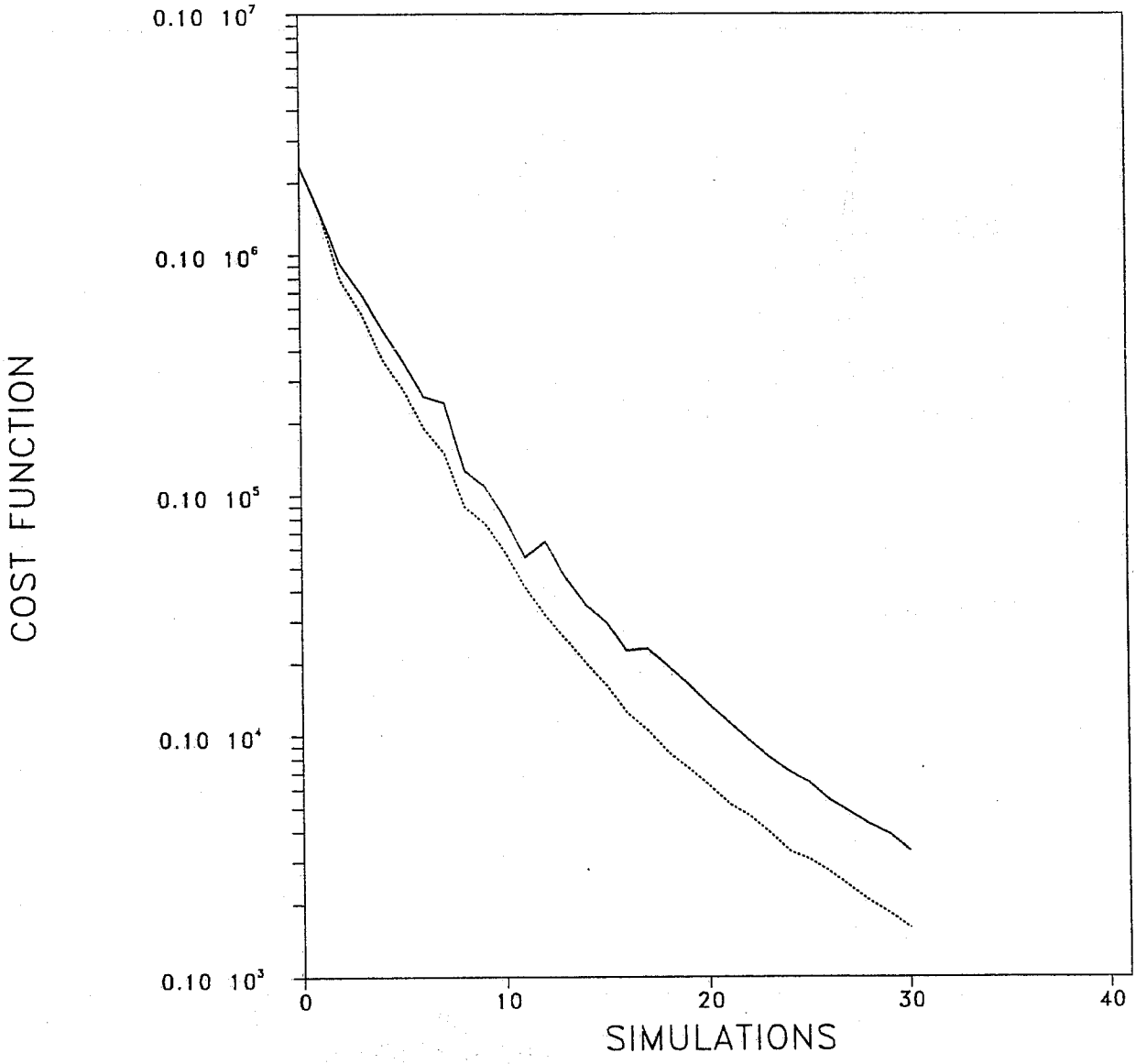

Fig. 1 Variation of the cost function during the course of the minimization without (solid) and with (dot) preconditioning. The minimization problem consists of a 24 h inversion of a T21L19 adiabatic PE model.

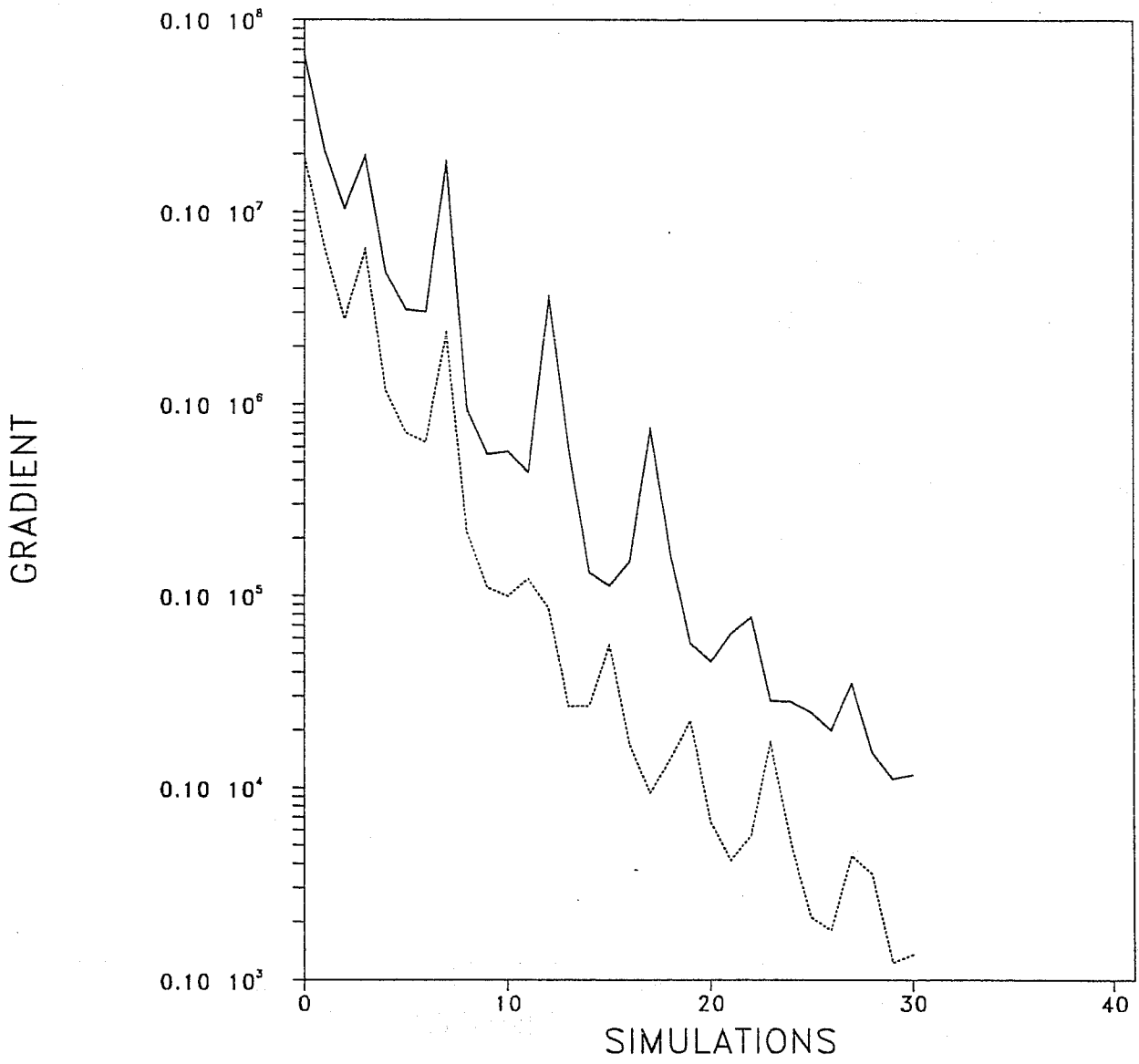

Fig. 2 Same as Fig. 1 but for the variation of the square norm of the gradient.

