VARIATIONAL ASSIMILATION OF METEOROLOGICAL OBSERVATIONS WITH THE DIRECT AND ADJOINT SHALLOW-WATER EQUATIONS

Philippe Courtier* and Olivier Talagrand+
European Centre for Medium-Range Weather Forecasts
Reading, U.K.

Abstract

Experiments of variational assimilation, similar to those already performed by the authors on a vorticity equation model (Q.J.R. Meteorol.Soc., 113, 1329-1347, 1987), are performed on a shallow-water equation model. The variational algorithm requires the computation of the gradient of the distance function to be minimized with respect to the model state at the beginning of the assimilation period. As in the previous experiments, this gradient is computed by using the adjoint equations of the model.

Northern Hemisphere observations of wind and geopotential, distributed at the 500 mb level over a 24 hour time period, are assimilated with a pseudo-spectral model truncated at degree 21. The results confirm the results previously obtained, namely that the variational process reconstructs to a satisfactory degree of accuracy the meteorological structures of the flow. In addition:

- Gravity wave noise can be efficiently eliminated by adding an appropriate penalty term to the distance function, and by introducing in the variational process a nonlinear normal mode initialization algorithm. The latter has the effect of improving the numerical conditioning of the variational process.

- The quality of forecasts produced from the results of variational assimilation is similar to the quality of shallow-water equation forecasts produced from the results of operational assimilations, which use many more data and more realistic models.

Assimilations of observations at the 300 mb level produce similar results, with differences which can be ascribed to the fact that a shallow-water equation model is a poorer simulator of the 300 mb circulation. Assimilations performed with a model truncated of degree 42 also produce similar results. They also show that the numerical efficiency of the variational process, as measured by the number of descent steps necessary to reach convergence, is almost insensitive to the dimension of the model phase space.

Finally, study of the variations of the distance-function suggests that, as in the case of the vorticity equation, the tangent linear approximation to the model equations is valid in the conditions of data assimilation.

---

*Current affiliation: Direction de la Météorologie, Paris, France
+Current affiliation: Laboratoire de Météorologie Dynamique, Paris, France

1. INTRODUCTION

Variational assimilation of meteorological or oceanographical observations has been studied in the last few years by a number of authors (see, e.g., Lewis and Derber, 1985; Hoffman, 1986; Le Dimet and Talagrand, 1986; Derber, 1987; Talagrand and Courtier, 1987; Courtier and Talagrand, 1987; Lorenc, 1988a and b, Thacker and Long, 1988). Variational assimilation is one of the many applications of optimal control, i.e. of the development of variational methods brought about by the simultaneous development of numerical analysis and computer power. A basic book of the theory of optimal control is Lions (1971). The principle of variational assimilation is extremely simple. A set of observations being given over some time interval (t_0, t_1), together with a numerical model of the atmospheric or oceanic flow, one first defines a scalar function which, for any solution of the model over (t_0, t_1), measures the "distance" between that solution and the observations. One then seeks the particular model solution which minimizes that distance function. In view of the complexity and the extremely large size of the corresponding minimization problem, the only practical way to numerically determine the minimizing solution is apparently to take the model initial conditions at time t_0 as the control variables of the problem. In the established terminology of optimal control, this means that the distance function is considered, through the model equations, as a function of these initial conditions only. A descent algorithm, which computes successive approximations of the minimizing values of the initial variables, is then implemented. Each step of the descent requires the explicit knowledge of the local vector of the partial derivatives, or gradient, of the distance function with respect to the control variables. As explained in detail in Talagrand and Courtier (1987), the most economical way of determining that gradient in the case of variational assimilation is to integrate the adjoint equations of the numerical model.

The various numerical experiments which have been performed so far show that variational assimilation does numerically converge to a solution which minimizes the distance function. They also show, as far as can be judged from the relative simplicity of the models used so far, that the results thus obtained are physically quite reasonable. The present article, a large part of which has already been presented in Courtier (1987), describes experiments which are very similar, in their conception, their implementation, and also in their conclusions, to experiments previously described by the authors in Courtier and Talagrand (1987) (hereafter referred to as CT). The main difference is that the experiments described here have been performed with the shallow-water equations, while the experiments described in CT had been performed with the vorticity equation. The particular problem of enforcing an appropriate quasi-geostrophic balance between the mass and velocity fields is therefore addressed here. The conclusion is that it is relatively easy to implement variational assimilation in such a way as to avoid an unrealistically large amount of gravity waves in the results. A similar conclusion has been reached by Lorenc (1988a). Another conclusion presented here is that the quality of the variational

assimilation, as judged from the quality of a subsequent 24 hour shallow-water equation forecast, is similar to the quality of operational assimilations.

After a brief reminder of how the adjoint equations of a numerical model can be used in the context of variational assimilation (Section 2), we describe the model and the observations used in the numerical experiments (Sections 3 and 4). The results of these experiments are then presented and discussed, especially in terms of the amount of gravity waves they contain (Section 5). It is shown in particular that the Machenhauer condition for nonlinear geostrophic balance can be successfully introduced in the assimilation process. A number of additional results are presented in Sections 6 to 8, especially as regards the quality of the forecast produced from the results of variational assimilations (Section 6). Numerical aspects are discussed in Section 9. General conclusions are given, and some of the remaining problems are discussed, in Section 10.

2. THE USE OF ADJOINT EQUATIONS AS AN EFFICIENT TOOL FOR COMPUTING GRADIENTS

We summarize in this section the main results of Section 2 of Talagrand and Courtier (1987). A numerical model is available which explicitly integrates, for given initial conditions $x(t_0)$ at time t_0, a set of differential equations written in a synthetic form as

$$\frac{dx}{dt} = F(x) \tag{2.1}$$

In this equation, x(t) is the state vector of the model at time t, which belongs to a phase space E on which an inner product, denoted by $< , >$ has been defined. We consider a scalar function defined for any solution x(t) ($t_0 \leq t \leq t_1$) of the model by

$$= \int_{t_0}^{t_1} H\,[x(t), t]\, dt \tag{2.2}$$

where $(x,t) \rightarrow H[x,t]$ is a regular scalar-valued function defined on E (t_0, t_1). is a uniquely defined function of the initial conditions $u = x(t_0)$ and we want to determine the gradient of with respect to u. This gradient, which we shall denote ∇_u, is characterized by the property that for any variation δu of u, the corresponding first order variation δ of is given by the Taylor formula

$$\delta = < \nabla_u , \delta u > \tag{2.3}$$

Linearizing eq. (2.1) in the vicinity of a particular solution x(t) leads to the corresponding tangent linear equation:

$$\frac{d\,\delta x}{dt} = F'(t)\,\delta x \qquad (2.4)$$

whose variable is δx, and where, for any time t, F'(t) is the Jacobian operator determined by differentiating F(x) with respect to x at point x(t) in the phase space **E**. The tangent linear equation describes the temporal evolution of a perturbation $\delta x(t)$ imposed on the solution x(t), to first order with respect to the initial value $\delta u = \delta x(t_0)$ of δx.

We will denote F'*(t) the adjoint of F'(t) with respect to the inner product $< , >$.

These definitions being given, the distance function and its gradient ∇_u can be numerically determined for given u by performing the following operations

a) Starting from the initial conditions $x(t_0) = u$, integrate the basic equation (2.1) from t_0 to t_1, compute from the solution x(t) thus obtained, and store the values of the solution over the entire time interval (t_0, t_1).

b) Starting from the "final" condition $\delta'x(t_1) = 0$, integrate the adjoint equation

$$-\frac{d\,\delta'x}{dt} = F'^*(t)\,\delta'x + \nabla_x H(t) \qquad (2.5)$$

backward in time from t_1 to t_0. The result $\delta'x(t_0)$ obtained at time t_0 is the gradient ∇_u . In (2.5), $\nabla_x H(t)$ is the gradient with respect to x of the integrand H(x,t) of (2.2), taken at point x(t). Both terms F'*(t) $\delta'x$ and $\nabla_x H(t)$ of (2.5) are computed at time t of the adjoint integration from the value x(t) computed in the direct integration a).

Remark At any time t, the quantity $\delta'x(t)$ produced by the adjoint integration (2.5) is the gradient, with respect to x(t), of the integral

$$\int_t^{t_1} H\,[x(t'), t']\,dt'$$

limited to the interval (t,t_1).

