551.509.313.22

A strategy for operational implementation of 4D-Var, using an incremental approach

By P. COURTIER*, J.-N. THÉPAUT and A. HOLLINGSWORTH
European Centre for Medium-range Weather Forecasts, UK

SUMMARY

An order of magnitude reduction in the cost of four-dimensional variational assimilation (4D-Var) is required before operational implementation is possible. Preconditioning is considered and, although it offers a significant reduction in cost, it seems that it is unlikely to provide a reduction as large as an order of magnitude. An approximation to 4D-Var, namely the incremental approach, is then considered and is shown to produce the same result at the end of the assimilation window as an extended Kalman filter in which no approximations are made in the assimilating model but in which instead a simplified evolution of the forecast error is introduced. This approach provides the flexibility for a cost–benefit trade-off of 4D-Var to be made.

1. INTRODUCTION

The development of variational four-dimensional assimilation (4D-Var) from the stage of being a theoretical possibility to being a practical reality is progressing at a rapid pace. The first results of four-dimensional variational assimilation using real observations were provided by Thépaut *et al.* (1993b) using an adiabatic primitive-equation model at truncations T21 and T42. More recently Andersson *et al.* (1994) used 4D-Var with a T63 model to assimilate remotely-sensed data such as infrared and microwave TOVS radiance measurements, while Thépaut *et al.* (1993a) used 4D-Var with the same model to assimilate normalized radar backscatter cross-section measurements from the ERS-1 scatterometer.

The main practical problem to be solved for an operational implementation of 4D-Var is to be able to reduce to an affordable level the time needed to do the 4D-Var calculations. In current operational practice at the European Centre for Medium-range Weather Forecasts (ECMWF), the cost of 24 hours of data assimilation is equivalent to the cost of four days of integration of the model. With the introduction of 3D-Var, this could increase to the equivalent of the cost of six days. If 30 iterations of the minimization algorithm are necessary in 4D-Var, the central processor unit (CPU) time of a 24-hour 4D-Var assimilation is equivalent to a CPU time of 100 days of integration of the model. To achieve this amount of computation within operational deadlines would require a significantly faster computer or a substantial algorithmic improvement, or both.

This paper, which is a polished version of the paper by Courtier *et al.* (1993), discusses the scientific and practical problems to be solved before one can envisage an operational implementation of 4D-Var. We consider two approaches for reducing the cost of 4D-Var. Section 2 examines preconditioning as a way to speed up the minimization. In section 3 we present an approximate and cost-effective formulation of the 4D-Var problem in terms of increments. This formulation should also help to solve some of the scientific problems. In section 4 we discuss some enhancements of the incremental 4D-Var formulation, and in section 5 present numerical results of incremental 4D-Var. Our conclusions are summarized in section 6, the main conclusion being that the formulation of the 4D-Var algorithm in terms of increments offers good prospects of success.

* Corresponding author: European Centre for Medium-range Weather Forecasts, Shinfield Park, Reading, Berkshire, RG2 9AX, UK.

2. PRECONDITIONING OF VARIATIONAL ASSIMILATION

Variational data assimilation can be regarded as an attempt to solve a minimization problem. The use of the adjoint technique allows an efficient computation of the gradient of the cost function (Le Dimet and Talagrand 1986). However, the number of iterations of the minimization process can be large when dealing with real observations. Courtier (1987) and Courtier and Talagrand (1990) used 30 iterations in two-dimensional problems. Thépaut and Courtier (1991), Thépaut *et al.* (1993b), Rabier and Courtier (1992) and Rabier *et al.* (1993) also used about 30 iterations in a three-dimensional problem for minimization problems of size 10^5; it was found that convergence of the minimization was not saturated by doubling the number of iterations. In a problem similar to that of Thépaut and Courtier (1991), Navon *et al.* (1992) used 60 iterations. In their operational implementation of 3D-Var, Parrish and Derber (1992) used more than 100 iterations, although most of the convergence was achieved in about 50 iterations. M. Zupanski (1993a) and D. Zupanski (1993) both obtained good convergence with 15 iterations, but their achievements should be qualified since they used full fields and not a real geographic distribution of observations. Even if cost is far less an issue for 3D-Var than for 4D-Var, nevertheless efficient convergence of the minimization is important for 3D-Var and it is a key point in the operational implementation of 4D-Var.

In an operational implementation of 4D-Var, it would probably not be necessary to achieve a level of convergence such that the distance to the true minimum would be a small fraction of the standard deviation of analysis error; this pragmatic relaxation providing a less stringent requirement will probably be adequate. However, the requirements as regards convergence criteria are more stringent in research work on 4D-Var because if there are problems in the scientific formulation they will generally show up in the later stages of the minimization. Good convergence in the research phase is therefore necessary to identify and remove possible weaknesses in the scientific formulation.

(a) Earlier applications of preconditioning

Preconditioning has been widely used in applications of minimization. Even with the help of the powerful algorithms now available, conditioning still remains an issue in variational analysis. Davidon (1959) introduced the idea of the variable-metric or quasi-Newtonian minimization algorithms; Navon and Legler (1987) described the most popular modern algorithms, while Golub and O'Leary (1989) have provided a comprehensive review of the literature up to 1976. The ideal preconditioning for a quadratic problem is achieved through the matrix of the second derivatives of the cost function (the Hessian). The Hessian transforms an elliptically shaped cost function by an appropriate change of metric (or of variable) into a circularly shaped one. After preconditioning by the Hessian the gradient points towards the minimum, whereas before preconditioning it can be almost orthogonal to the direction of the minimum if the cost function is strongly elliptic.

Appropriate preconditioning speeds up the minimization process in meteorological problems (Navon and Legler 1987) and, since the paper by Hestenes and Stiefel (1952), has been applied to the conjugate gradient algorithm (Gill *et al.* 1981). When studying the subject of nonlinear satellite radiance inversion, 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 it was possible to get good preconditioning by using only the diagonal of the Hessian. Heckley *et al.* (1992) showed how an appropriate choice of control variable can significantly enhance the conditioning of the

3D-Var problem. M. Zupanski (1993b) carried out a preconditioning of a 4D-Var problem using *a priori* knowledge.

The minimization algorithms used by those engaged in meteorological applications usually belong to the quasi-Newtonian family. These methods improve the pre-conditioning during the course of the minimization; they are often called 'variable-metric algorithms'. We now discuss methods of estimating the Hessian in weakly nonlinear problems. We call weakly nonlinear problems, problems in which the tangent-linear dynamics is a 'good' approximation to the nonlinear dynamics throughout the assimilation interval, and for meteorological orders of magnitude of the background errors. As a consequence, the cost function is close to quadratic.

(b)   *Relation between the Hessian matrix (the inverse of the covariance matrix of analysis error) and 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 to calculate error bars on the solution of their 4D-Var problem. Here we shall summarize the main results. Given an unbiased background field x_b (realization of a random variable of covariance B) and a set of unbiased observations y (realization of a random variable of covariance O, which includes instrumental and representativeness errors as described by Lorenc 1986), and given the observation operator H which computes the model equivalent Hx of the observation y, the variational analysis attempts to solve the minimization problem, viz.

$$\mathscr{P}: \text{minimize } \mathscr{J}(x) = \tfrac{1}{2}(x - x_b)^T B^{-1}(x - x_b) + \tfrac{1}{2}(Hx - y)^T O^{-1}(Hx - y). \tag{1}$$

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 quasi-linear case using the tangent-linear operator H' of H is straightforward and does not introduce anything of interest other than an enlargement of the validity of our discussion. It has to be pointed out that, outside the range of validity of the tangent-linear approximation, the Hessian would be more complicated and the extension of the results presented below not straightforward. However, Lacarra and Talagrand (1988), Vukicevic (1991) and Rabier and Courtier (1992) have shown that the tangent-linear approximation has a wide range of application. Nevertheless, this is a bold assumption and its range of validity will have to be reassessed while the complexity of the assimilating model increases (e.g. inclusion of highly nonlinear physics) or when data nonlinearly related to the model variables are used (e.g. humidity-sensitive TOVS channels).