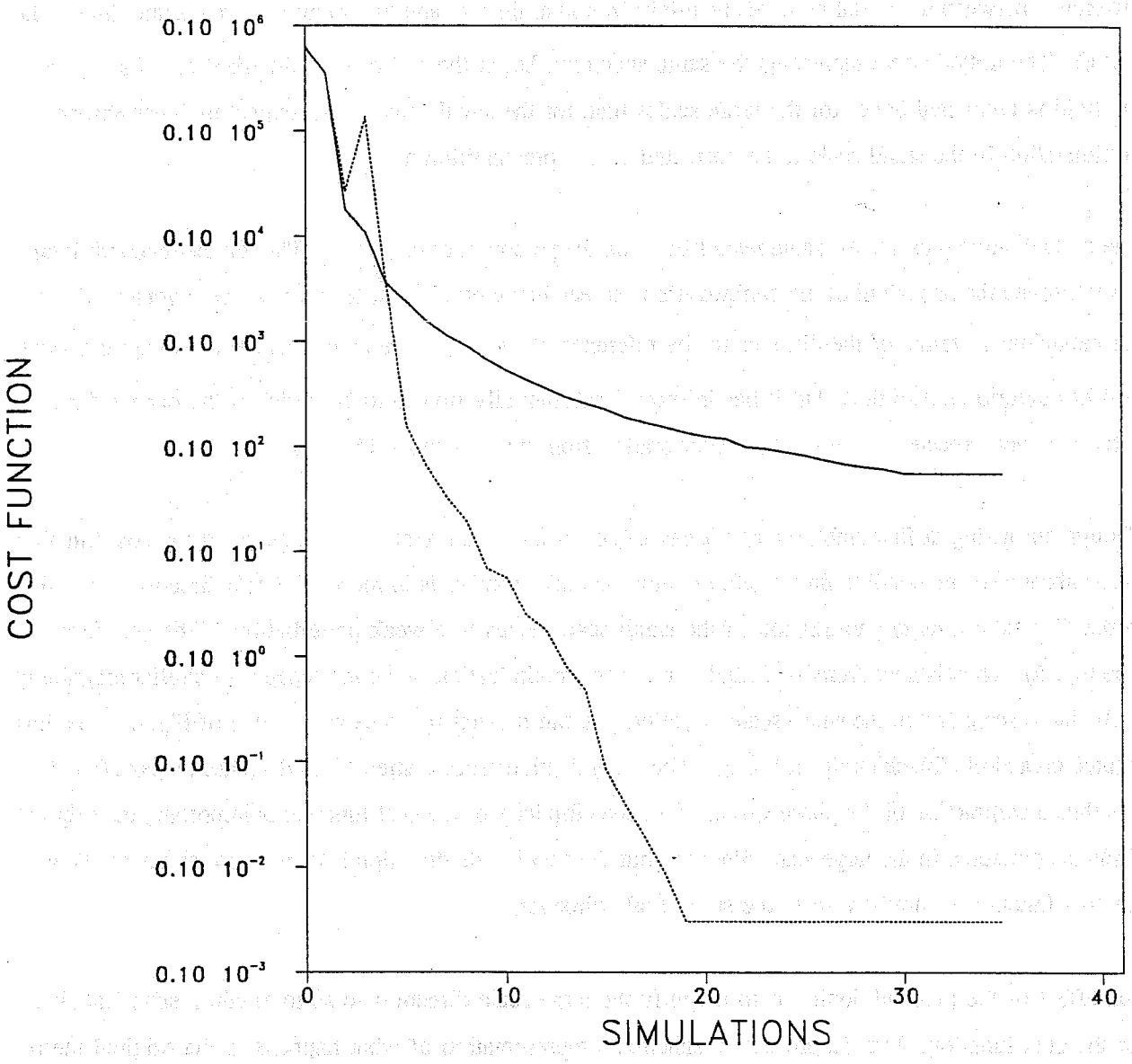

Fig. 3 Same as Fig. 1 but the model consists of only horizontal diffusion.

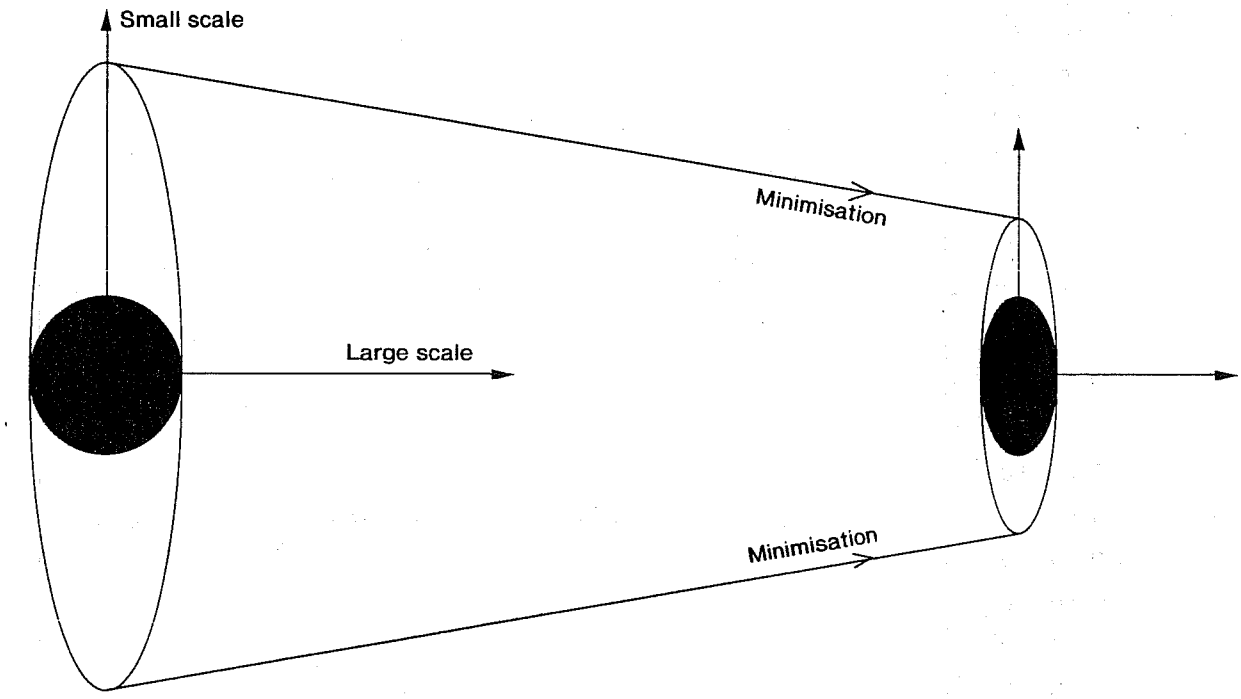

Fig. 5 Schematic representation of the effect of the minimization in phase space.