3. THE NUMERICAL MODEL AND ITS ADJOINT

The shallow-water equations at the surface of a rotating sphere Σ with radius a, expressed in terms of the vorticity ζ and the divergence η of the velocity field and of the geopotential φ of the fluid, assumed to be inviscid, read

$$\frac{\partial \zeta}{\partial t} = J(\zeta+f,\Delta^{-1}\zeta) - \nabla.[(\zeta+f)\nabla\Delta^{-1}\eta]$$

$$\frac{\partial \eta}{\partial t} = J(\zeta+f,\Delta^{-1}\eta) + \nabla.[(\zeta+f)\nabla\Delta^{-1}\zeta] - \Delta\varphi - \Delta K \qquad (3.1)$$

$$\frac{\partial \varphi}{\partial t} = J(\varphi,\Delta^{-1}\zeta) - \nabla.(\varphi\nabla\Delta^{-1}\eta)$$

where K is the kinetic energy per unit mass

$$K = \tfrac{1}{2}[\nabla\Delta^{-1}\zeta.\nabla\Delta^{-1}\zeta + \nabla\Delta^{-1}\eta.\nabla\Delta^{-1}\eta + 2J(\Delta^{-1}\zeta,\Delta^{-1}\eta)] ,$$

f is the vorticity of the basic rotation, ∇ now denotes the first order differential operator along Σ, and Δ and J are respectively the Laplacian and Jacobian operators along Σ. The state of the flow at a given time is entirely defined by the knowledge of ζ, η and φ along Σ at that time.

For a given solution $(\zeta(t), \eta(t), \varphi(t))$ of eqs. (3.1), the tangent linear equations, analogous to (2.4), read

$$\frac{\partial \delta\zeta}{\partial t} = J(\delta\zeta,\Delta^{-1}\zeta) + J(\zeta+f,\Delta^{-1}\delta\zeta) - \nabla.[\delta\zeta\nabla\Delta^{-1}\eta] - \nabla.[(\zeta+f)\nabla\Delta^{-1}\delta\eta)]$$

$$\frac{\partial \delta\eta}{\partial t} = J(\delta\zeta,\Delta^{-1}\eta) + J(\zeta+f,\Delta^{-1}\delta\eta) + \nabla.(\delta\zeta\nabla\Delta^{-1}\zeta) + \nabla.[(\zeta+f)\nabla\Delta^{-1}\delta\zeta)]$$

$$- \Delta[\delta\varphi + \nabla\Delta^{-1}\zeta.\nabla\Delta^{-1}\delta\zeta + \nabla\Delta^{-1}\eta.\nabla\Delta^{-1}\delta\eta] \qquad (3.2)$$

$$- \Delta[J(\Delta^{-1}\delta\zeta,\Delta^{-1}\eta) + J(\Delta^{-1}\zeta,\Delta^{-1}\delta\eta)]$$

$$\frac{\partial \delta\varphi}{\partial t} = J(\delta\varphi,\Delta^{-1}\zeta) + J(\varphi,\Delta^{-1}\delta\zeta) - \nabla.(\delta\varphi\nabla\Delta^{-1}\eta) - \underline{\nabla.(\varphi\nabla\Delta^{-1}\delta\eta)}$$

As for the definition of the inner product on the space of all possible states of the flow, without which the adjoint of (3.2) cannot be defined, we have chosen the inner product associated with the quadratic form

$$E = \frac{1}{a^2}\int_\Sigma (\frac{1}{2}\varphi^2 + \Phi_0 K)\, d\Sigma$$

$$= \frac{1}{a^2}\int_\Sigma [\tfrac{1}{2}\varphi^2 + \tfrac{1}{2}\Phi_0(\nabla\Delta^{-1}\zeta.\nabla\Delta^{-1}\zeta + \nabla\Delta^{-1}\eta.\nabla\Delta^{-1}\eta]\, d\Sigma \qquad (3.3)$$

where Φ_0 is a constant geopotential. E is an invariant of the tangent linear equations in the vicinity of the state of rest ($\varphi = \Phi_0$, $\zeta = \eta = 0$). It is in these circumstances the second order approximation of the total energy $\int_\Sigma (\frac{1}{2}\varphi^2 + \varphi\ K)d\Sigma$, which is an invariant of eqs. (3.1). The norm associated to the quadratic form (3.3) will be denoted $\|\ \|$. It has physical dimension (length)^2 (time)$^{-2}$. The adjoint of eqs. (3.2) with respect to the inner product defined by (3.3) can be determined by systematic use of Green formulae (integration by parts). As an example, the derivation of the adjoint of the term underlined in eqs. (3.2) is given in Appendix A. After the adjoints of all the individual terms have been taken, we finally obtain for the adjoint of eqs. (3.2) the following equations

$$-\frac{\partial\delta'\zeta}{\partial t} = J(\Delta^{-1}\delta'\zeta,\zeta+f) + \Delta J(\Delta^{-1}\zeta,\Delta^{-1}\delta'\zeta) + \Delta(\nabla\Delta^{-1}\eta.\nabla\Delta^{-1}\delta'\zeta)$$

$$+ \Delta J(\Delta^{-1}\eta,\Delta^{-1}\delta'\eta) - \Delta(\nabla\Delta^{-1}\zeta.\nabla\Delta^{-1}\delta'\eta) + \nabla.(\delta'\eta\nabla\Delta^{-1}\zeta)$$

$$+ \nabla.[(\zeta+f)\nabla\Delta^{-1}\delta'\eta] + J(\delta'\eta,\Delta^{-1}\zeta) - \frac{1}{\Phi_0} J(\delta'\varphi,\varphi)$$

$$-\frac{\partial\delta'\eta}{\partial t} = J[\Delta^{-1}\delta'\eta,\zeta+f] + \nabla.(\delta'\eta\nabla\Delta^{-1}\eta) + J(\Delta^{-1}\zeta,\delta'\eta) \qquad (3.4)$$

$$- \nabla.[(\zeta+f)\nabla\Delta^{-1}\delta'\zeta] + \underline{\frac{1}{\Phi_0}\nabla.(\varphi\nabla\delta'\varphi)}$$

$$-\frac{\partial\delta'\varphi}{\partial t} = \Phi_0\delta'\eta + J(\Delta^{-1}\zeta,\delta'\varphi) + \nabla\Delta^{-1}\eta.\nabla\delta'\varphi$$

These equations are the analogue of eqs. (2.5). The inhomogeneous term $\nabla_x H(t)$, which will depend on the choice of the scalar function , is still ignored for the time being.

Remark The particular inner product (3.3) was originally chosen because the authors thought it desirable to use a "physically significant" inner product. They have since then realized that in the present context, where inner products are not used primarily for estimating "norms" , but for evaluating first order variations, that requirement was meaningless. Indeed, the terms containing $\Phi_0$in eqs. (3.4) are dimensionally incorrect, if Φ_0 is considered as a geopotential and if we remember that the adjoint variables $\delta'\zeta$, $\delta'\eta$, $\delta'\varphi$ are partial derivatives of the distance function with respect to ζ, η, φ. As will be discussed below in Section 9, the numerical consequences of that choice was negligible.

In the numerical experiments described below, eqs. (3.1) have been integrated with a pseudo-spectral model built on the spherical harmonics Y_n^m, with triangular truncation at some degree N. Since observations over the Northern Hemisphere only were used, the model was in effect made hemispherical by retaining only symmetric components (n-m even) for the geopotential and the divergence fields, and antisymmetric components (n-m odd) for the vorticity field. According to the classical procedure used in pseudo-spectral models (Eliasen et al., 1970), the quadratic advection terms in eqs. (3.1) were computed in physical space, on a collocation grid with enough resolution to avoid aliasing errors. It is known (Talagrand and Courtier, 1987, Appendix B) that the property of "adjointness" is conserved by unaliased discretization, in the sense that the adjoint of an unaliased discretized operator is the unaliased discretized form of the adjoint of the original (non discretized) operator. Under these conditions, it suffices, in order to obtain the adjoint of the discretized tangent linear equations (3.2), to replace in (3.4) the various differential operators by their unaliased discretized analogues.

The temporal integration of the model uses a leap-frog scheme modified by a semi-implicit treatment of gravity waves (Robert et al., 1972) and a time filter (Asselin, 1972).

Two different truncations, at degrees N=21 and 42 respectively, have been used. For N=21, the timestep was taken equal to 1 hr, and the dimension of the model state vector was equal to 735. For N=42, the timestep was taken equal to 30 minutes, and the state vector dimension was equal to 2793.

In all experiments, the mean geopotential was taken equal to $\Phi_0=10^5$ m^2s^{-2}. This value was chosen in order to obtain a realistic speed of propagation for gravity waves. The same value was used in the inner product (3.3) and in the corresponding adjoint equations (3.4).