Result 1.

The Hessian $\mathscr{J}''$ of the cost function $\mathscr{J}$ at the minimum is given by the equation

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

as can be seen by differentiating (1) twice.

Result 2.

The analysis error covariance matrix is the inverse of the Hessian, as can be found in Appendix B of Rabier and Courtier (1992). Calling x_a the result of the minimization (the analysis), and x_t the truth

$$\langle (x_a - x_t)(x_a - x_t)^T \rangle = \mathscr{J}''^{-1} = (B^{-1} + H^T O^{-1} H)^{-1}$$

where $\langle\ \rangle$ stands for the mathematical expectation. This result holds regardless of whether the error statistics are Gaussian or non-Gaussian. However if they are Gaussian, the analysis is the minimum variance estimate, and if they are non-Gaussian, then it is the minimum variance only for the linear estimates.

Result 3.

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

$$\langle\nabla\mathcal{J}\nabla\mathcal{J}^T\rangle = \mathcal{J}''. \tag{3}$$

Thus the error covariance matrix of $\nabla\mathcal{J}$ is equal to the Hessian.

(c)   Application to the preconditioning

In practice one of the difficulties of using (3) is that one does not know *a priori* the result x_a of the minimization. However, given that $\mathcal{J}''$ is independent of the values of the observations (since $\mathbf{H}$ is linear), we can define a minimization problem $\mathcal{P}_1$ for which we know the solution and which has the same Hessian as the original problem $\mathcal{P}$

$$\mathcal{P}_1: \text{minimize } \mathcal{J}_1(x) = \tfrac{1}{2}(x - x_b)\mathbf{B}^{-1}(x - x_b)^T + \tfrac{1}{2}(\mathbf{H}x - y_b)\mathbf{O}^{-1}(\mathbf{H}x - y_b)^T \tag{4}$$

with $y_b = \mathbf{H}x_b$. Clearly, $\mathcal{J}_1$ and $\mathcal{J}$ have the same Hessian given by (2). In addition, the minimum of $\mathcal{J}_1$ occurs when $x = x_b$. Defining the random variables $\tilde{x}_b = N(x_b, \mathbf{B})$ and $\tilde{y}_b = N(y_b, \mathbf{O})$ as Gaussian variables with expectations respectively x_b and y_b and covariances $\mathbf{B}$ and $\mathbf{O}$, we have

$$\mathcal{J}'' = \mathcal{J}_1'' = \langle\nabla\mathcal{J}_1 \nabla\mathcal{J}_1^T\rangle$$

with $\nabla\mathcal{J}_1 = \mathbf{B}^{-1}N(0, \mathbf{B}) + \mathbf{H}^T\mathbf{O}^{-1}N(0, \mathbf{O})$, 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 realizations δx_b^i and δy_b^i of the above random variables, we have

$$\mathcal{J}'' \approx \mathcal{J}_p'' = \frac{1}{p}\sum_{i=1}^{p} \nabla\mathcal{J}_1^i \nabla\mathcal{J}_1^{iT} \tag{5}$$

with

$$\nabla\mathcal{J}_1^i = \mathbf{B}^{-1}\delta x_b^i + \mathbf{H}^T\mathbf{O}^{-1}\delta y_b^i.$$

$\mathcal{J}_p''$ converges toward $\mathcal{J}''$ when p becomes large. It follows a Wishart law (Aitchison and Dunsmore 1975; Wishart 1952) which is the generalization of the multi-dimensional case of the χ^2 law. In practice, p will have to remain small. One immediately sees that the rank of $\mathcal{J}_p''$ is, at most, p, which implies that we cannot directly use $\mathcal{J}_p''$ as a preconditioning. Nevertheless we can extract useful information from this matrix.

Following Thépaut and Moll (1990) we shall, in the following, concentrate on the diagonal of the Hessian 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 put $p = 60$ to evaluate the potential of the approach.

(d)   Numerical results

(i) *Effectiveness of the diagonal preconditioning.* Following 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, however, remain meteorologically small. Extra attention had to be paid to the validation of the adjoint since the curl of the gradient of geopotential is no longer explicitly set to zero, and this led to roundoff errors in the test of the gradient. The technical details can be found in the paper by Courtier (1991).

All the experiments are designed along the same lines and belong to an identical twin framework. A reference forecast is made from $t_0 = 0$ to $t = 24$ h and the results of the integration at $t = 24$ h are used as 'observations'. One then tries to recover the state at time t_0. The initial condition of the reference forecast is the operational ECMWF initialized analysis valid for 1200 UTC 30 December 1991, truncated at total wave number 21, and the initial point of the minimization is a 6-hour forecast valid for the same time. Again following Thépaut and Courtier (1991), we define the cost function as the energy of the difference between the actual trajectory at time $t = 24$ h and the reference. In the stochastic interpretation this would imply that the background error was considered to be homogeneous and weakly correlated in space. However, with the addition of a geostrophic constraint, energy equipartition leads to realistic spatial structures (Phillips 1986). In this paper all the experiments have been performed using the M1QN3 minimization algorithm described by Gilbert and Lemaréchal (1989). In the experiments using preconditioning, (5) is used with an ensemble of 60 randomly chosen state vectors to estimate the diagonal of the Hessian. It should be noted that this experimental framework could not have resisted the impacts of deficiencies in the tangent-linear approximation validity in the preconditioning since the 'observations' in this case are an exact model solution with zero departure at the optimum.

Figure 1 shows the variation of the cost function during the course of the minimization with (thick line) and without (thin line) preconditioning. No horizontal diffusion was used. The preconditioning clearly has a positive impact: the same decrease of the cost function is obtained after 24 iterations with the preconditioning as with 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 by which the norm of the gradient is defined is not the same in the two cases because the preconditioning changes the metric. That is the reason why we have considered only relative variations here, and not absolute values.

If the Hessian were diagonal, then we should expect the present approach to be even more successful. 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 four hours in the troposphere, decreasing in the stratosphere roughly as the square root of the density of the standard atmosphere. The resolvent of this dynamic is diagonal in spectral space, and so the Hessian also is diagonal. Figure 3 shows the decrease of the cost function with (thick line) and without (thin line) preconditioning. The impact of the preconditioning is that it leads 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
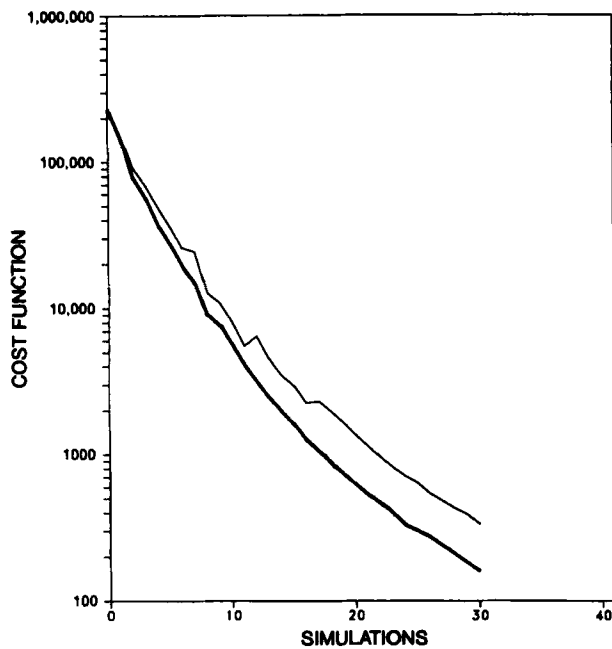
Figure 1.   Variation of the cost function during the course of the minimization without (thin line) and with (thick line) preconditioning. The minimization problem consists of a 24-hour inversion of a T21L19 adiabatic PE model.
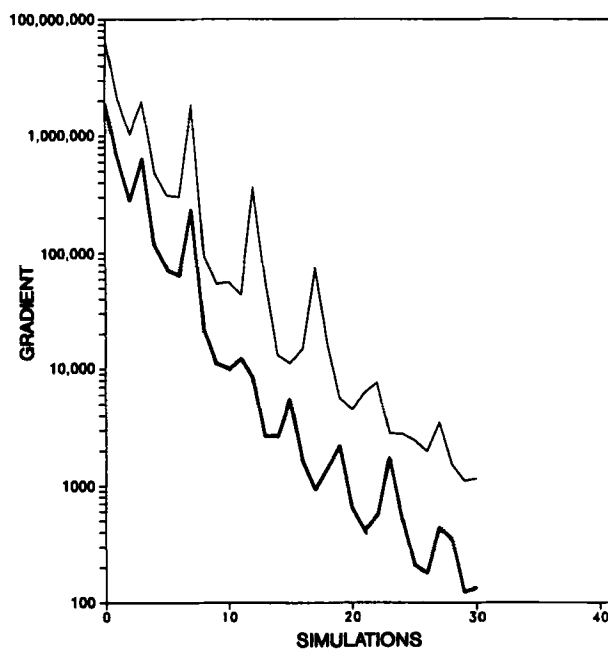