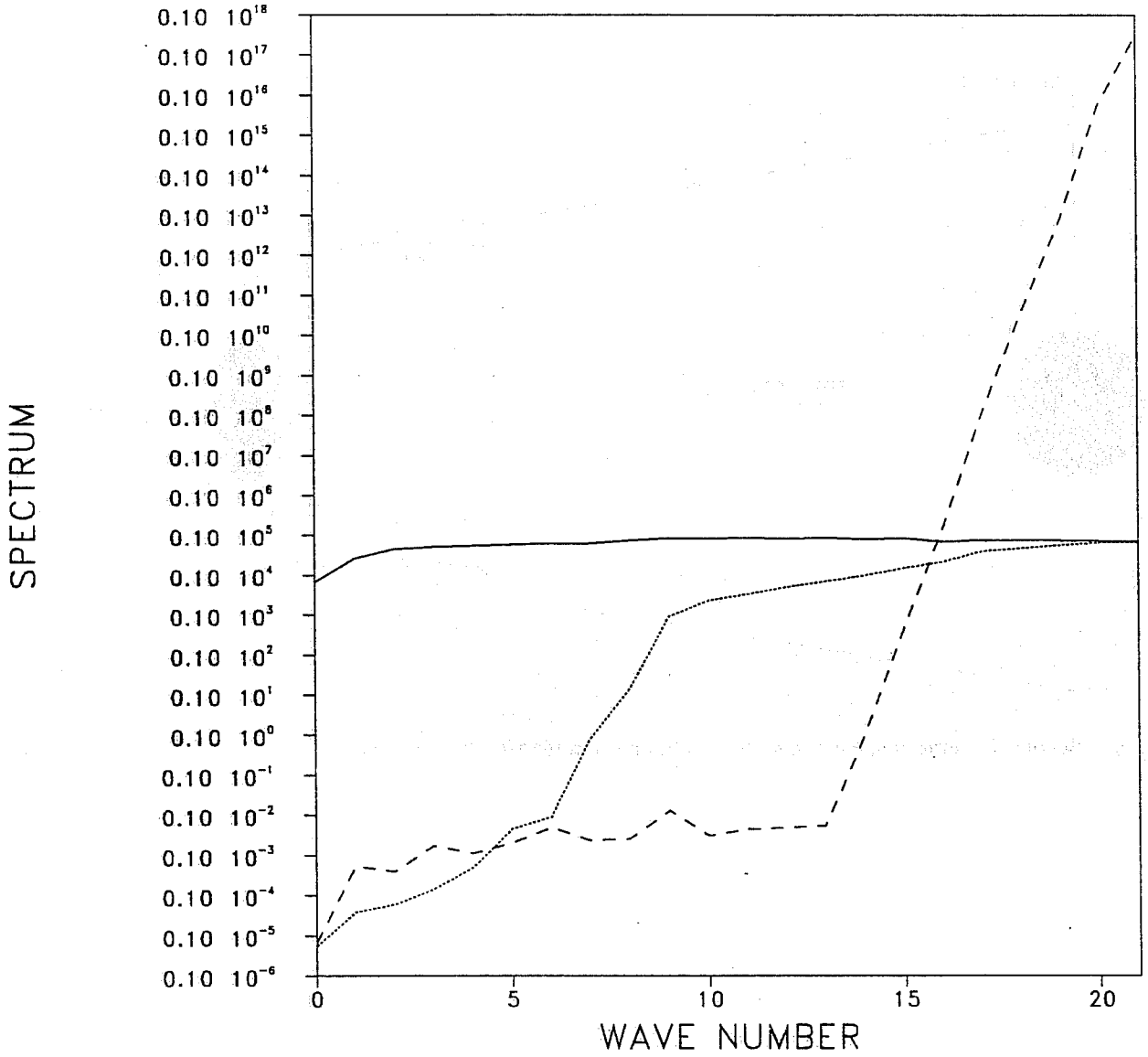

Fig. 4 Spectrum with respect to the total wave number of the energy of the difference between the reference at time t_o and
solid: the initial point of the minimisation
dotted: the result of the minimization with no preconditioning
dashed: the result of the minimization with no preconditioning.

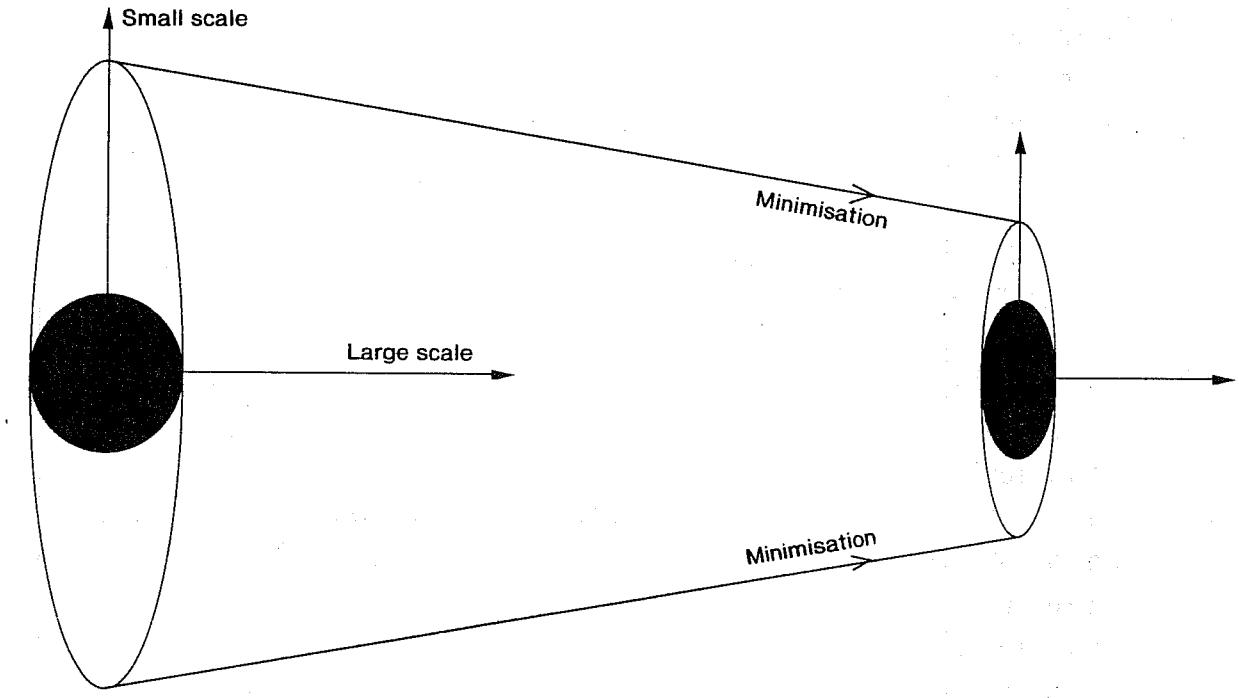

Fig. 5 Schematic representation of the effect of the minimization in phase space.