4. THE OBSERVATIONS

Two independent series of experiments have been performed, using observations distributed over the Northern Hemisphere and over the 24 hour time period starting at 00:00Z, 18 March 1985 (initial and final times included). The two series of experiments used observations at the 500 mb and 300 mb levels respectively. In each series, all the observations of wind and geopotential available for the corresponding level in the data base of Direction de la Météorologie, Paris, were used. This data base is made up of all the observations transmitted on the Global Telecommunication System (GTS), with rejection of only a few observations which lie outside reasonable climatic ranges (a number of erroneous observations were therefore certainly included in the present experiments and some observations in the data base had effectively been rejected by the operational assimilation scheme of Direction de la Météorologie). At the 500 mb level, 1752 individual observations of the geopotential, and 2048 observations of each of the two components of the wind field, were retained in the data base. Most of the observations had been obtained from radiosonde (TEMP and PILOT) reports, with a few wind observations coming from SATOB and AIREP reports (87 and 3 respectively). In a few experiments, 72 additional estimates of the geopotential, obtained by adding 1000-500 mb SATEM thicknesses to values of the 1000 mb geopotential produced by the operational assimilation scheme of Direction de la Météorologie, have been used together with the observations extracted from the GTS data base. These additional estimates were all obtained for points located over the Atlantic Ocean.

At the 300 mb level, 1729 observations of geopotential and 2899 observations of each of the two components of the wind were used. One third approximately of the wind observations were obtained from SATOB and AIREP reports (435 and 566 respectively), the other observations having been obtained from radiosonde reports.

The geographical distribution of these various observations is shown in Figs. 1 and 2. The observations are concentrated over the middle latitude continental areas, with the exception of the 300 mb wind observations which, because of the SATOB an AIREP data, are relatively more uniformly distributed. The temporal distribution of observations is shown in Fig. 3. It is seen that the observations are concentrated at the synoptic and subsynoptic hours, with the exception of the 300 mb wind observations which, again, are relatively more uniformly distributed.

It must be noted that, at both the 500 mb and the 300 mb levels, the total number of individual scalar observations (5848 and 7527 respectively) was much larger than the dimension of the state vector of the model, either in its N=21 or in its N=42 version. The

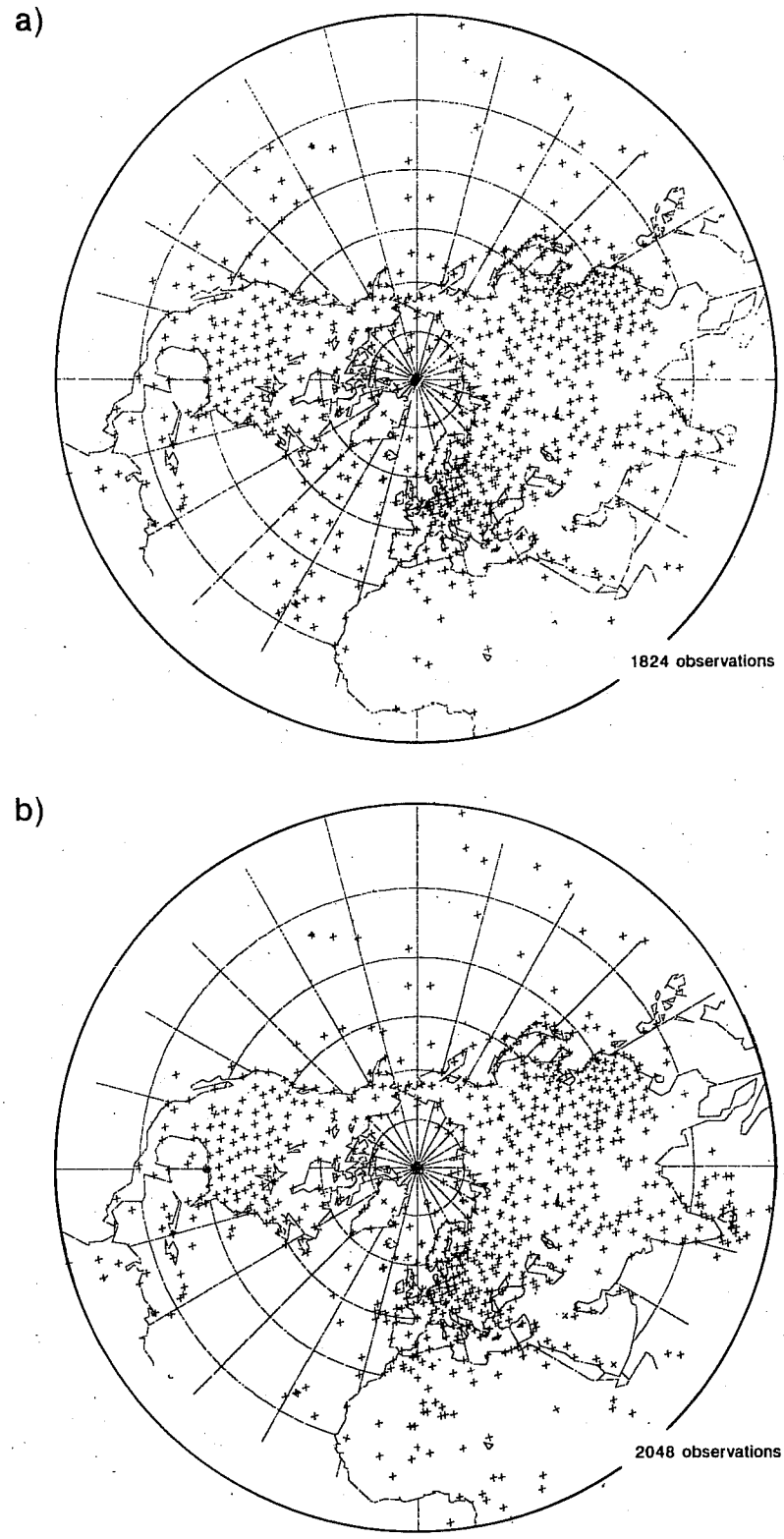

Fig. 1 Geographical distribution of the 500 mb geopotential and wind observations used in the assimilation experiments (panels a and b respectively). Most of the geopotential observations over the Atlantic Ocean have been obtained from SATEM thicknesses.

problem of fitting the model to the observations was therefore overdetermined, and no model solution could be expected to be exactly compatible with the observations.

Figs. 4 and 5 show respectively the 500 mb and the 300 mb geopotential fields at 00:00Z, 19 March 1985 (i.e. at the end of the assimilation period), as produced by the assimilation system of the operational EMERAUDE forecasting programme of Direction de la Météorologie. This assimilation system, which is described in detail in Pailleux et al. (1982), uses a multivariate three-dimensional optimal interpolation scheme and a three-dimensional primitive equation model, with 15 levels in the vertical and spectral triangular truncation at degree 79 in the horizontal. For easier comparison with the results to be presented below, the fields in Figs. 4 and 5 have been truncated at degree 21. These fields are certainly erroneous in several respects, and in particular in planetary scales, as is at present the case for most assimilation systems (Cats et al., 1986). However, Figs. 4 and 5, having been produced with a primitive equation model from a much larger (in particular three-dimensional) set of observations, are useful references for the evaluation of the results to be presented below.

5. THE NUMERICAL EXPERIMENTS

5.1 The minimization process

All experiments consisted in minimizations performed in the space of the model state at the initial time t_0 of the assimilation period (00:00Z, 18 March 1985). Except mention to the contrary, all minimizations were started from the EMERAUDE analysis at that initial time. The minimizations used an algorithm of the quasi-Newton type, described by Buckley and Lenir (1983) (we will explain in Section 9 what a quasi-Newton algorithm is). In all experiments the minimization was interrupted after 30 computations of the gradient at time t_0 had been performed.