Figure 2.   Same as Fig. 1, but for the variation of the square norm of the gradient.
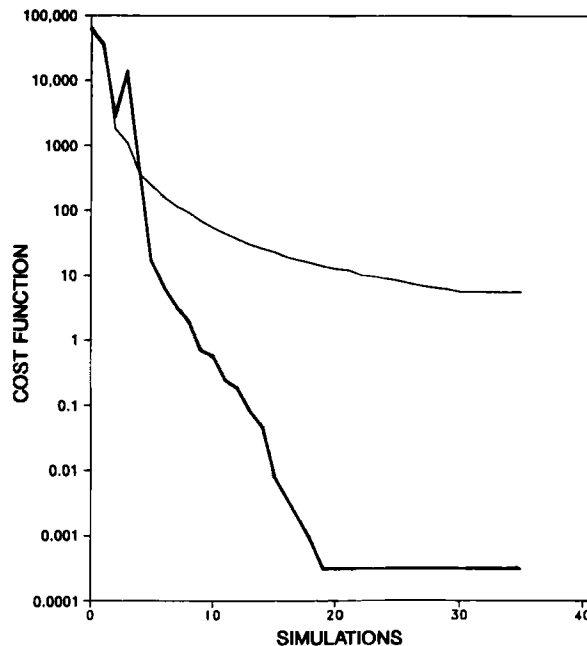
Figure 3.   Same as Fig. 1, but the model consists of only horizontal diffusion.

the use of the Kalman filter applied to barotropic models that this is certainly not the case, since the spatial variation of the estimation error arising from nonlinearities in the dynamics is substantial (Gauthier *et al.* 1993).

(ii)   *Scale dependence of the minimization.* Thépaut and Courtier (1991) also studied the ability of the minimization to recover the fields at time t_0. We begin by looking at the previous experiment with horizontal diffusion only and no preconditioning of the minimization. Figure 4 shows the spectrum with respect to the total wave number of the energy of the difference between the initial point of the minimization at time t_0 and the reference at the same time (thin line). The thick line represents the same spectrum, but at the end of the minimization. As expected, the field is recovered better for the large scales than for the small scales. We expect the near absence of minimization in the small scale to be corrected by the preconditioning.

The dashed line presents the same results but with the preconditioning active. Though the preconditioning has improved the behaviour of the minimization, as was seen in the previous subsection, it has adversely affected the behaviour in terms of the distance to the reference at time t_0. The convergence for wave numbers 0 to 4 has deteriorated slightly, but it has improved substantially from 6 to 14, which is the expected result. What was not expected is the marked deterioration from wave number 16 to 21.

Though surprising at first sight, this result has a geometrical interpretation. The shape of the cost function in the absence of preconditioning is elliptic, with the major semi-axis in the small-scale direction, since, owing to the diffusion, a large perturbation of the small scales leads to a weak perturbation of the cost function (see Fig. 5). The distance from the initial point of the minimization to the reference is of similar magnitude both in the large-scale and small-scale directions; as can be seen from the thin line of Fig. 4. It is thus located on a circle (black circle in Fig. 5). The minimization process
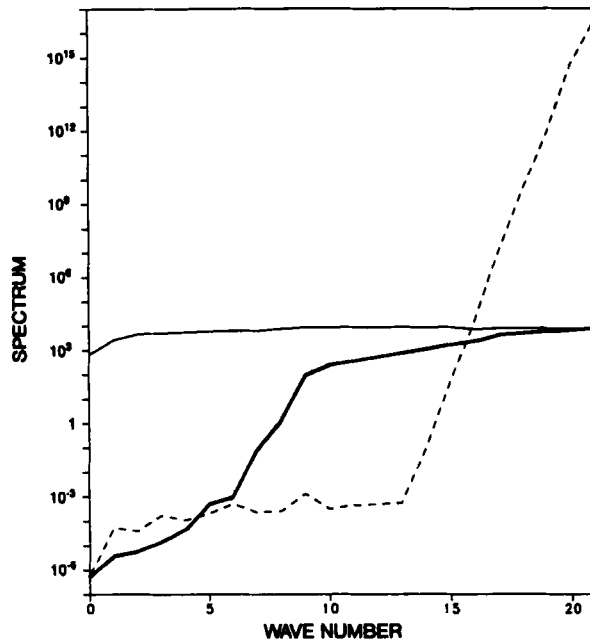
Figure 4. Spectrum with respect to the total wave number of the energy of the difference between the reference at time t_0 and (thin line) the initial point of the minimization; (thick line) the result of the minimization with no preconditioning; (dashed line) the result of the minimization with preconditioning.
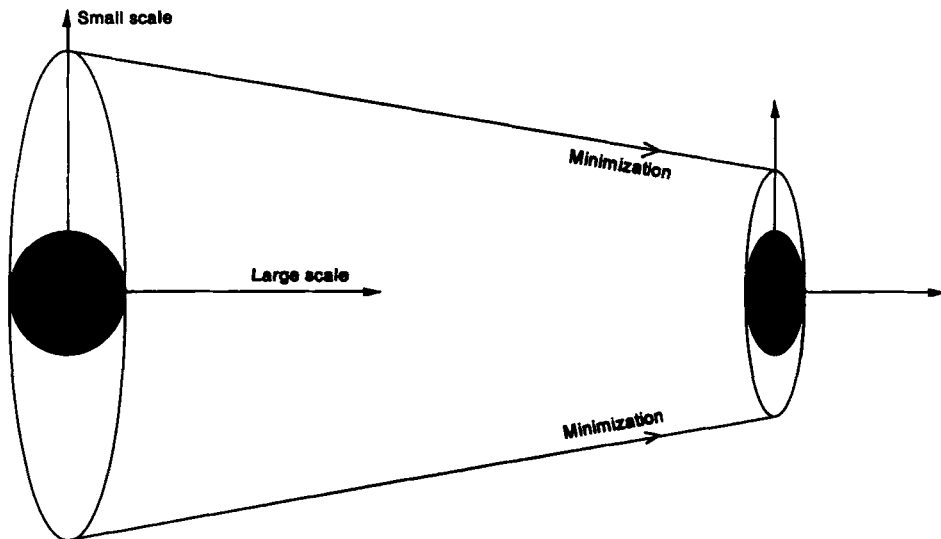

Figure 5. Schematic representation of the effect of the minimization in phase space.

attempts to decrease the cost function; it is thus a contraction of the phase space. Since the ellipticity of the cost function is important, the original circle is contracted in the large-scale direction but the small-scale direction is unaffected, since the gradient of the cost function is almost zero in the small-scale direction.

The effect of the preconditioning is to magnify the large-scale direction so as to obtain a spherical shape for the cost function. Figure 6(a) shows a schematic representation of what happens to the original metric, while in Fig. 6(b) is the modified metric. Again the minimization algorithm is a contraction, but in the modified metric. The small scales are now modified since the effect of preconditioning is to have the same order of magnitude for the gradient component in every scale. The result of the minimization can be located anywhere in the circle which, in the original metric, is strongly elliptic. As a consequence the small scales can be further from the minimum after the minimization than they were at the initial point, as we saw in Fig. 4.