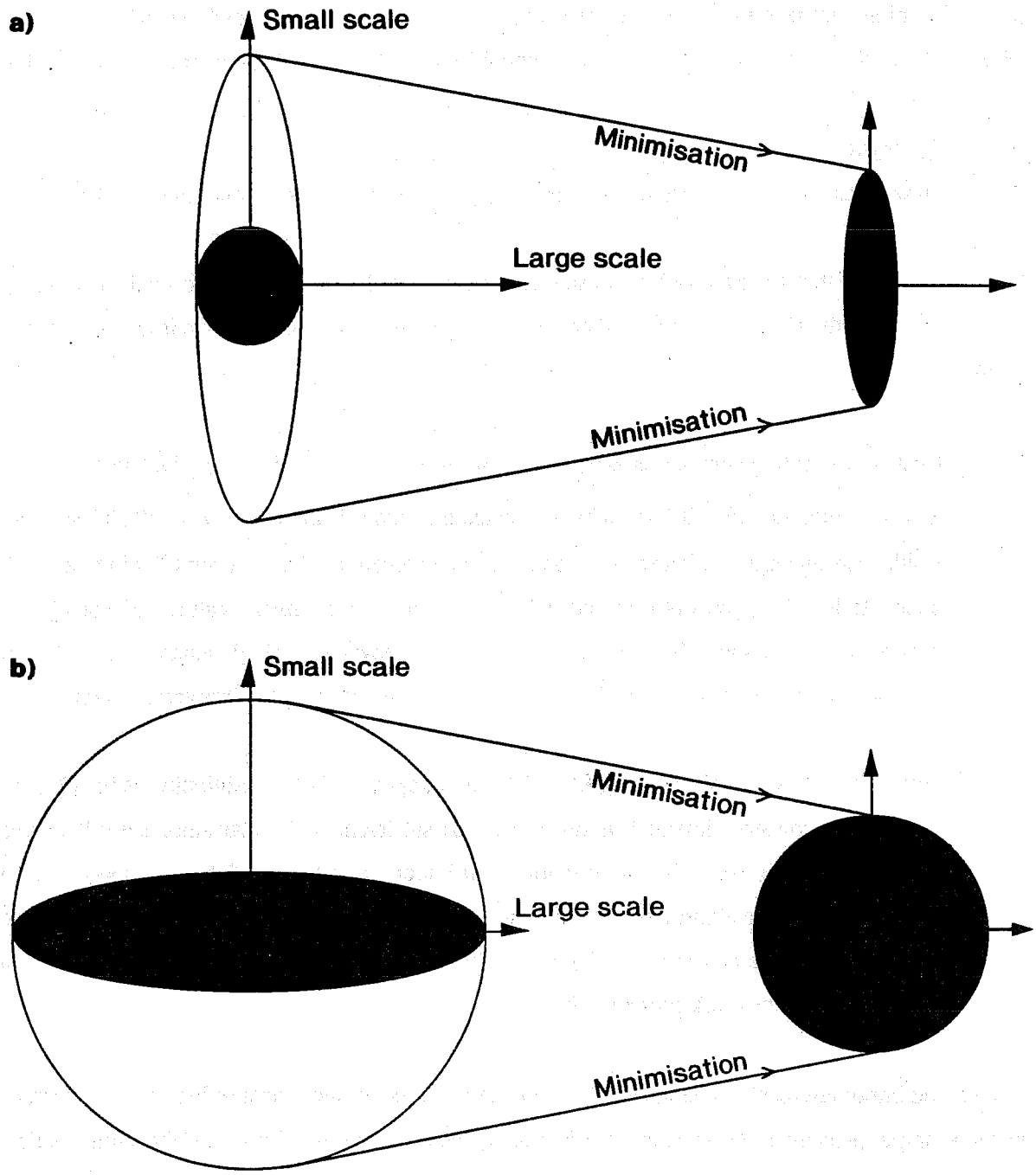

Fig. 6 Same as Fig. 5 but with the preconditioning
panel a: in the original metric
panel b: in the metric defined by the preconditioning

Similar to Fig. 4, Fig. 7 presents the energy spectrum of the difference of the reference to the initial point of the minimization (solid), the result of the minimization with no preconditioning (dotted) and with preconditioning (dashed) in the case of the inversion of a T21L19 PE adiabatic model (no horizontal diffusion). In the absence of diffusion, the preconditioning leads to an improvement for all scales.

3.5 Discussion

Two important results have emerged from our study of the effect of preconditioning on the minimization.

The first result is that the approach proposed does indeed work: it is possible to evaluate approximately some elements of the Hessian and to use them as an efficient preconditioning. Two issues, however, remain open:

i) Cost: In the experiment we have presented, we used $p = 60$ which is equivalent to the cost of 60 gradient computations. It is possible to reduce this number and using $p = 30$ did not change the results significantly (not shown). Moreover, the evaluation of the preconditioning can perhaps be done off line, independently of the value of the data. The main features of the observational network are fairly stable from day to day, so one can envisage a single adaptive algorithm in order to build up a preconditioning valid for today based on yesterdays' observation distribution.

ii) Sufficiency: A spectrally based (and therefore homogeneous) preconditioning is insufficient in the model inversion experiment; it is unlikely to be sufficient in 4D-Var since a similar dynamic is present. Furthermore, the observation distribution is rather variable in space. A natural enhancement of the algorithm is to use information on the diagonal of the Hessian expressed both in grid point space and in spectral space. This is not too difficult to implement and is a natural continuation of the work presented here.

The second important result is the potential for a negative effect of the preconditioning (on the small scales in the example presented). The negative effect arises because the initial point of the minimization is not random (according to the analysis error covariance matrix) but is so close to the minimum in all directions that the analysis minus first-guess difference can be far smaller than the variance of analysis error in some specific phase space directions. This is likely to happen in practice in 3D-Var or 4D-Var. In data sparse areas the main source of information is the background field, which is also the initial point of the minimization (i.e. the first-guess) in an area where the variance of analysis error is maximum. In the extreme case of no data in an area, we are already at the minimum in this area and one would not like to modify the fields during the course of the minimization. Special attention will have to be paid to finding a preconditioning which preserves the correct features of the initial point of the minimization throughout