5.2 Assimilation without balance constraint

In a first class of experiment, the distance function was defined as a simple weighted sum of squared differences between the observations and the corresponding model values

$$= \frac{1}{\sigma_\varphi^2} \sum_1^{N_\varphi} [\varphi-\varphi_0-\bar{\varphi}]^2 \quad + \quad \frac{1}{\sigma_v^2} \sum_1^{N_v} [(u-u_0)^2 + (v-v_0)^2] \tag{5.1a}$$

$$\text{with } \bar{\varphi} = \frac{1}{N_\varphi} \sum_1^{N_\varphi} (\varphi-\varphi_0) \tag{5.1b}$$

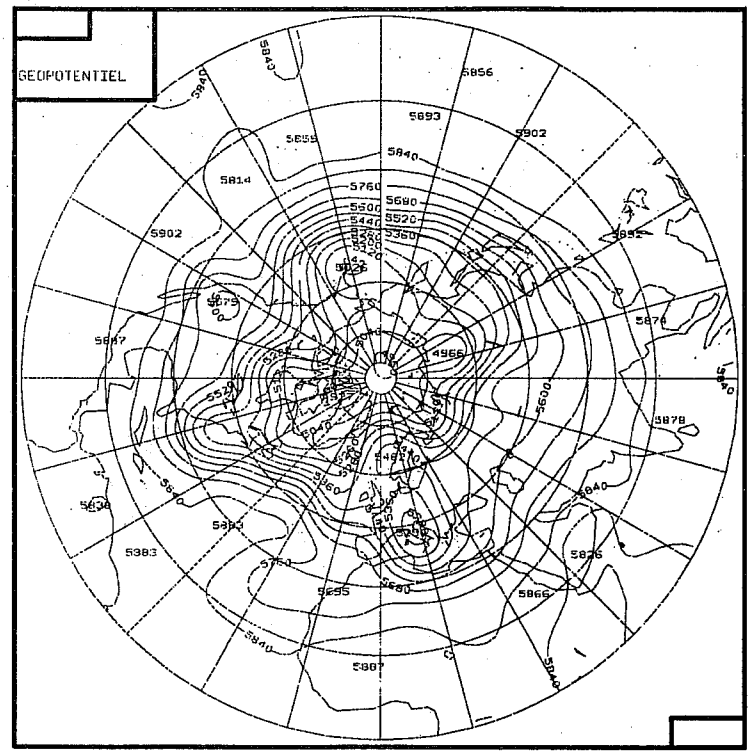

Fig. 4 500 mb height field for 00:00Z, 1 March 1985, as produced by the operational assimilation and forecast EMERAUDE system of Direction de la Météorologie, Paris (unit:metre).

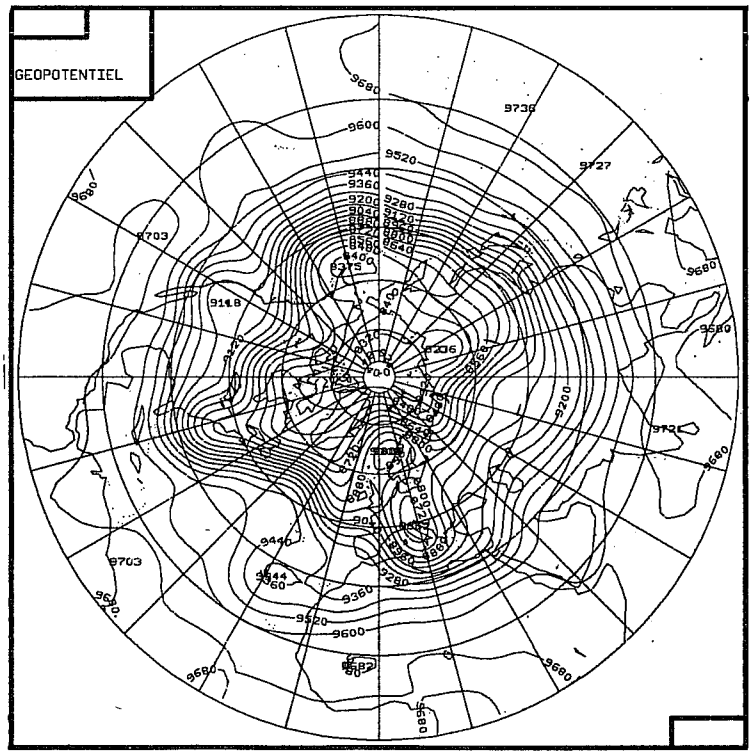

Fig. 5 As Fig. 4, but for the 300 mb level.

where N_v is the total number of wind vector observations available over the assimilation period (t_0,t_1), and N_φ is the total number of geopotential observations. The quantities u_o, v_o and φ_o are the observed values for northward wind component, eastward wind component and geopotential respectively, while the quantities u, v, φ are the corresponding model values. The latter were obtained from the model values of the wind components and geopotential on the collocation grid in physical space through bilinear interpolation with respect to latitude and longitude. It would have been possible, and would have arguably been more consistent with the spectral character of the model, to directly compute u, v, φ at the observation locations from the spectral components of the model fields. But, as already mentioned in CT, this would have forbidden the use of Fast Fourier Transforms. Moreover, because of the high resolution of the collocation grid, the corresponding differences on the estimated u, v, φ would presumably have been small.

In (5.1a), σ_φ and σ_v are estimates of the statistical root-mean-square observational errors on geopotential and wind components respectively. The values σ_φ = 110 m^2s^{-2} and σ_φ = 155 m^2s^{-2} have been taken at the 500 mb and 300 mb levels respectively, and the value σ_v = 2.7 ms^{-1} has been taken at both levels.

The term (5.1b) has been included in (5.1a) in order to obtain for an expression which was independent of the spatial average of the model geopotential field. As in CT, where a similar term was introduced, the basic reason for the presence of this term is of course that geopotential is only defined up to an additive constant. In the present case, the mean geopotential Φ_o = 10^5 m^2s^{-2} had been chosen, as noted above, in order to obtain a realistic phase velocity for the gravity waves and was therefore not allowed to vary in the minimization process. But this value is inconsistent with observed values of geopotential, at least if these values are estimated, as usual, with respect to sea surface level. It was therefore necessary to define for an expression which was independent of the mean geopotential. Another possibility would have been to introduce in a fictitious mean geopotential, independent from the dynamics, to be determined, together with the other components of the geopotential field, by the minimization process. But it is easy to verify that the final minimizing value of that mean geopotential would precisely have been equal to $-\bar{\varphi}$, so that the final result would have been exactly the same as when directly minimizing (5.1a). It is the value $-\bar{\varphi}$ which has been used as mean geopotential for plotting the height charts, produced by variational assimilation, which will be presented below.

Fig. 6 shows the height field produced at the final time of the assimilation period, when assimilating 500 mb observations through minimization of (5.1a). The root-mean-square fits (per individual observation) of the minimizing model solution to the observations are e_φ = 20.4 m for height observations, e_u = 6.1 ms^{-1} for u-component wind observations, and e_v = 5.5 ms^{-1} for v-component observations. The corresponding values at the beginning of the descent process (i.e. for the shallow-water model solution obtained from the EMERAUDE analysis at 00:00Z, 18 March 1985) are respectively e_φ =41.2 m, e_u = 6.8 ms^{-1} and e_v = 6.1 ms^{-1}. The temporal variations of the instantaneous root-mean-square fits over the assimilation period are shown in Fig. 7 for both solutions. It is seen that the minimization has not only decreased, on all three components u, v, φ, the fit of the model to the observations, but has produced final instantaneous fits which are much more uniform in time. This clearly shows the capability of assimilation to adjust a model to a set of observations distributed in time.

Comparison between Figs. 6 and 4 shows that the main features of the mid-latitude circulation (depression over Western Europe and ridge to the west of that depression, depression over Southern Greenland, trough over Eastern North America, depressions off the Western Coast of North America and in the Aleutian area) are the same in both charts, but with somewhat different positions and intensities, in particular for the Aleutian depression. But the main difference between the two charts is the presence, in Fig. 6, of numerous noisy structures in the low latitudes. These structures, which are mostly concentrated in the smallest scales resolved by the model, are very similar to the small scale noise already observed in CT in data-void areas. In the present case, these structures turn out to consist mostly of gravity waves. The descent algorithm uses all the degrees of freedom of the model in order to minimize the distance function, and puts in the fields as much gravity waves as necessary to reach the minimum. A clear indication to this effect is given by the variations of the squared norm $\| \frac{dG}{dt} \|^2$ (estimated according to (3.3)) of the time derivative of the gravity component G of the flow. This quantity is known to give a fair measure of the ageostrophic character of the flow. Evaluated at time t_o, it increases in the present case from 0.47 $m^4\ s^{-6}$ to 1.63 $m^4\ s^{-6}$ between the beginning and the end of the minimization.