Similar to Fig. 4, Fig. 7 shows the energy spectrum of the difference between the reference and the initial point of the minimization (thin line), the result of the minimization with no preconditioning (thick line), 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.
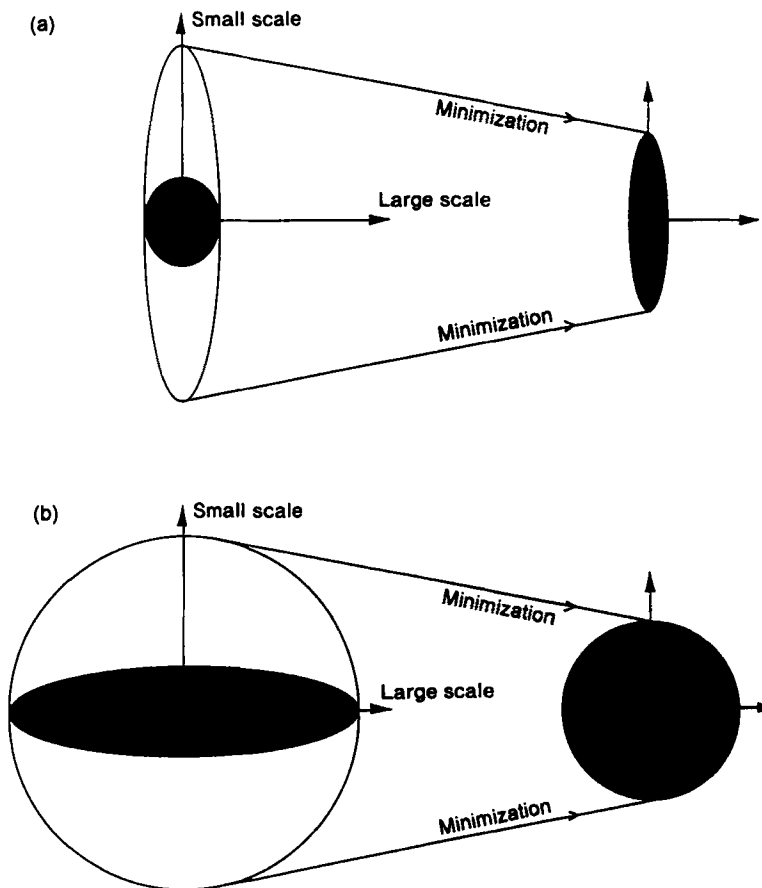

Figure 6.   Same as Fig. 5, but with the preconditioning (a) in the original metric; (b) in the metric defined by the preconditioning.
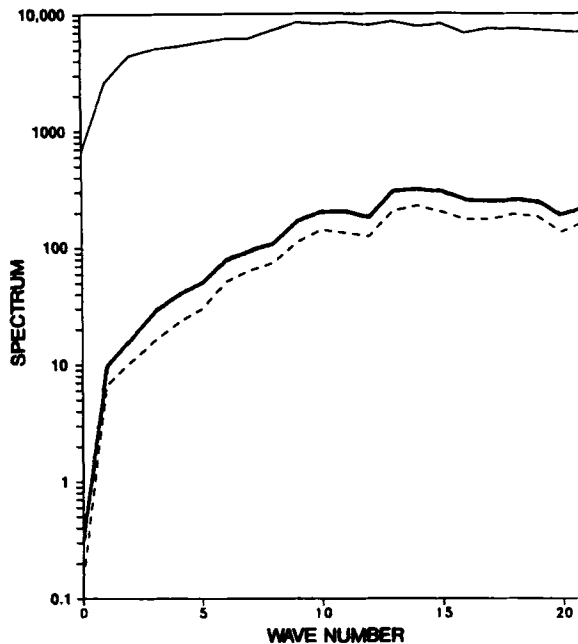
Figure 7.   Same as Fig. 4, but for the 24-hour inversion of a T21L19 adiabatic PE model.

(e)   Discussion

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

The first 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 unresolved:

(i)   *Cost*. In the experiment that we have described we used $p = 60$, which is equivalent to the cost of 60 gradient computations. Although it is possible to reduce this number, nevertheless using $p = 30$ did not change the results significantly (not shown). Moreover, the evaluation of the preconditioning could 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 for building up a preconditioning valid for today, based on yesterday's observation distribution. Another way of reducing the cost would be to evaluate the preconditioning valid for the larger scales (T21 for a T106 minimization problem).

(ii)   *Sufficiency*. A spectrally based (and therefore homogeneous) preconditioning was 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 would be to use information on the diagonal of the Hessian, expressed both in gridpoint space and in spectral space. This would not be too difficult to implement in practice.

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 most 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, and it is there that the variance of analysis error is maximum. In the extreme case of there being no data in an area, we would be at the minimum in this area already, and it would be unwise 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 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 in the practical implementation of variational algorithms.

In the current ECMWF implementation of 3D-Var, the preconditioning relies on the matrix $\mathbf{B}^{-1}$ only. In that context, the Hessian takes the form $\mathbf{I} + \mathbf{Q}$ where $\mathbf{Q}$ is a positive semi-definite matrix. All eigenvalues of the Hessian are then greater than unity. Use of this property in an algorithm to bound the spectrum may improve the conditioning while preserving the monotonicity of the convergence. If a Monte Carlo approach, as is the case here, is used to get a low-rank estimation of $\mathbf{Q}$, only the Sherman–Morrisson–Woodbury formulae will provide us with the inverse of $\mathbf{I} + \mathbf{Q}$, which is necessary for a preconditioning through a change of variable.

A further application of the algorithm, combined with the use of the Sherman–Woodbury–Morrison formulae, would be an evaluation of the analysis error variances, the Hessian being the inverse of the covariance matrix of analysis error as shown in section 2. Rabier and Courtier (1992) estimated only the diagonal of the inverse of the analysis error covariance, which is simply related to the variance field only if one makes assumptions about the off-diagonal elements. The variances are often used in operational assimilation to provide a spatial dependency of background errors used in the subsequent assimilation cycle.

3. 4D-VAR IN TERMS OF INCREMENTS

We now consider an alternative way to reduce the cost of 4D-Var. Let us denote by $\mathcal{M}(t_2, t_1)$ the model integrated from time t_1 to t_2; it is used to carry in time the state of the atmosphere, $\mathbf{x}$, i.e.

$$\mathbf{x}(t_2) = \mathcal{M}(t_2, t_1)\mathbf{x}(t_1). \tag{6}$$

Suppose that it is our intention to perform a 4D-Var assimilation over a period $(t_0, t_0 + T = t_N)$, (to fix ideas we let $T = 24$ hours). N is the number of time-steps necessary to integrate the model from time t_0 to time $t_0 + T$.

Over this time interval, observations $\mathbf{y}_i$ are available at each time t_i. We assume that all observations available between times $t_{(i-\frac{1}{2})}$ and $t_{(i+\frac{1}{2})}$ are valid for time t_i. This is not a serious limitation since we are already within the timescales resolved by the model. The observation $\mathbf{y}_i$ is linked to the model state variable $\mathbf{x}(t_i)$ by the observation operator $\mathbf{H}_i$, viz.

$$\mathbf{y}_i = \mathbf{H}_i \mathbf{x}(t_i) + \varepsilon_i. \tag{7}$$

The observational error, ε_i, of covariance matrix $\mathbf{O}_i$ is defined by (7), and consists of the sum of the measurement errors and the representativeness errors (Lorenc 1986).

The 4D-Var then consists of the minimization problem

$$\mathcal{P}_{4D}: \text{minimize } \mathcal{J}(\mathbf{x}(t_0)) = \tfrac{1}{2}\{\mathbf{x}(t_0) - \mathbf{x}_b\}^T \mathbf{B}^{-1}\{\mathbf{x}(t_0) - \mathbf{x}_b\} +$$

$$+ \tfrac{1}{2}\sum_{i=0}^{N}\{\mathbf{H}_i\mathbf{x}(t_i) - \mathbf{y}_i\}^T \mathbf{O}_i^{-1}\{\mathbf{H}_i\mathbf{x}(t_i) - \mathbf{y}_i\} \qquad (8)$$

with

$$\mathbf{x}(t_i) = \mathcal{M}(t_i, t_0)\mathbf{x}(t_0)$$

$\mathbf{x}_b$ is the background information valid for time t_0, which summarizes all the information used before time t_0, and $\mathbf{B}$ is the error covariance matrix of $\mathbf{x}_b$.