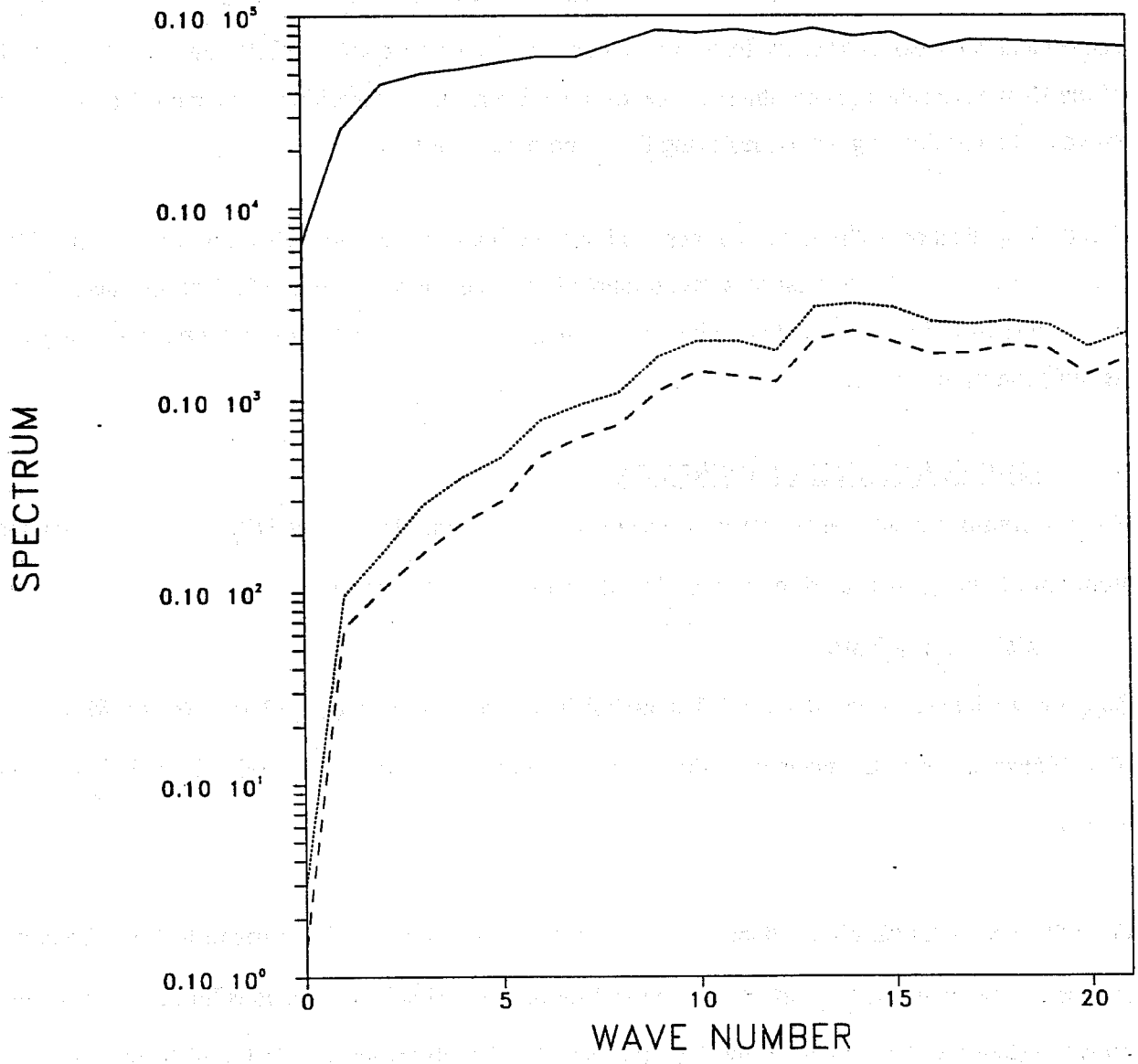

Fig. 7 Same as Fig. 4 but for the 24 h inversion of a T21L19 adiabatic PE model.

the course of the minimization. In other words, we seek "monotonic" behaviour of the minimization, monotonic from the initial point of the minimization towards the minimum. *Lorenc* (1988) pointed out the importance of this point for practical implementation of variational algorithms.

In the current ECMWF implementation of 3D-Var, the preconditioning relies on the matrix $\boldsymbol{B}^{-1}$ only. In that context, the Hessian takes the form $\boldsymbol{I} + \boldsymbol{Q}$ where $\boldsymbol{Q}$ is a positive semi-definite matrix. All eigenvalues of the Hessian are then greater than 1. Use of this property in an algorithm to bound the spectrum may improve the conditioning while preserving the monotonocity of the convergence.

A further application of the algorithm presented is an evaluation of the analysis error variance, the Hessian being the inverse of the covariance matrix of analysis error as shown in section 2. The variances are often used in operational assimilation to provide a spatial dependency of background errors used in the subsequent assimilation cycle.

4. 4D-VAR IN TERMS OF INCREMENTS

We now consider an alternative way to reduce the cost of 4D-Var. Denote by $\boldsymbol{M}(t_2,t_1)$ the model integrated from time t_1 to t_2. It is used to carry in time the state of the atmosphere $\boldsymbol{x}$:

$$\boldsymbol{x}(t_2) = \boldsymbol{M}(t_2,t_1)\, \boldsymbol{x}(t_1) \tag{6}$$

Suppose we intend to perform a 4D-Var assimilation over a period $(t_o, t_o+T=t_N)$, (to fix ideas we let $T = 24hours$). N is the number of time steps necessary to integrate the model from time t_o to time $t_o + T$.

Over this time interval, observations y_i are available at each time t_i. We assume that all observations available between times $t_{(i-½)}$ and $t_{(i+½)}$ are valid for time t_i. This is not a serious limitation since we are already below the time scales resolved by the model. The observation y_i is linked to the model state variable $\boldsymbol{x}(t_i)$ by the observation operator $\boldsymbol{H}_i$

$$y_i = \boldsymbol{H}_i\, \boldsymbol{x}(t_i) + \varepsilon_i \tag{7}$$

Eq. (7) defines the observation error ε_i of covariance matrix $\boldsymbol{O}_i$, which consists of the sum of the measurement errors and the representativeness errors (*Lorenc*, 1986).

4D-Var then consists of the minimization problem

$$\wp_{4D}\text{: minimise } J(x(t_o)) = \tfrac{1}{2}(x(t_o)-x_b)^t\, B^{-1}(x(t_o)-x_b) + \tfrac{1}{2}\sum_{i=0}^{N} (H_i x(t_i)-y_i)^t\, O_i^{-1}(H_i x(t_i)-y_i) \qquad (8)$$

with $x(t_i) = M(t_i,t_o)x(t_o)$

x_b is the background information valid for time t_o which summarises all the information used before time t_o, and B is the error covariance matrix of x_b's.

A classical result, assuming a perfect model and linearity of H and M, is that if $x^*(t_o)$ is the result of $\wp_{4D}$, then $x^*(t_N) = M(t_N,t_o)\, x^*(t_o)$ can also be obtained applying the Kalman filter to the same statistical estimation problem (*Jazwinski*, 1970; *Lorenc*, 1986; see *Thépaut and Courtier*, 1991 or *Rabier et al.*, 1992 for a detailed presentation). In meteorological applications, however, H and M are weakly nonlinear. Assuming that the tangent linear operators $\mathcal{R}$ and H' of respectively M and H satisfy, to acceptable accuracy, the relations

$$\begin{aligned} M(t_i,t_o)\,(x(t_o)+\delta x(t_o)) &= M(t_i,t_o)\, x(t_o) + \mathcal{R}(t_i,t_o)\delta x(t_o) \\ H_i(x(t_i) + \delta x(t_i)) &= H_i x(t_i) + H'\cdot\delta x(t_i) \end{aligned} \qquad (9)$$

for perturbation $\delta x(t_o)$, then the 4D-Var problem $\wp_{4D}$ is equivalent to the so-called extended Kalman filter (see previous references). This consists of two steps (^{f} and ^{a} denote respectively forecast and analysis)

<u>the forecast step</u>

$$x^f(t_{i+1}) = M(t_{i+1},t_i)x^a(t_i) \qquad (10a)$$

$$B^f(t_{i+1}) = \mathcal{R}(t_{i+1},t_i)\, B^a(t_i)\, \mathcal{R}^t(t_{i+1},t_i) \qquad (10b)$$

where the state vector is advanced by the full model (Eq. 10a) and the forecast error covariance is advanced by the tangent linear model (Eq. 10b).

<u>the analysis step</u>

$$x^a(t_i) = x^f(t_i) + K_i(y_i - H_i x(t_i)) \qquad (11a)$$

$$B^a(t_i) = (I-K_i H'_i)\, B^f(t_i) \qquad (11b)$$

where $K_i = B^f(t_i)H_i^t[H_i B^f(t_i)\, H'^t_i + O_i]^{-1}$ (11c)

is given by the minimum variance optimality condition.