5.3 Assimilation with a balance constraint

Two methods have been used, first independently, and then in combination, in order to reduce the amount of gravity waves in the assimilated fields, and to impose an appropriate balance between the mass and velocity fields. Both methods follow the logic of the nonlinear normal mode initialization procedure of Machenhauer (1977) and tend to reduce to 0 the time derivative of the gravity wave component of the flow

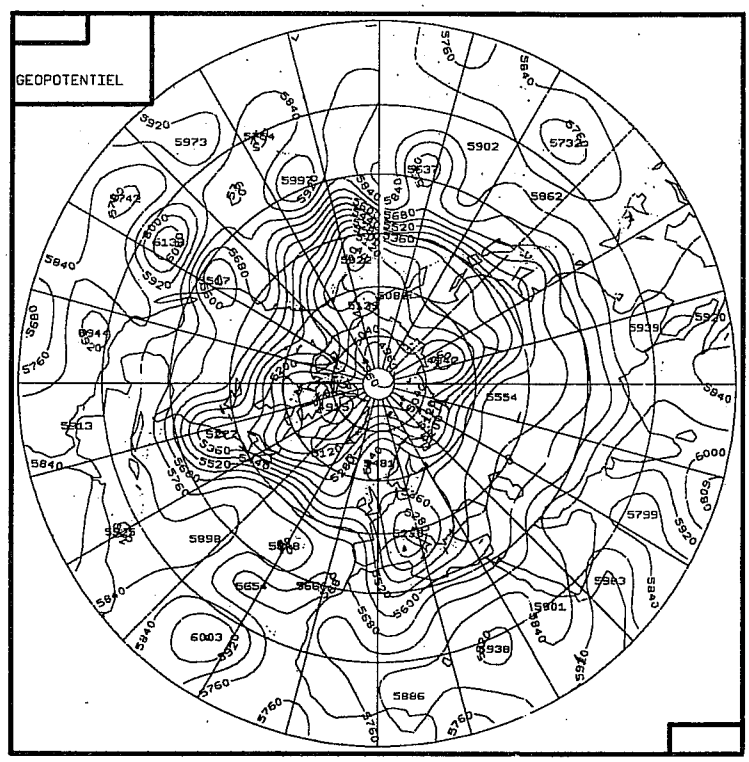

Fig. 6 500 mb height field produced at the end of the assimilation period (00:00Z, 19 March 1985) by minimization of the distance function (5.1).

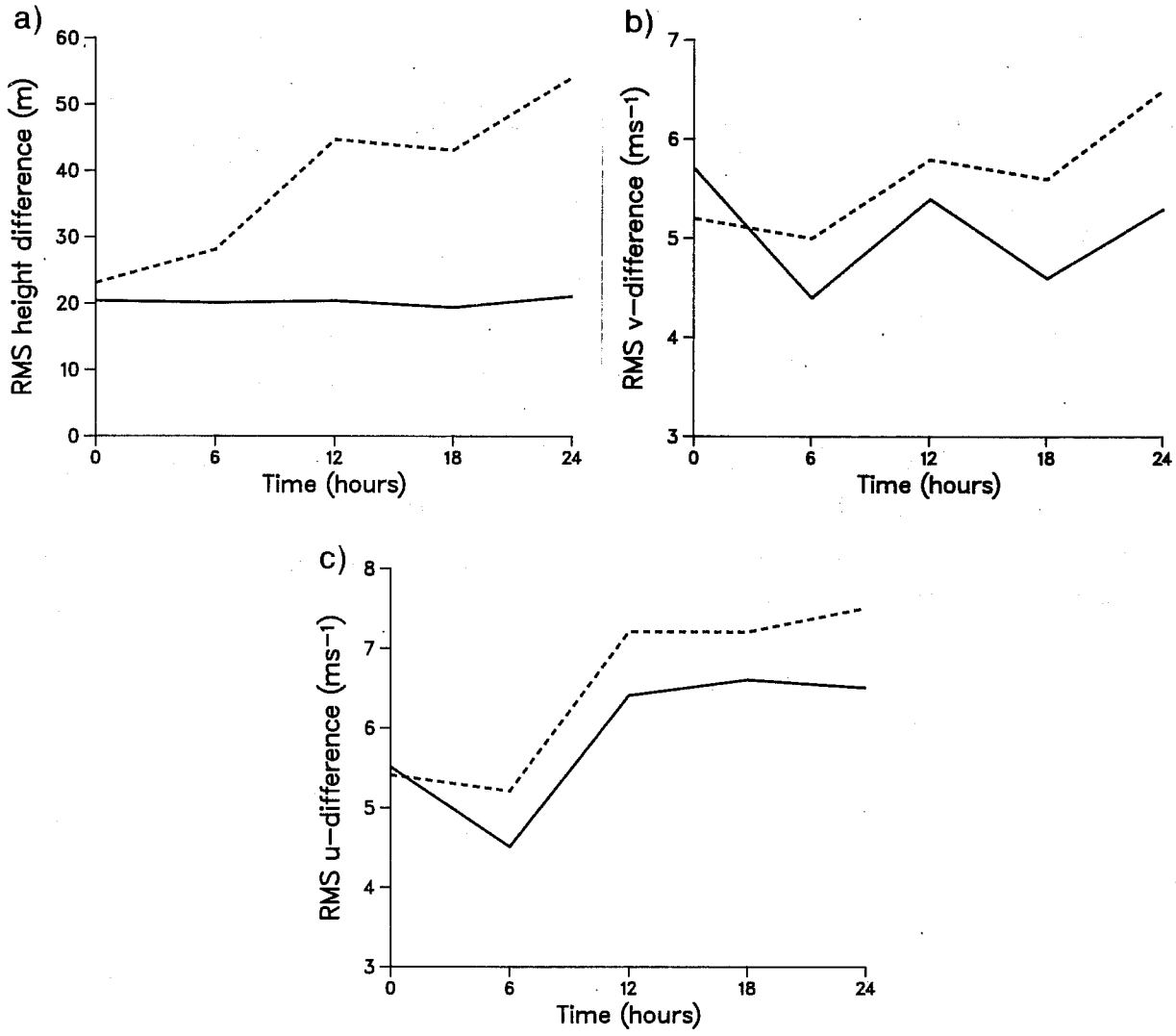

Fig. 7 Temporal variations, over the assimilation period (t_0, t_1), of the root-mean-square differences between the observations and (i) the model solution produced from the EMERAUDE analysis at time t_0 (dashed curves) (ii) the model solution minimizing the distance function (5.1) (full curves). Panel a: height. Panel b: northward (v) component of the wind. Panel c: eastward (u) component of the wind.

$$\| \frac{dG}{dt} \| = 0 \qquad (5.2)$$

The first of the two methods can be described, in the now classical terminology introduced by Sasaki (1970), as enforcing condition (5.2) as a "weak constraint", while the second is intended at enforcing (5.2) as a "strong constraint".

The subset of the phase space made up of the points which satisfy condition (5.2) is known, as first shown by Machenhauer (1977), to be practically invariant by integration of the equations governing the atmospheric flow. This subset will be called here the "slow manifold". The expression slow manifold was first introduced by Leith (1980); we stress that our slow manifold is not identical with Leith's, but is what he called the first approximation **M** to his slow manifold. Fig. 8 is a is a well-known schematic (and two-dimensional) representation of the phase space **E** and of the slow manifold, which will be denoted by **S**. Each point of **E** is defined in this representation by its projections R and G onto the subspaces spanned by the Rossby modes and gravity modes respectively.

In the first of the two methods used for limiting the amount of gravity waves, a penalty term of the form

$$P = \alpha_G \| \frac{dG}{dt} \|^2 \qquad (5.3)$$

where α_G is a positive numerical coefficient, has been added to the distance function . Implementation of the minimization then requires the explicit computation of the gradient of P with respect to the model fields: this gradient contributes to the term $\nabla_x H(t)$ in eq. (2.5). No practically usable analytical expression can be obtained for this gradient, but it can be numerically determined, following the general logic of adjoint equations, by using the adjoint of the numerical process which, starting from the model fields at a given time, leads to $\| \frac{dG}{dt} \|^2$ at the same time. That process basically consists of one timestep of the model and, accordingly, its adjoint basically consists of one timestep of the adjoint model.