A classical result, assuming a perfect model and linearity of $\mathbf{H}$ and $\mathcal{M}$, is that if $\mathbf{x}^*(t_0)$ is the result of $\mathcal{P}_{4D}$, then $\mathbf{x}^*(t_N) = \mathcal{M}(t_N, t_0)\mathbf{x}^*(t_0)$ can also be obtained by applying the Kalman filter to the same statistical estimation problem (Jazwinski 1970; Ghil *et al.* 1981; Lorenc 1986. See Thépaut and Courtier 1991 or Rabier *et al.* 1993 for a detailed presentation using the same notations as here). In meteorological applications, however, $\mathbf{H}$ and $\mathcal{M}$ are weakly nonlinear: we assume that the tangent-linear operators $\mathcal{R}$ and $\mathbf{H}'$ of, respectively, $\mathcal{M}$ and $\mathbf{H}$ satisfy, to acceptable accuracy for meteorological orders of magnitude of the perturbation $\delta\mathbf{x}(t_0)$ (magnitude compared to the estimation error), the relations

$$\mathcal{M}(t_i, t_0)\{\mathbf{x}(t_0) + \delta\mathbf{x}(t_0)\} \approx \mathcal{M}(t_i, t_0)\mathbf{x}(t_0) + \mathcal{R}(t_i, t_0)\delta\mathbf{x}(t_0)$$

$$\mathbf{H}_i\{\mathbf{x}(t_i) + \delta\mathbf{x}(t_i)\} \approx \mathbf{H}_i\,\mathbf{x}(t_i) + \mathbf{H}' \cdot \delta\mathbf{x}(t_i). \qquad (9)$$

Under the linearity hypothesis (if (9) are exact) and the perfect model and fixed-lag-interval assumptions, the 4D-Var problem $\mathcal{P}_{4D}$ and the so-called Kalman filter are equivalent: they provide the same result at time t_N (see previous references). Under the weakly nonlinear hypothesis, the results of the 4D-Var problem $\mathcal{P}_{4D}$ and of the extended Kalman filter are expected to remain close. The extended Kalman filter consists of two steps denoted by the superscripts ^{f} and ^{a}, respectively, indicating forecast and analysis:

The forecast step.

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

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

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

The analysis step.

$$\mathbf{x}^a(t_i) = \mathbf{x}^f(t_i) + \mathbf{K}_i\{\mathbf{y}_i - \mathbf{H}_i\,\mathbf{x}(t_i)\} \qquad (11a)$$

$$\mathbf{B}^a(t_i) = (\mathbf{I} - \mathbf{K}_i\mathbf{H}_i')\mathbf{B}^f(t_i) \qquad (11b)$$

where

$$\mathbf{K}_i = \mathbf{B}^f(t_i)\mathbf{H}_i'^T\{\mathbf{H}_i'\mathbf{B}^f(t_i)\mathbf{H}_i'^T + \mathbf{O}_i\}^{-1} \qquad (11c)$$

is given by the minimum variance optimality condition.

The initial conditions $\mathbf{x}^f(t_0)$ and $\mathbf{P}^f(t_0)$ of the Kalman filter are

$$\mathbf{x}^f(t_0) = \mathbf{x}_b \qquad \mathbf{B}^f(t_0) = \mathbf{B}.$$

In current operational practice (10a) is treated exactly and (11a) and (11c) are solved to a good accuracy; however (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 implicitly uses flow-dependent structure functions in (10b) (Thépaut *et al.* 1993a), 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 (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 (10b) no source terms Q are present (Talagrand 1988; Cohn and Parrish 1991; Daley 1991; Wergen 1992), nevertheless Derber (1989) and Zupanski (1993a) demonstrated how to deal with a model bias in 4D-Var. Secondly, we do not have access to the analysis-error covariance $B^a(t_N)$. Here we suggest that if (10b) is approximate anyway, it is not scientifically worthwhile solving it exactly. This idea has been followed by most of the national weather prediction centres which implemented optimal interpolation; the dynamics $\mathcal{R}$ was replaced by the identity or a simple law for the temporal evolution of the variances. In the Kalman filter context, this was discussed e.g. by Dee (1991) and Cohn (1992). In other words, it may be scientifically acceptable to replace $\mathcal{R}$ by a linear model close to the tangent-linear model in (10b), provided this approximation is smaller than the approximation consisting of neglecting the model-error-source term Q.

Let us suppose from now on that $\mathcal{R}$ is any linear operator verifying $\mathcal{R}(t_0, t_0) = I$ (I being the identity operator) and that $\mathcal{R}(t'', t')\mathcal{R}(t', t) = \mathcal{R}(t'', t)$, for which we shall later stipulate the link with the model $\mathcal{M}$. We define the 4D-Var problem thus

$$\mathcal{P}'_{4D}: \text{minimize } \mathcal{J}(\delta x(t_0)) = \tfrac{1}{2}\delta x(t_0)^T B^{-1}\delta x(t_0) + \tfrac{1}{2}\sum_{i=0}^{N} (H_i x(t_i) - y_i)^T O_i^{-1}(H_i x(t_i) - y_i)$$

(12)

with

$$x(t_i) = \mathcal{M}(t_i, t_0)x_b + \mathcal{R}(t_i, t_0)\delta x(t_0)$$

(we then have $\delta x(t_0) = x(t_0) - x_b$).

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

Remark 2. If $\mathcal{R}$ is any linear operator which we assume would describe exactly the forecast error evolution, then $\mathcal{P}'_{4D}$ leads to the same result as the Kalman filter described by (10) and (11). However, in a nonlinear problem there is no linear operator which describes the error evolution exactly; introducing $\mathcal{R}$ will then remain an approximation.

Remark 3. $\mathcal{P}'_{4D}$ is better than $\mathcal{P}_{4D}$ so far as an operational implementation is concerned, since we keep the original model $\mathcal{M}$ for propagating in time the state of the atmosphere, but have introduced the possibility of using an approximate propagation in time of the errors, thus introducing some flexibility on the cost of 4D-Var.

Remark 4. A variant of $\mathcal{P}'_{4D}$ is the quadratic problem:

$$\mathcal{P}''_{4D}: \text{minimize } \mathcal{J}(\delta x(t_0)) = \tfrac{1}{2}\delta x(t_0)^T B^{-1}\delta x(t_0) +$$

$$+ \tfrac{1}{2}\sum_{i=0}^{N} (y_{b,i} + H'_i \delta x(t_i) - y_i)^T O_i^{-1}(y_{b,i} + H'_i \delta x(t_i) - y_i)$$

(13)

with

$$\delta x(t_i) = \mathcal{R}(t_i, t_0)\delta x(t_0)$$

and

$$\mathbf{y}_{\mathrm{b},i} = \mathbf{H}_i\{\mathcal{M}(t_i, t_0)\mathbf{x}_\mathrm{b}\}.$$

The costs of $\mathcal{P}'_{\mathrm{4D}}$ and $\mathcal{P}''_{\mathrm{4D}}$ are similar but the storage requirement for the background trajectory is different: in $\mathcal{P}'_{\mathrm{4D}}$, it is the background vertical column at the observation point, and in $\mathcal{P}''_{\mathrm{4D}}$ it is the observation equivalent of the background which has to be stored. In $\mathcal{P}'_{\mathrm{4D}}$, $\mathcal{R}$ is an approximate linearization of $\mathcal{M}$, similarly in $\mathcal{P}''_{\mathrm{4D}}$, $\mathbf{H}'_i$ is an approximate linearization of $\mathbf{H}_i$. Approximating the full problem by a quadratic problem has theoretical advantages since the solution in principle involves solving only linear equations, and is guaranteed to be unique.