The initial conditions $x^f(t_o)$ and $P^f(t_o)$ of the Kalman filter are

$$x^f(t_o) = x_b$$
$$B^f(t_o) = B$$

In current operational practice Eq. (10a) is treated exactly and Eqs. 11a and 11c are solved to a good accuracy; however Eq. (10b) is very crudely approximated with the so-called structure functions. These are the specified spatial error correlations which are kept constant in time in current practice, while the analysis variances are amplified with a very simple rule. By contrast, 4D-Var uses implicit flow-dependent structure functions in Eq. (10b) (*Thepaut et al.* 1992), so that 4D-Var is a scientific improvement on the current operational implementation. Moreover, 4D-Var is also an algorithmic improvement on the Kalman filter (9, 10) where the equation (10b) has to be solved explicitly.

There are two main weaknesses in the 4D-Var implementation. First the model is assumed to be perfect: in Eq. (10b) no source terms Q are present. Secondly, we do not have access to the analysis error covariance $B^a(t_N)$; in section 4.5, we shall discuss some possible ways of accounting for Q. Here we suggest that if (10b) is approximate anyway, it is not scientifically worthwhile solving it exactly. In other words, it may be scientifically acceptable to replace $\mathfrak{R}$ by an approximate tangent linear model in (10b) provided this approximation is smaller than the approximation of neglecting the model error source term Q.

Let us assume from now on that $\mathfrak{R}$ is any linear operator, for which we will later stipulate the link with the model M. We define the 4D-Var problem:

$$\mathcal{P}'_{4D}: \text{minimise } J(\delta x(t_o)) = \tfrac{1}{2}\delta x(t_o)^t\, B^{-1}\, \delta x(t_o) + \tfrac{1}{2}\sum_{i=0}^{N} (H_i x(t_i)-y_i)^t\, O_i^{-1}(H_i x(t_i)-y_i) \qquad (12)$$

with $x(t_i) = M(t_i,t_o)\, x_b + \mathfrak{R}(t_i,t_o)\, \delta x(t_o)$

remark 1 If $\mathfrak{R}$ is the tangent linear model, $\mathcal{P}'_{4D}$ and $\mathcal{P}_{4D}$ are equivalent to within the accuracy of the tangent linear approximation.

remark 2 If $\mathfrak{R}$ is any linear operator which we assume describes exactly the forecast error evolution, $\mathcal{P}'_{4D}$ leads to the same result as the Kalman filter described by equations (10) and (11).

remark 3 Φ'_{4D} is better than Φ_{4D} as far as an operational implementation is concerned since we keep the original model M for propagating in time the state of the atmosphere, but use an approximate propagation in time of the errors, thus introducing some flexibility on the cost of 4D-Var.

remark 4 A variant of Φ'_{4D} is the quadratic problem

$$\Phi''_{4D}: \text{minimise } J(\delta x(t_o)) = \tfrac{1}{2}\delta x(t_o)^t B^{-1}\delta x(t_o) + \tfrac{1}{2}\sum_{i=o}^{N} (y_{b,i}+H'_i\delta x(t_i)-y_i)^t O_i^{-1} (y_{b,i} + H'_i\delta x(t_i)-y_i) \quad (13)$$

with $\delta x(t_i) = \mathfrak{R}(t_i,t_o)\delta x(t_o)$

and $y_{b,i} = H[M(t_i,t_o)x_b]$

The cost of Φ'_{4D} and Φ''_{4D} are similar but the storage requirement for the background trajectory is different: in Φ'_{4D} it is the background vertical column at the observation point and in Φ''_{4D} it is the observation equivalent of the background. In Φ'_{4D}, $\mathfrak{R}$ is an approximate linearisation of M, similarly in Φ''_{4D}, H'_i is an approximate linearisation of H.

The structure functions used in the current operational T213 optimal interpolation have a cut-off at wave number 63 (*Lönnberg*, 1988). If we were to use a T106 truncation for $\mathfrak{R}$, this would already be an enhancement in terms of resolution. An adiabatic version for $\mathfrak{R}$ with some basic simplified diabatic processes like horizontal and vertical diffusion and surface friction would produce the same benefits in terms of implicit flow dependent structure functions as obtained by *Thépaut et al.* (1992a).

The CPU cost of an adiabatic semi-Lagrangian T106 L31 model is typically 1/16 of the CPU cost of the T213 L31 version. However, as the tangent linear integration is twice as expensive as the direct integration, the effective gain is reduced by 4/3 (we do not have to apply this factor to the adjoint for which it was already taken into account), the gain is then 12 which is larger than the factor of 8 we were looking for. We shall then have scope for some further enhancements of 4D-Var to be discussed in the next section.

5. FUTURE DEVELOPMENTS

5.1 Improving $\mathfrak{R}$

There are two ways of improving $\mathfrak{R}$. Firstly one could increase the horizontal resolution: the main drawback here is the cost involved since the CPU follows a power law close to 3. In addition the trajectory

storage and thus the IO also follow a cubic law (quadratic at a given time step but the number of time steps increases linearly).

Secondly, it is necessary to take into account the physics. The experiments performed so far (*Thépaut et al.*, 1992a and *Rabier*, 1992) have used only horizontal and vertical diffusion with a simple surface friction. *Rabier et al.*, 1992 showed that large-scale condensation is essential in order to get reasonable humidity fields in the upper troposphere. More generally, it is expected that the important feedback loops present in the model M will have to be described to a reasonable accuracy with $\mathfrak{R}$. We expect the automatic methods developed at INRIA to assist us in formulating a series of tangent linear models including progressively more effects of the physics.

In terms of cost this will eventually double the CPU cost of 4D-Var (as the cost of the physical parametrizations is about 50% of the cost of the model) but it will immediately double the storage required for the trajectory (and the related IO). Currently, only t values are stored since the dynamics are nonlinear only with respect to these t values and not the $t - \Delta t$. Since, the physics are nonlinear with respect to $t - \Delta t$ values, they too will have to be stored. It should be pointed out, however, that a 2 time-level semi-Lagrangian scheme would not require this extra storage.

The physics is far more nonlinear than the dynamics. As a consequence, the tangent linear approximation is likely to be less valid for the full model than for the adiabatic version. This means that $\mathcal{P}'_{4D}$ or $\mathcal{P}''_{4D}$ are not necessarily a very good approximation of $\mathcal{P}_{4D}$. A simple way for accounting some of the nonlinearities in the final analysis is to define a sequence $\mathcal{P}''_{4D}(n)$ of assimilation:

$$\mathcal{P}''_{4D}(n)\text{: minimise } J(\delta x^n(t_o)) =$$

$$\tfrac{1}{2}\,(\delta x^n(t_o)^t + x^{n-1}-x_b)^t\, B^{-1}\,(\delta x^n(t_o) + x^{n-1}-x_b) + \tfrac{1}{2}\sum_{i=o}^{N}(y_i^{n-1}+H_i'\,\delta x^n(t_i)-y_i)^t\, O_i^{-1}(y_i^{n-1}+H_i'\,\delta x^n(t_i)-y_i) \qquad (14)$$

$$\text{with} \quad \delta x(t_i) = \mathfrak{R}(t_i,t_o)\delta x(t_o) \qquad (15)$$

$$\text{and} \quad y_i^{n-1} = H[M(t_i,t_o)\,(x^{n-1}+\delta^{*n-1}x(t_o))] \qquad (16)$$

$\delta^{*n-1}x(t_o)$ is the result of the (approximate) minimization of $\mathcal{P}''_{4D}(n-1)$ and $\delta^{*o}x(t_o) = 0$

$$\text{and} \quad x^{n-1} = x^{n-2} + \delta^{*n-1}x(t_o) \qquad \begin{matrix} x^o = x_b \\ x^{-1} = x_b \end{matrix}$$

This algorithm can be seen as a pair of nested loops. The outer loop uses the complete model in Eq. 16 to re-define the model trajectory at each iteration of the outer loop. The innter loop uses the tangent linear and adjoint of a simpler (e.g. adiabatic) model (Eq. 15) to minimize the cost function (Eq. 14) for the increments calculated with respect to the re-defined trajectory.