With the penalty term (5.3) included in the distance function, the minimization process converged to a solution which looked acceptable in terms of the amount of gravity waves it contained, but the convergence turned out to be extremely slow. The reason for that is that because of the presence of the penalty term, the distance function is very sensitive to the presence of possible gravity waves. The iso- surfaces are accordingly extremely elongated in the phase space in the R-direction (curve Γ' of Fig. 9). It is easy to realize

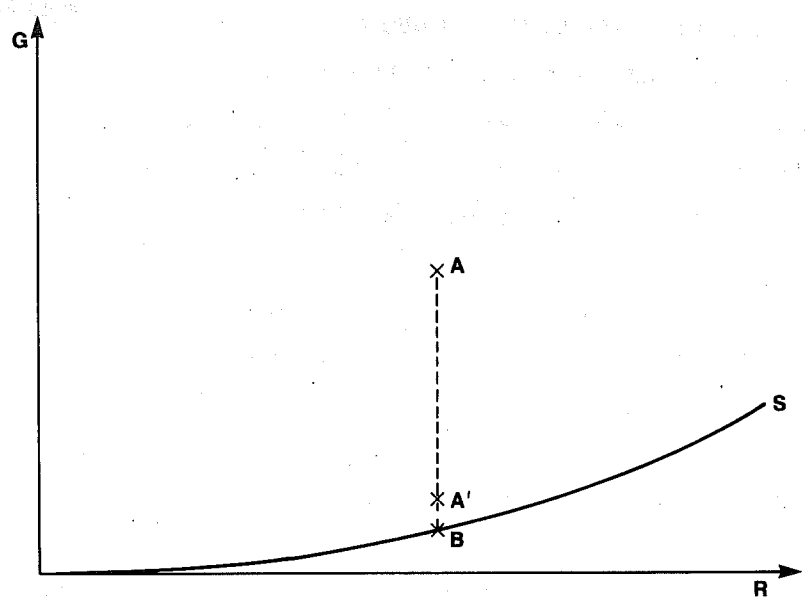

Fig. 8 Schematic representation of the phase space **E**, extracted from Leith (1980). Each point in **E** is defined by its projections **R** and **G** onto the subspaces spanned by the Rossby modes and gravity modes respectively. **S** is the slow manifold, i.e. the subset of points satisfying condition (5.2). Starting from a given point A, the Machenhauer nonlinear normal mode initialization algorithm would lead, in the limit of infinitely many iterations, to point B, located on the slow manifold. Because the algorithm must in practice be interrupted after a finite number of iterations, it only leads to point A'.

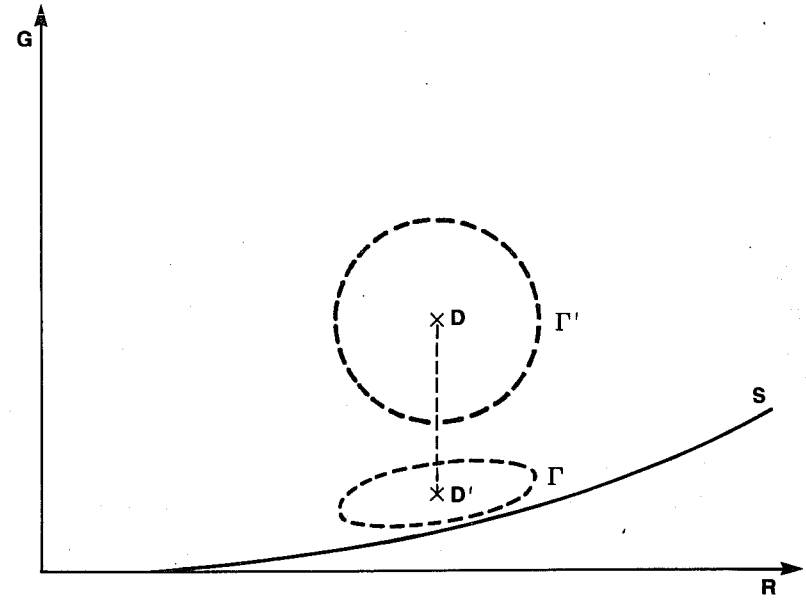

Fig. 9 Same schematic representation of the phase-space **E** as in Fig. 8. Points D and D' correspond respectively to the "uninitialized" and "initialized" model states leading to the solution which minimizes the distance function penalized by term (5.3). D' is the image f D through the Machenhauer algorithm. Similarly, Γ and Γ' are level curves for the penalized distance function in the spaces of "uninitialized" and "initialized" model states respectively. Γ' is the image of Γ through the Machenhauer algorithm.

that this situation leads to ill-numerical conditioning of the minimization and to slow convergence. Although the unfavourable consequences of such a situation are alleviated to some extent by the use of a quasi-Newton algorithm, which basically modifies the metric defined by (3.3) in order to make the iso- surfaces more spherical, they were in the present case severe enough to lead to an unacceptably slow convergence.

Remark Numerical tests have shown that the results which have just been described are insensitive to the particular instant of the assimilation period (t_0, t_1) at which the time derivative $\frac{dG}{dt}$ is evaluated in (5.3).

In the second method, no penalty term was added to the distance function, but a Machenhauer nonlinear normal mode initialization, intended at exactly enforcing condition (5.2), was introduced at each step of the minimization before the integration of the model, and it is the fields before the initialization which were taken as control variables with respect to which the minimization was performed. We recall here that the Machenhauer algorithm is an iterative algorithm for projection of the phase space E onto the slow manifold. This projection is performed at constant Rossby wave component R, so that, starting from an uninitialized state A (Fig. 8), the Machenhauer algorithm will lead, in the limit of infinitely many iterations, to the initialized state B. Now, the algorithm must in practice be interrupted after a finite number of iterations. This leads to a point A', which does not exactly lie on the slow manifold. Machenhauer's algorithm does not therefore define an exact projection of the phase space onto the slow manifold, but only a contraction, along the G-direction, towards the slow manifold. And, as such, the algorithm is invertible in the sense that to any point A' in E there corresponds a point A (which may correspond to unrealistic physical fields) whose image by the algorithm is A'.

Once the uninitialized fields are taken as control variables, it is necessary, in order to determine the gradient of the distance function with respect to these new control variables, to apply the adjoint of the initialization algorithm on the result $\delta'x(t_0)$ of the adjoint integration (2.5). Each step of the initialization algorithm essentially consists of one timestep of the basic model. Therefore, in a way very similar to what has been said above about the gradient of the penalty term (5.3), the adjoint of the initialization algorithm is an iterative algorithm, each step of which essentially consists of one timestep of the adjoint model.

With this second method, the minimization first apparently converged to a solution which was free of gravity waves, but after a number of descent steps (about 15 when the initialization contained two iterations), the amount of gravity waves started increasing and the minimization finally converged to the same model solution over (t_0,t_1) as when no

balance constraint was imposed. The reason for that result is simple: If in Fig. 8, A' is the point corresponding to the initial condition minimizing the non-penalized distance function (5.1), minimisation with respect to "initialized" fields will lead to point A whose image by the initialization algorithm is A', so that the final model solution will be the same as before. This means that the minimization algorithm is in fact capable of inverting the initialization. Increasing the number of iterations in the initialization algorithm has only the effect of delaying the eventual convergence to C, but does not significantly change the final results.

It therefore appears that, owing to the iterative and approximate character of the initialization algorithm, condition (5.2) cannot in practice be enforced as an exact constraint. It is by combining together the two methods described above that properly balanced conditions have been obtained in an acceptable number of descent steps. Inclusion of the penalty term in the distance function ensures that the minimizing solution will not contain an unacceptable amount of gravity waves, while performing the minimization on uninitialized fields leads to better numerical conditioning. This is illustrated in Fig. 9. With the penalty term included in the distance function, the iso- surfaces, as already said, are strongly elongated in the R-direction. The corresponding surfaces in the space of uninitialized fields are obtained through the inverse of the initialization algorithm. The latter being a contraction in the G-direction towards the slow manifold, its inverse is a dilatation in the G-direction. The iso- surfaces turn out to be more spherical in the space of uninitialized fields than in the space of initialized fields. This ensures more rapid convergence.

Fig. 10 shows the height field thus obtained at time t_1. The coefficient α_G has been taken equal to .19 x 10^4 m^{-4} s^6. The unrealistic small scale features of Fig.6 are no more visible in Fig. 10, while the meteorologically significant structures of the middle latitudes are still present. Indeed, some of these middle latitude structures, such as for instance the Aleutian depression, are now closer to the EMERAUDE analysis (Fig. 4) than they were in Fig. 6. This suggests that, because of the additional link imposed by the penalty term between the mass and velocity fields, more of the information contained in each of these two fields has effectively been used in the reconstruction of the other.

A measure of the efficiency of the constraint imposed by (5.3) is given by the corresponding variations of $\| \frac{dG}{dt} \|^2$ which (estimated after the nonlinear normal mode initialization has been performed), decreases from 0.47 10^{-2} m^4 s^{-6} to 0.47 10^{-5} m^4 s^{-6} in the course of the minimization. At the end of the minimization, the ratio of the penalty term to the "pure" distance term (5.1) is equal to 6. 10^{-3}.