The structure functions used in the current operational at ECMWF T213 optimal interpolation have a cut-off at wave number 63 (Lönnberg 1988). If we were to use a T106 truncation for $\mathcal{R}$, this would already be an enhancement in terms of resolution. An adiabatic version for $\mathcal{R}$ with some basic simplified physical processes such as 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.* (1993a).

The CPU cost of an adiabatic semi-Lagrangian T106 L31 model is, typically, 1/16 of the CPU cost of the T213 L31 version. The gain is then of the order of one magnitude.

Remark 5. Another variant of $\mathcal{P}'_{\mathrm{4D}}$ can be obtained by replacing $\mathbf{H}'_i\,\delta\mathbf{x}(t_i)$ in (13) by a finite difference, viz.

$$\mathcal{P}'''_{\mathrm{4D}} \text{ minimize } \mathcal{J}(\delta\mathbf{x}(t_0)) = \tfrac{1}{2}\delta\mathbf{x}(t_0)^{\mathrm{T}}\mathbf{B}^{-1}\delta\mathbf{x}(t_0) +$$

$$+ \tfrac{1}{2}\sum_{i=0}^{N} (\mathbf{y}_{\mathrm{b},i} - \tilde{\mathbf{y}}_{\mathrm{b},i} + \tilde{\mathbf{y}}_i - \mathbf{y}_i)^{\mathrm{T}}\mathbf{O}_i^{-1}(\mathbf{y}_{\mathrm{b},i} - \tilde{\mathbf{y}}_{\mathrm{b},i} + \tilde{\mathbf{y}}_i - \mathbf{y}_i)$$

with $\tilde{\mathbf{y}}_{\mathrm{b},i} = \mathbf{H}_i(\tilde{\mathcal{M}}(\tilde{\mathbf{x}}_\mathrm{b}(t_0)))$ being the model equivalent of the observation obtained from a simplified (low-resolution, adiabatic) background trajectory and $\tilde{\mathbf{y}}_i$ being the model equivalent of the observation obtained from the simplified trajectory originating from $\tilde{\mathbf{x}}_\mathrm{b}(t_0) + \delta\mathbf{x}(t_0)$. The term $\tilde{\mathbf{x}}_\mathrm{b}(t_0)$ is the background, but at lower resolution. Using the Taylor formula, to first order $\tilde{\mathbf{y}}_i - \tilde{\mathbf{y}}_{\mathrm{b}i} = \mathbf{H}'_i\,\delta\mathbf{x}(t_i)$. The practical advantage of this formulation over $\mathcal{P}''_{\mathrm{4D}}$ or $\mathcal{P}'_{\mathrm{4D}}$ is that there is less technical development required once the full 4D-Var problem has already been implemented. It is with this formulation that all the numerical experimentations presented later have been performed. These three implementations $\mathcal{P}'_{\mathrm{4D}}$, $\mathcal{P}''_{\mathrm{4D}}$ and $\mathcal{P}'''_{\mathrm{4D}}$ are equivalent in the linear case. They are expected to behave differently in the presence of strong nonlinearities, however we have not seen any reason why one should be superior to the others.

4. FURTHER DEVELOPMENTS

There are two ways of improving the model $\mathcal{R}$. Firstly one could increase the horizontal resolution: the main drawback here is the cost involved, since the CPU time follows a power law close to 3 as a function of linear resolution. In addition the trajectory storage and thus the input/output 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.* 1993a; Rabier 1992) have used only horizontal and vertical diffusion with a simple surface friction. Rabier *et al.* (1993) showed that large-scale condensation is essential to get reasonable humidity fields in the upper troposphere. More generally, it is expected that the important feedback loops present in the model $\mathcal{M}$ will have to be described to a reasonable accuracy with $\mathcal{R}$; this is expected to be of particular importance in the tropics. Zou *et al.* (1993)

and D. Zupanski (1993) have done feasibility studies using the adjoint of physical parametrizations. The automatic methods developed at INRIA will assist us in formulating a series of tangent-linear models including progressively more and more effects of the physics (Rostaing *et al.* 1993).

This will eventually double the CPU cost of 4D-Var compared to an adiabatic version (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 input/output). Currently, only *t*-values are stored, since the dynamics are nonlinear only with respect to these *t*-values, and not $(t - \Delta t)$. 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 two-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 $\mathcal{P}''_{4D}$, is not necessarily a very good approximation of $\mathcal{P}_{4D}$. A simple way of accounting for some of the nonlinearities in the final analysis is to define a sequence $\mathcal{P}''_{4D}(n)$ of assimilation, viz.

$$\mathcal{P}''_{4D}(n): \text{minimize } \mathcal{J}(\delta \mathbf{x}^n(t_0)) = \tfrac{1}{2}\{\delta \mathbf{x}^n(t_0) + \mathbf{x}^{n-1} - \mathbf{x}_b\}^{\mathrm{T}}\mathbf{B}^{-1}\{\delta \mathbf{x}^n(t_0) + \mathbf{x}^{n-1} - \mathbf{x}_b\} +$$

$$+ \tfrac{1}{2}\sum_{i=0}^{N}\{\mathbf{y}_i^{n-1} + \mathbf{H}_i'\,\delta \mathbf{x}^n(t_i) - \mathbf{y}_i\}^{\mathrm{T}}\mathbf{O}_i^{-1}\{\mathbf{y}_i^{n-1} + \mathbf{H}_i'\,\delta \mathbf{x}^n(t_i) - \mathbf{y}_i\} \qquad (14)$$

with

$$\delta \mathbf{x}(t_i) = \mathcal{R}(t_i, t_0)\delta \mathbf{x}(t_0) \qquad (15)$$

and

$$\mathbf{y}_i^{n-1} = \mathbf{H}_i\{\mathcal{M}(t_i, t_0)(\mathbf{x}^{n-2} + \delta^{*n-1}\mathbf{x}(t_0))\} = \mathbf{H}_i\{\mathcal{M}(t_i, t_0)(\mathbf{x}^{n-1})\} \qquad (16)$$

$\delta^{*n-1}\mathbf{x}(t_0)$ is the result of the (approximate) minimization of $\mathcal{P}''_{4D}(n-1)$; $\delta^{*-1}\mathbf{x}(t_0) = 0$ and

$$\mathbf{x}^n = \mathbf{x}^{n-1} + \delta^{*n}\mathbf{x}(t_0) \qquad (\mathbf{x}^0 = \mathbf{x}_b, \quad \mathbf{x}^{-1} = \mathbf{x}_b).$$

This algorithm can be seen as a pair of nested loops. At each iteration of the outer loop, the complete model is used through (16) to redefine the model trajectory. The inner loop uses the tangent-linear and adjoint of a simpler (e.g. adiabatic) model (15) to minimize the cost function (14) for the increments calculated with respect to the redefined trajectory.

This approach allows a progressive inclusion of physical processes without having to deal 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 $\delta^{*n}\mathbf{x}(t_0)$ will converge. Experimental work is necessary to tackle 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. The term $\mathcal{R}$ 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 included increase with n.