This approach allows a progressive inclusion of physical processes without dealing with large-scale non-differentiable minimization problems, of which little is known in practice. The drawback is that we have no guarantee that the sequence $\delta^{*n}x(t_o)$ will converge. Experimental work is necessary to address this issue but we have to be pragmatic. Highly non regular problems will remain intractable for a long time but we have here a reasonable approach that is probably robust.

Remark $\mathcal{R}$ does not have to be kept constant in this iterative process and one can imagine a sequence $\mathcal{R}^n$ where the resolution and the number of physical processes dealt with increase with n.

5.2 Quality control

Any form of quality control makes use of redundant information in order to identify erroneous data. The strength of 4D-Var compared to 3D-Var for quality control comes from the fact that we treat simultaneously 24 hours of data and not just 6 hours as is done in current operational practice. Our intention is to work along the lines developed in *Dharssi et al.* (1992), see also *Lorenc* (1988). To account for the presence of gross errors they use a cost function which reaches a plateau at a certain distance from the origin instead of increasing quadratically, as approrpiate for Gaussian error statistics (*Lorenc and Hammon*, 1988). Several issues remain open like what to do with multiple minima or how to include the possibility of gross error in the background field.

5.3 Attractor

The forced dissipative nature of the atmosphere implies the existence in phase space of an attractor to which atmospheric phase-space trajectories converge. If a numerical forecast is started from an initial state which is not on the model's attractor, there is a "spin-up" or adjustment period during which the model produces, for example, somewhat unrealistic rainfall rates. At the end of the adjustment period the model reaches its equilibrium regime, where the large-scale flow is in dynamical balance (geostrophic in mid-latitudes) and the water cycle is stabilised.

The existence of an attractor is implicit in the equations. However the time scales of the forcing and dissipation are longer than the characteristic time of the assimilation. As a result one must explicitly constrain the analyzed state to lie on the model's attractor, not only in OI or 3D-Var but also in 4D-Var.

The description of the attractor is a fundamental problem in dynamic meteorology. The model's attractor is currently approximated through the dynamical balance imposed in diabatic normal mode initialisation. Theoretical advances are needed to describe it more accurately. We shall formulate the problem of physical initialisation in a variational framework. In particular precipitation estimates will be compared with model estimates provided by a linearised version of the mass flux convection scheme. This will be a first step in coupling the humidity with the other variables in the analysis.

In rain-free tropical areas we intend to evaluate statistically the balance of radiative and adiabatic temperature tendencies. If the balance is statistically reliable, it would be used to define a weak constraint in the cost function.

5.4 Estimation of the background Error in 4D-var

The main advantage of 4D-Var compared to 3D-Var is that less weight is given to the background term since more observations are treated simultaneously. However, in data sparse areas the background still remains the main source of information and in order to do extended assimilations with 4D-var it will be necessary to provide an estimate of the background error. x_b should obviously be the result at time t_N of the previous assimilation but the statistics of error $\boldsymbol{B}$ remain to be specified. We see three possible ways of addressing this issue.

a) Use a simple algorithm such as the one in current operational practice; one would not wish to live with this solution for too long.

b) Provide an estimate of $\boldsymbol{B}$ (or better $\boldsymbol{B}^{-1}$ as it is what we need) of the 4D-Var problem. As noted in section 2, the inverse of the covariance matrix of analysis error is the Hessian of the minimization problem. One could get an estimate of the Hessian, following *Rabier et al.* (1992). However, it then has to be transported in time at the end of the assimilation period. An algorithm for this purpose remains to be found. It will also be necessary to account in a crude way for the presence of a source term $\boldsymbol{Q}$, otherwise $\boldsymbol{B}^{-1}$ would diverge towards infinite values.

c) Implement a simplified Kalman filter to provide the estimation errors at time t_N and then to use them as $\boldsymbol{B}$. This has the theoretical advantage of providing a proper estimation of $\boldsymbol{B}$ accounting both for a source term $\boldsymbol{Q}$ and for the data distribution. The difficulties are twofold:

- Due to cost, it will be possible to do this only at low resolution. Even if it can be done off-line (since it uses data of the past period of time for getting the $\boldsymbol{B}$ valid for the current

period of time) it has to be done in less than 24 hours! An efficient implementation of Kalman filter is necessary; the ideas expressed by *Cohn* (1992) for hyperbolic systems might be a solution to overcome that difficulty.

- Once we have got $\boldsymbol{B}$, it is necessary to get $\boldsymbol{B}^{-1}$ in order to compute the background terms of the variational problem. This is not a trivial problem in practice because of the size of $\boldsymbol{B}$.

Bearing in mind those two points, it is likely that a simplified Kalman filter would be used at first to propagate in time the variances and maybe to improve the vertical structure functions. It is, however, likely that one would keep a parametric description for $\boldsymbol{B}$ as currently used in 4D-var.

5.5 Accounting for Model Errors

As we have seen in section 3, the current implementation of 4D variational assimilation assumes the model to be perfect. This is a scientific limitation of the approach and one must consider ways of dealing with this problem (*Wergen* 1992).

As the Kalman filter is a particular (sequential) algorithm for solving a linear estimation problem, a variational formulation of this same problem exists. In this interpretation, the model is to be considered as part of the observation operator. One has to realise that the result of the minimization is not optimal globally, but only for a chosen time. Modifying the $\boldsymbol{O}$ matrix of problem $\mathcal{P}_{4D}$ would then lead to the proper observation term. As t_N is our time of interest in numerical weather prediction, the background term also has to be modified accordingly. How to modify those two terms is another issue, *Courtier and Talagrand* (1990) gave less weight to past information than to more recent data. This can be justified theoretically if one considers only the effect of $\boldsymbol{Q}$ on the variances, neglecting the horizontal propagation. The appropriate weight to be given to data can then be evaluated analytically if one knows the local magnitude of $\boldsymbol{Q}$ and the local growth rate of the errors (*Talagrand*, 1985, personal communication). We intend to pursue this issue as it is relatively easy to implement in the current formulation of 4D-Var. A further degree of sophistication would lead to horizontal and temporal cross-correlations of observation errors. Even if we were able to get an estimate of those, a practical implementation would be extremely difficult.

It is more important to account for $\boldsymbol{Q}$ when it is dominant compared to the right hand side of Eq. 10b. This is likely to be the case if Eq. 10b is contracting in some phase space direction. A possible solution would be to insert the background information term at several instants t : this would prevent having the effective

variance of background error becoming small enough so that observations at the final time are useless. This would not change the effective structure functions as computed by *Thépaut et al.* (1992a) when there is a large amplification of error. The theoretical justification of inserting a distributed-in-time background information is not easy, since in the case of a perfect model, we lose the theoretical equivalence with the Kalman filter. However it seems a reasonable approach in order to circumvent an obvious weakness.