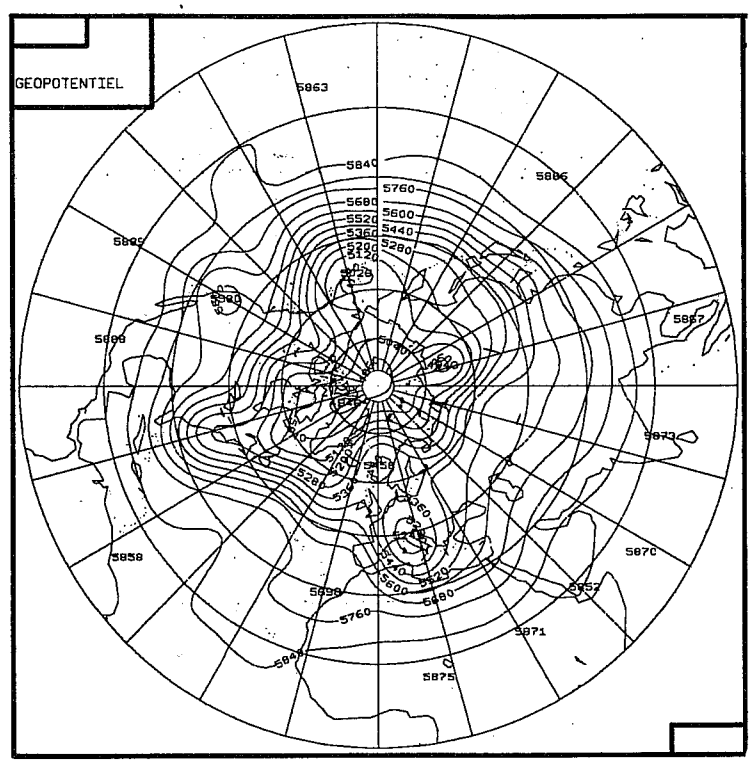

Fig. 10 As Fig. 6, but for a distance function penalized with term (5.3).

As for the fit of the minimizing solution to the observations, it is now 26.5 m per individual height observation, and 6.2 ms^{-1} and 5.6 ms^{-1} per individual u and v wind observation component respectively. These values are larger than the values obtained when no penalty term was included in the distance-function (20.4 m, 6.1 ms^{-1} and 5.5 ms^{-1} respectively). This is normal, since the global fit to the observations is necessarily degraded by the presence of the penalty term. But it is interesting to note that the change is mostly in the fit to the height observations. This is in agreement with the linear theory of geostrophic adjustment, which predicts that in most scales resolved by the model, and especially at low latitudes, it is the mass field which must adjust to the velocity field.

Combining a penalty term (5.3) and a Machenhauer integration algorithm is therefore a successful way for obtaining a properly "balanced" model solution in a relatively small number of descent steps. It is noteworthy that it is only the penalty term (5.3) which imposes the balance constraint. The Machenhauer algorithm is in effect only a way to perform a change of variable which speeds up the convergence of the minimization. There is of course no special reason to think that the particular (and rather arbitrary) procedure used here is the best way for imposing a proper balance on the assimilated fields. But the very fact that present initialization techniques can in effect be simply and successfully introduced in a variational assimilation is of extremely great interest. Indeed, it suggests that the general problem considered by Daley (1978), of finding the point on the slow manifold lying closest to a given analysis, must be solvable in the broader context of an assimilation including the temporal dimension.

Lorenc (1988a) has also found, in experiments performed with a one-dimensional shallow-water model, that adding an appropriate penalty term to the distance function could effectively constrain the solution of a variational assimilation to be "slow". He does not mention however any problem of ill-conditioning arising from the presence of the penalty term.

6. ADDITIONAL RESULTS

6.1 The effect of temporal weighting on the observations

The results presented so far were obtained with a distance function in which the weights given to the observations were independent of time. The difference between the final minimizing solution and the corresponding EMERAUDE analyses shows some tendency to take smaller values at intermediate times of the assimilation period (dashed curves of Fig. 11). This effect, which is made visible on the rotational wind difference is easily understandable. The model, not being perfect, cannot adjust uniformly to the whole set of observations, and minimizes the distance function by adjusting preferentially to observations at intermediate times. However, in the case of assimilation, where one wants to make a forecast from the final time of the assimilation period, better adjustment to later

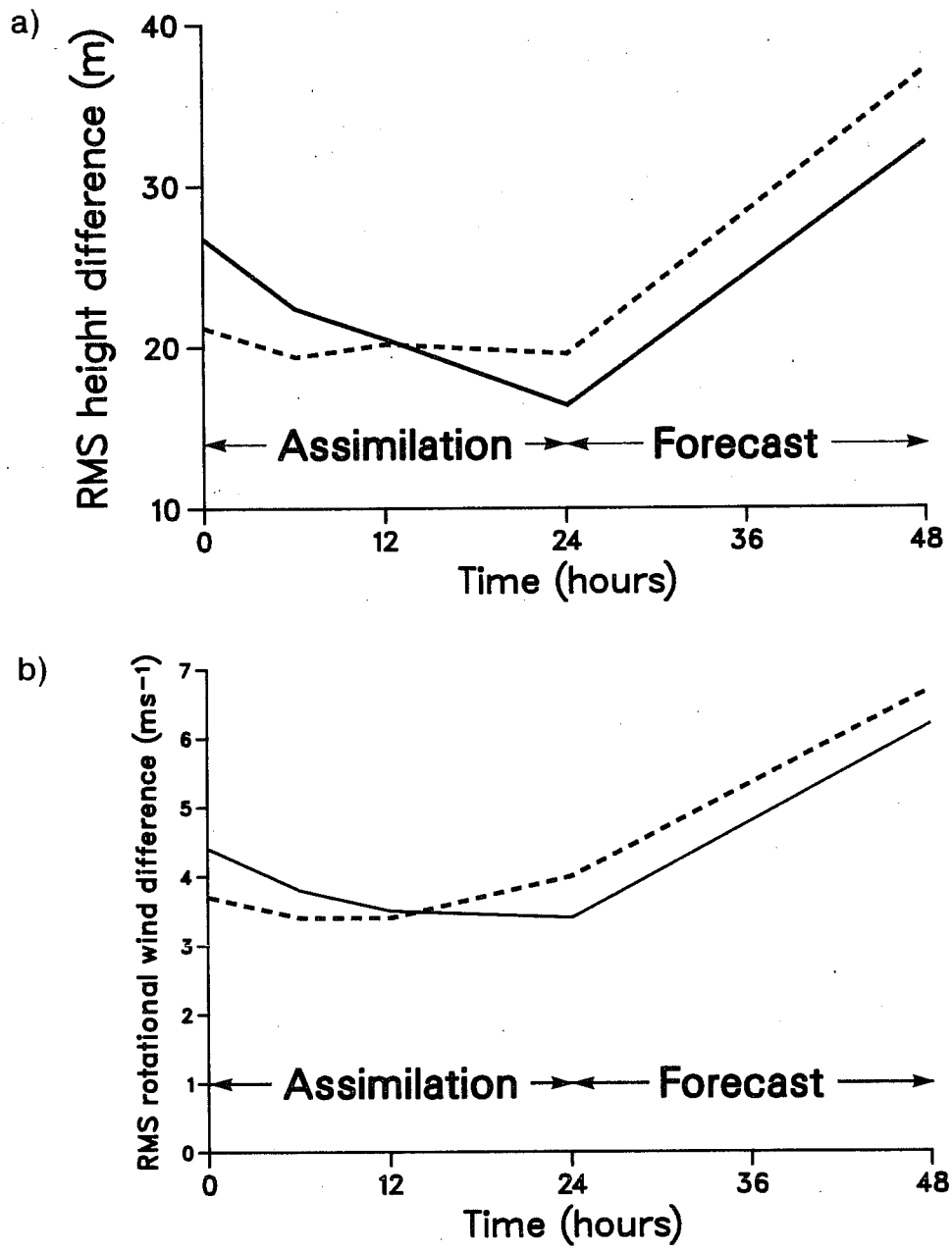

Fig. 11 a) Temporal variation, over the assimilation period (from 0 to 24 hours) and over a subsequent 24 hour forecast (from 24 to 48 hours), of the root-mean-square difference between (i) the height fields of a model solution produced by variational assimilation and (ii) the corresponding 500 mb height fields of the operational EMERAUDE analyses. Dashed curves: assimilation with constant temporal weighting of the observations. Full curves: assimilation with variable weighting of the observations. b) As a) but for rotational wind root-mean-square difference.

observations is obviously preferable, and one must give larger weights to more recent observations in the definition of the distance function. The problem of determining the appropriate temporal variation of the weights, knowing the statistical performance of the model, does not appear simple. Useful ideas can probably be drawn in that respect from the theory of Kalman filtering (Ghil et al., 1981), which provides a systematic approach for dealing with model errors. In the present case, an experiment has been performed in which weights given to individual observations varied linearly with time in the ratio 3 to 80 between the beginning and the end of the assimilation period (the total sum of the weights was kept at the same value as in the previous experiments in order not to modify the relative importance of the gravity wave penalty term).