5. NUMERICAL RESULTS WITH THE INCREMENTAL APPROACH

Figure 8 presents the variation of the contribution to the cost function of both the background term $\mathcal{J}_b$ and the departure to the observations $\mathcal{J}_o$. We are applying the
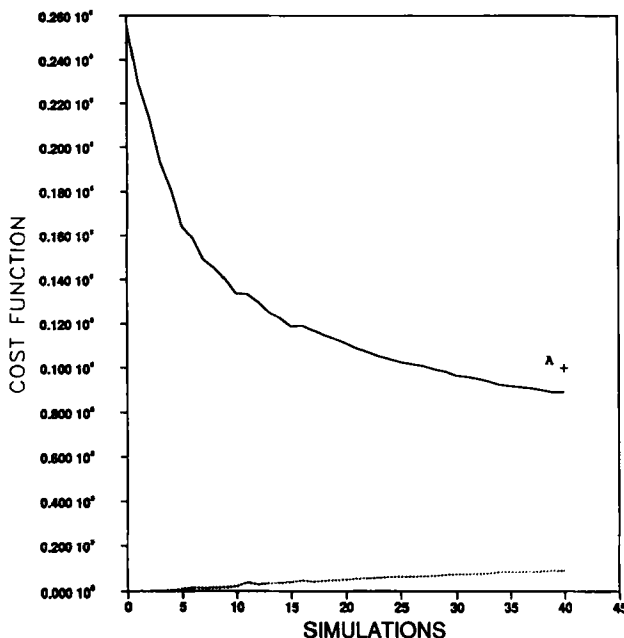
Figure 8. Variation of the contribution to the cost function of both the background term (. . .) and the observation term (—). Point A measures the distance to the observations after an update of the trajectory at full resolution.

formulation described in remark 5 of section 3 with increments at truncation T42. The background trajectory is performed at T63 with the physical parametrization of the operational ECMWF model. The increment's evolution is governed by a T42 adiabatic model with horizontal diffusion and a single surface drag/vertical diffusion scheme. The time interval is 12 hours and we are using all conventional observations as in the paper by Thépaut *et al.* (1993b). The $\mathcal{J}_b$ formulation is described precisely by Courtier *et al.* (1993) and is similar in spirit to the spectral statistical interpolation implementation of Parrish and Derber (1992).

As expected, $\mathcal{J}_o$ decreases at the expense of a $\mathcal{J}_b$ increase (the initial point of the minimization is the background, so $\mathcal{J}_b$ is equal to zero at the beginning of the minimization). The minimization was stopped after 40 iterations and then the distance to the observation was recomputed at full resolution (point A). The increase is small but is present, showing that, indeed, owing to some nonlinearities, the incremental approach is suboptimal. However, the suboptimality is not likely to be critical since the increase in $\mathcal{J}_o$ remains small.

In section 4, we presented a way of accounting for some of the nonlinearities, namely with an update of the reference trajectory. Figure 9, similarly to Fig. 8, presents the variation of $\mathcal{J}_o$ and $\mathcal{J}_b$, but with an update of the reference trajectory every 10 iterations of the minimization scheme. The result, in terms of the magnitude of $\mathcal{J}_o$, is worse than without the update. However, one immediately notices the concave curvature of the curve after each update. This result can be easily interpreted bearing in mind that we are using a limited memory quasi-Newtonian minimization scheme which builds an approximation of the Hessian during the course of the descent. At every restart of the minimization, the information on the Hessian previously accumulated is lost. The naïve idea to overcome the problem is thus, after the trajectory update, to keep the information
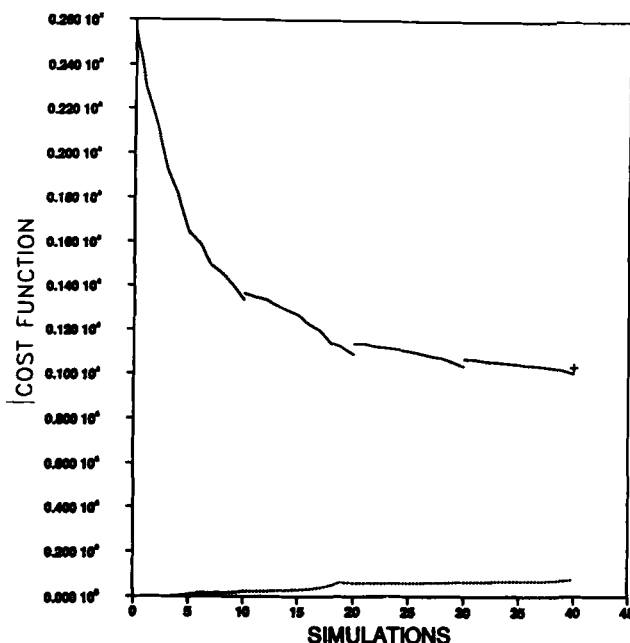
Figure 9. Same as Fig. 8, but with four updates of the trajectory.

accumulated during the previous 10 iterations. J.-C. Gilbert (INRIA) made the necessary developments to M1QN3; his results are presented in Fig. 10. As expected, the decrease of $\mathcal{J}_o$ now remains convex. It is worth noticing that the distance to the observations of the full-physics trajectory (point B) is now smaller than without updates (point A). At first sight the improvement might appear to be small; however, it should bring robustness to the system since $\mathcal{R}$, with the trajectory updates, is a linearization of a trajectory which is constrained by the observations. In other words, the implicit structure functions, as described by Thépaut et al. (1993a), are obtained by a dynamic consistent with the observations. The above results have shown that we do minimize the cost functions well. The effect of the approximations on actual analysis quality remains to be judged. Courtier et al. (1993) addresses this point.

Courtier (1987) mentioned the idea that, as the observational network is quasi-periodic with a 12-hour and 24-hour period, it should be possible to use the information on the Hessian accumulated the day before for preconditioning today's problem. In Fig. 11, one sees the improvement in the decrease of $\mathcal{J}_o$ brought about by using the information accumulated, during a 12-hour 4D-Var problem, for the 12-hour 4D-Var minimization that followed. There is a remarkable speeding up, thus showing the potential for further improvements through preconditioning.

6. CONCLUSION

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. Owing to the large dimension of the problem, it is only possible to improve the conditioning to
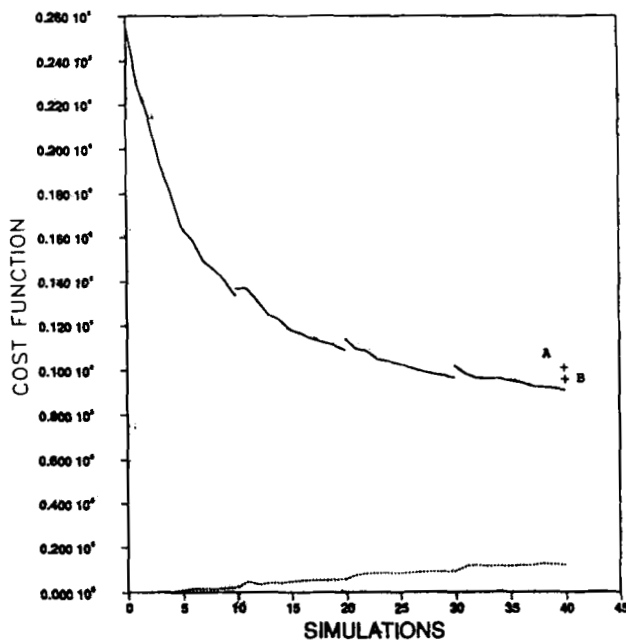
Figure 10. Same as Fig. 9, but keeping the information of the Hessian accumulated during the previous minimization. Point A of Fig. 8 has been added to this figure; the difference between point A and point B measures the (small) improvement brought about by the update of the trajectory.
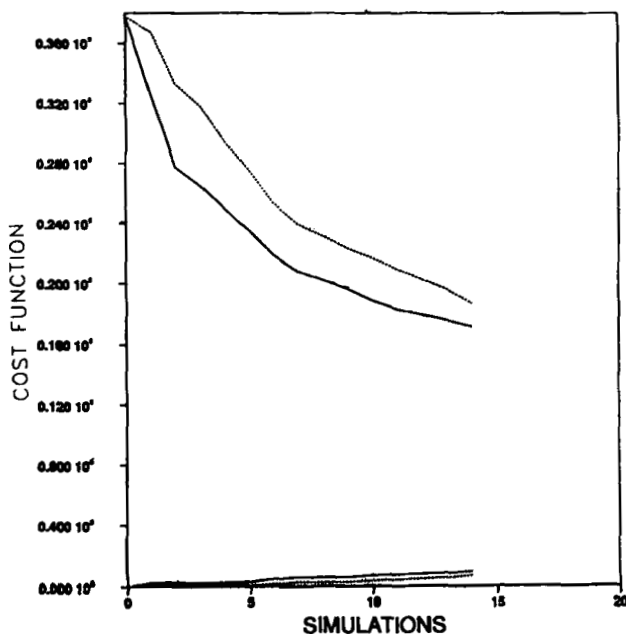

Figure 11. The dotted curve is the same as Fig. 8. The dotted curves represent the variation of the background term (bottom curve) and the observation term (top curve) when no preconditioning is used in the minimizations. The full line is obtained using the information on the Hessian accumulated during the previous 12 hours.