Contrary to the classical implementation of Kalman filter where the forecast errors are assumed to be uncorrelated in time, one would assume them to be fully correlated in time. The forecast step becomes, instead of Eq. 10a

$$x(t_{i+1}) = M(t_{i+1},t_i)\, x(t_i) + \lambda(t_i)V$$

where $\lambda(t_i)$ is a given function of time and V (the forecast error) as $x(t_o)$ becomes part of the control variable. This is an immediate generalisation of *Derber* (1989) where only V was part of the control variable. Though part of the forecast errors are certainly systematic, this approach has two drawbacks:

i) There may not be enough data over 24 h to estimate both V and $x(t_o)$?

ii) This approach doubles the size of the control variable, which might lead to efficiency problems.

Even with these two restrictions, it might nevertheless be worth investigating this approach with for V a field truncated in the large scales (***T10***?) which would rule out reservations i) and ii).

A fourth approach is to use the dynamics as a weak rather than strong constraint. This might be the proper scientific solution; we foresee, however, severe algorithmic difficulties. First it significantly increases the size of the control variable. Secondly, weak constraints lead to ill conditioning when the constraint is known to be accurately satisfied.

6. CONCLUSION

In this paper, we have shown that major algorithmic improvements are necessary if one is to implement an operational 4D-Var on the next generation of computers.

A feasibility study of preconditioning shows that though preconditioning is the mathematical solution for our problem, it is not easy to implement in practice. Due to the large dimension of the problem, it is only possible to improve the conditioning to a limited extent. Preconditioning might also have adverse effects and one should seek monotonic behaviour of the convergence.

We have proposed 4D-Var in terms of increments as a pragmatic approach which allows us to trade cost versus benefits. Even so, 4D-Var remains expensive. Furthermore, as explained in the last section, several scientific issues are opened which will require a substantial experimental programme. Nevertheless this approach offers good prospects for success.

ACKNOWLEDGMENTS: We thank J Derber who hosted us in 1990 at NMC. It was during a discussion with him that he expressed the idea of working in terms of increments to avoid developing the adjoint of the physics. After some time, this eventually became section 4.

We are indebted to the IFS team and particularly Mats Hamrud, who implemented the IFS on the C90 very quickly.

Thanks to Carole Edis, who has been able to transform apparently random black ink on white paper into legible text.

7. REFERENCES

Andersson, E., J-N. Thépaut, J. Eyre, A.P. McNally, G. Kelly, P. Courtier and J. Pailleux, 1992: Use of radiances in 3D/4D variational assimilation. This volume.

Cohn, S.E., 1992: Dynamics of short-term univariate forecast error covariances. Submitted to Mon.Wea.Rev.

Courtier, P., 1987: Application du contrôle optimal à la prévision numérique en météorologie. Thèse de doctorat de l'université Paris VI.

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

Courtier, P., 1991: Two difficulties encountered in the validation of cycle 7. Note ARPEGE No. 24, 3 pp (available from METEO-FRANCE).

Davidon, W.C., 1959: Variable metric for minimization, Report ANL-5990, Argonne National Laboratory, Argonne, Illinois.

Derber, J.C., 1989: A variational continuous assimilation technique. Mon.Wea.Rev., 117, 2437-2446.

Derber, J.C. and D.F. Parrish, 1992: The National Meteorological Center's Spectral Statistical Interpolation Analysis System. Mon.Wea.Rev., 120.

Dhassi, I., A.C. Lorenc and N.B. Ingleby, 1992: Treatment of gross errors using a maximum probability theory. Q.J.R.Meteorol.Soc., 118, 1017-1036.

Forsythe, G.E. and E.G. Strauss, 1955: On best conditioned matrices. Proceedings of the Amer.Math.Soc., 6, 340-345.

Gauthier, P., 1992: Chaos and quadri-dimensional data assimilation: a study based on the Lorenz model. Tellus, 44A, 2-17.

Gauthier, P., P. Courtier and P. Moll, 1992: Assimilation of simulated wind lidar data with Kalman filter. To appear in Mon.Wea.Rev.

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

Golub, G.H. and D.P. O'Leary, 1989: Some history of the conjugate gradient and Lanczos algorithm: 1948-1976. Siam Review, 31, 50-102.

Heckley, W.A., P. Courtier, J. Pailleux and E. Andersson, 1992: On the use of background information in the variational analysis at ECMWF. This volume.

Hestenes, M.R. and E. Stiefel, 1952: Methods of conjugate gradients for solving linear systems. J.Res.Nat.Bur.Standards, 49, 409-436.

Jazwinski, A.H., 1970: Stochastic processes and filtering theory. Academic Press, New York, 376 pp.

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

Lönnberg, P., 1988: Developments in the ECMWF analysis system. 1988 ECMWF seminar on data assimilation and the use of satellite data, pp 75-119.

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

Lorenc, A.C., 1988: Optimal nonlinear objective analysis. Q.J.R.Meteorol.Soc., 114, 205-240.

Lorenc, A.C. and O. Hammon, 1988: Objective quality control of observations using Bayesian methods: Theory and a practical implementation. Q.J.R.Meteorol.Soc., 114, 515-543.

Navon, I.M. and D.M. Legler, 1987: Conjugate-gradients method for large-scale minimization in meteorology. Mon.Wea.Rev., 115, 1479-1502.

Navon, I.M., X. Zou, J. Derber and J. Sela, 1992: Variational data assimilation with an adiabatic version of the NMC spectral model. Mon.Wea.Rev., 120, 1433-1446.

Rabier, F., P. Courtier, J. Pailleux, O. Talagrand, J.-N. Thépaut and D. Vasiljevic, 1992: Comparison of Four-dimensional variational assimilation with simplified sequential assimilation. This volume.

Rabier, F. and P. Courtier, 1992: Four-dimensional assimilation in the presence of baroclinic instability. Q.J.R.Meteorol.Soc., 118, 649-672.

Rabier, F., P. Courtier, J. Pailleux, O. Talagrand, D. Vasiljevic, 1992: A comparison between four-dimensional variational assimilation and simplified sequential assimilation relying on three-dimensional variational analysis - submitted to Q.J.R.Meteorol.Soc.

Thépaut, J.-N. and P. Courtier, 1991: Four-dimensional variational data assimilation using the adjoint of a multilevel primitive-equation model. Q.J.R. Meteorol.Soc., 117, 1225-1254.

Thépaut, J.-N. and P. Moll, 1990: Variational inversion of simulated TOVS radiances using the adjoint technique. Q.J.R.Meteorol.Soc., 116, 1425-1448.

Thépaut, J.-N., P. Courtier and R. Hoffman, 1992a: Use of dynamical information in four-dimensional variational assimilation. This volume.

Thépaut, J.-N., D. Vasiljevic, P. Courtier and J. Pailleux, 1992b: Variational assimilation of conventional meteorological observations with a multilevel primitive equation model. To appear in Q.J.R.Meteorol.Soc.

Vasiljevic, D., C. Cardinali and P Undén, 1992: ECMWF 3D Variational data assimilation of conventional observations. This volume.

Wergen, W., 1992: The effect of model errors in variational assimilation. Tellus, 44A, 297-313.