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, several scientific issues remain open, such as quality control, a better specification of the background error and accounting for model errors, all of which will require a substantial experimental programme. Nevertheless this approach offers good prospects for success. In the last section we have indeed shown that the approach works numerically. The scientific evaluation is currently being done at ECMWF.

REFERENCES

Aitchison, J. and Dunsmore, I. R. 1975 *Statistical prediction analysis*. Cambridge University Press

Andersson, E., Pailleux, J., Thépaut, J.-N., Eyre, J. R., McNally, A. P., Kelly, G. A. and Courtier, P. 1994 Use of cloud-cleared radiances in three/four-dimensional variational assimilation. *Q. J. R. Meteorol. Soc.*, **120**, 627–653

Cohn, S. E. 1992 Short-term dynamics of forecast error covariances. Pp. 157–170 in Proc. ECMWF workshop on variational assimilation with special emphasis on three-dimensional aspects, 9–12 November 1992, Reading, UK

Cohn, S. E. and Parrish, D. F. 1991 The behaviour of forecast error covariances for a Kalman filter in two dimensions. *Mon. Weather. Rev.*, **119**, 1757–1785

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

1991 'Two difficulties encountered in the validation of cycle 7'. Note ARPEGE No. 24 (available from METEO-FRANCE)

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

Courtier, P., Thépaut, J.-N. and Hollingsworth, A. 1992 A strategy for operational implementation of 4D-Var. Pp. 437–464 in Proc. 1992 ECMWF workshop on variational assimilation with special emphasis on three-dimensional aspects, 9–12 November 1992, Reading, UK

Courtier, P., Andersson, E., Heckley, W., Kelly, G., Pailleux, J., Rabier, F., Thépaut, J.-N., Undén, P., Vasiljevic, D., Cardinali, C., Eyre, J., Hamrud, M., Haseler, J., Hollingsworth, A., McNally, A. and Stoffelen, A. 1993 Variational assimilation at ECMWF. ECMWF Tech. Memo 194 (available from ECMWF)

Daley, R. 1991 *Atmospheric data analysis*. Cambridge atmospherics and space science series, Cambridge University Press

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

Dee, D. P.	1991	Simplification of the Kalman filter for meteorological data assimilation. *Q. J. R. Meteorol. Soc.*, **117**, 365–384
Derber, J. C.	1989	A variational continuous assimilation technique. *Mon. Weather Rev.*, **117**, 2437–2446
Forsythe, G. E. and Strauss, E. G.	1955	On best conditioned matrices. *Proc. Am. 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., Courtier, P. and Moll, P.	1993	Assimilation of simulated wind lidar data with Kalman filter. *Mon. Weather Rev.*, **121**, 272–289
Ghil, M., Cohn, S., Tavanzis, J., Bube, K. and Isaacson, E.	1981	'Application of estimation theory to numerical weather prediction'. *Dynamic Meteorology: Data assimilation method*. Eds. L. Bengtsson, M. Ghil and E. Källen, Springer-Verlag
Gilbert, J. C. and Lemaréchal, C.	1989	Some numerical experiments with variable storage quasi-Newton algorithms. *Math. Prog.* **B25**, 407–435
Gill, P. E., Murray, W. and Wright, M. H.	1981	*Practical optimization.* Academic Press, London
Golub, G. H. and O'Leary, D. P.	1989	Some history of the conjugate gradient and Lanczos algorithm: 1948–1976. *SIAM Rev.*, **31**, 50–102
Heckley, W. A., Courtier, P., Pailleux, J. and Andersson, E.	1992	On the use of background information in the variational analysis at ECMWF. Proc. ECMWF workshop on variational assimilation with emphasis on 3D aspects. 9–12 November 1992, Reading, UK
Hestenes, M. R. and Stiefel, E.	1952	Methods of conjugate gradients for solving linear systems. *J. Res. Nat. Bur. Stand.*, **49**, 409–436
Jazwinski, A. H.	1970	*Stochastic processes and filtering theory.* Academic Press, New York
Lacarra, J. F. and Talagrand, O.	1988	Short-range evolution of small perturbations in a barotropic model. *Tellus*, **40A**, 81–95
Le Dimet, F.-X. and Talagrand, O.	1986	Variational algorithms for analysis and assimilation of meteorological observations. *Tellus*, **38A**, 97–110
Lönnberg, P.	1988	Developments in the ECMWF analysis system. Pp. 75–119 in Proc. 1988 ECMWF seminar on data assimilation and the use of satellite data, Reading, UK
Lorenc, A. C.	1986	Analysis methods for numerical weather prediction. *Q. J. R. Meteorol. Soc.*, **112**, 1177–1194
	1988	Optimal nonlinear objective analysis. *Q. J. R. Meteorol. Soc.*, **114**, 205–240
Navon, I. M. and Legler, D. M.	1987	Conjugate-gradients method for large-scale minimization in meteorology. *Mon. Weather Rev.*, **115**, 1479–1502
Navon, I. M., Zou, X., Derber, J. and Sela, J.	1992	Variational data assimilation with an adiabatic version of the NMC spectral model. *Mon. Weather. Rev.*, **120**, 1433–1446
Parrish, D. F. and Derber, J. C.	1992	The National Meteorological Center's spectral statistical interpolation analysis system. *Mon. Weather Rev.*, **120**, 1747–1763
Phillips, N. A.	1986	The spatial statistics of random geostrophic modes and first-guess errors. *Tellus*, **38A**, 314–332
Rabier, F. and Courtier, P.	1992	Four-dimensional assimilation in the presence of baroclinic instability. *Q. J. R. Meteorol. Soc.*, **118**, 649–672
Rabier, F., Courtier, P., Pailleux, J., Talagrand, O. and Vasiljevic, D.	1993	A comparison between four-dimensional variational assimilation and simplified sequential assimilation relying on three-dimensional variational analysis. *Q. J. R. Meteorol. Soc.*, **119**, 845–880
Rostaing, N., Dalmas, S. and Galligo, A.	1993	Automatic differentiation in Odyssée. *Tellus*, **45A**, 558–568
Talagrand, O.	1988	Four-dimensional variational assimilation. Proc. 1988 ECMWF seminar on data assimilation and the use of satellite data, Reading, UK
Thépaut, J.-N. and Courtier, P.	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 Moll, P.	1990	Variational inversion of simulated TOVS radiances using the adjoint technique. *Q. J. R. Meteorol. Soc.*, **116**, 1425–1448

Thépaut, J.-N., Hoffman, R. and Courtier, P.	1993a	Interactions of dynamics and observations in a four-dimensional variational assimilation. *Mon. Weather Rev.*, **121**, 3393–3414
Thépaut, J.-N., Vasiljevic, D., Courtier, P. and Pailleux, J.	1993b	Variational assimilation of conventional meteorological observations with a multilevel primitive-equation model. *Q. J. R. Meteorol. Soc.*, **119**, 153–186
Vukicevic, T.	1991	Nonlinear and linear evolution of initial forecast errors. *Mon. Weather Rev.*, **119**, 1602–1611
Wergen, W.	1992	The effect of model errors in variational assimilation. *Tellus*, **44A**, 297–313
Wishart, J.	1952	Moment coefficients of the k-statistics in samples from a finite population. *Biometrika*, **39**, 1–13
Zou, X., Navon, I. M. and Sela, J.	1993	Variational data assimilation with moist threshold processes using NMC spectral model. *Tellus*, **45A**, 370–387
Zupanski, D.	1993	The effects of discontinuities in the Betts-Miller cumulus convection scheme on four-dimensional variational data assimilation. *Tellus*, **45A**, 511–524
Zupanski, M.	1993a	Regional four-dimensional variational data assimilation in a quasi-operational forecasting environment. *Mon. Weather Rev.*, **121**, 2396–2408
	1993b	A preconditioning algorithm for large-scale minimization problems. *Tellus*, **45A**, 478–492