The full curves of Fig. 11 show the temporal variation of the final adjustment of the model to the EMERAUDE analyses. The adjustment is now closest to the latest observations. An objective proof that variable weighting of the observations has improved the quality of the assimilation is given by the root-mean-square differences between 24h-forecasts produced from the variational assimilations and the corresponding EMERAUDE analyses, also visible in Fig. 12. It is seen that variable weighting of the observations not only improves the fit of the assimilation to the analyses at the end of the assimilation period, but also leads to a slower growth rate of the forecast error.

6.2 The impact of additional SATEM observations

As already mentioned in Section 4, 72 estimates of the 500 mb geopotential have been used in some experiments in addition to the observations contained in the data base of Direction de la Météorologie. These additional "pseudo-observations", obtained by adding SATEM 1000-500 mb thickness measurements to 1000 mb heights produced by the EMERAUDE assimilation, were located over the Northern Atlantic (see Fig. 1a) and were valid at either 21:00Z or 22:00Z, 18 March 1985. Their impact on the assimilation can be seen in Fig. 12, which is in the same format as Fig. 11, and is relative to two assimilations (with variable weighting of the observations) performed with and without the pseudo-observations.

The impact of the additional geopotential estimates, although small, is clearly visible: on both the height and the rotational wind fields, their presence decreases the difference with the EMERAUDE analysis, by an amount which is maintained in the forecast. The fact that the difference is decreased in both fields is a new proof of the "multivariate" character, through the model dynamics and the balance condition, of the assimilation. (It can also be noted that the impact of the additional observations, through global adjustment of the model over the whole assimilation period, is already visible, at the rather low accuracy of the figure, at 12:00Z, i.e. 10 hours before the time of the measurement).

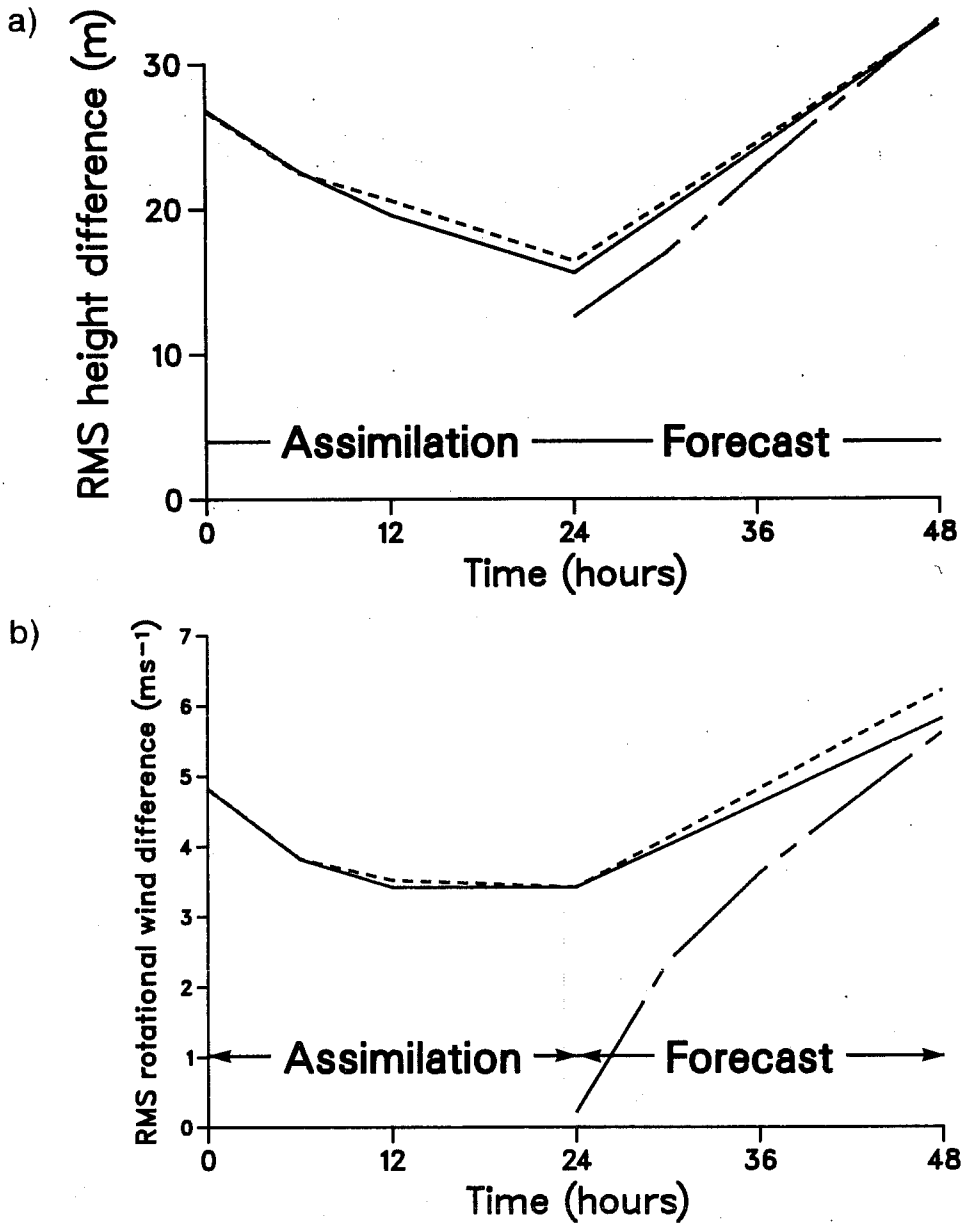

Fig. 12 As Fig. 11, but for assimilations without SATEM observations (dashed curves, identical with full curves of Fig. 11) and with SATEM observations (full curves). The dash-dotted curves show the time variation of the root-mean-square difference between (uninitialized) EMERAUDE analyses and a shallow-water forecast produced from the (initialized) EMERAUDE analysis at the end of the assimilation period. Panels a and b: height and rotational wind root-mean-square differences respectively.

6.3 Forecast quality

The dashed-dotted curves of Fig. 12 show the growth of the root-mean-square error in a 24 hour shallow-water equation forecast produced from the EMERAUDE analysis at the final time of the assimilation period. (Although EMERAUDE analyses are here taken as the "true" reference the initial root-mean-square error is not zero, especially in Fig. 12a, because the forecast is started from an initialized analysis while it is the uninitialized analyses which are used as references). It is seen that, in both figures, the growth of the error is smaller for forecasts produced from the results of variational assimilations than for the forecasts produced from the EMERAUDE analysis. If one remembers that the EMERAUDE analysis was the outcome of an assimilation process which had used a a three-dimensional primitive equation model, together with a three-dimensional set of observations, much larger than the set used in the variational assimilations, the results of Fig. 12 constitute a strongly positive indication as to the quality of variational assimilation.

7. ASSIMILATION OF OBSERVATIONS AT THE 300 MB-LEVEL

Assimilations of observations at the 300 mb level have been performed in a way very similar to assimilations at the 500 mb level. In particular, the temporal weighting of the observations and the procedure for imposing an appropriate balance in the assimilated fields (penalty term and Machenhauer algorithm) were the same. The only difference was, as said in Section 5, that the root-mean-square observational error on geopotential was now taken equal to 155 m^2s^{-2}, instead of 110 m^2s^{-2}. This resulted in a lower relative weighting of the geopotential observations. The results, which are described in more detail in Courtier (1987), are very similar to those obtained at 500 mb. The main difference is that the fit to the observations and to the EMERAUDE analyses is now much less accurate: the root-mean-square to the EMERAUDE analyses is roughly multiplied by 2, as can be seen by comparing Figs. 12 and 13. Another difference is that, even though the same temporal weighting was used, the variations of the fit to the EMERAUDE analyses is now rather different: for height, the fit now has a weekly marked minimum at 24:00Z, while for rotational wind, the fit is now closest at 12:00Z (Fig. 13a and b). All this is perfectly consistent with the idea that the shallow-water equations are a poorer model of the circulation at 300 mb than at 500 mb. In particular, it would be necessary, if one wanted for some reason to obtain the closest fit at 24:00Z, to give still a stronger relative weighting to the latest observations. As for the increase of the rotational wind difference between 12:00Z and 24:00Z, it is to be linked to the fact that wind observations at 300 mb are relatively more numerous at low latitudes (Fig. 2b), where a shallow-water model is certainly a very poor descriptor of the flow.

8. ASSIMILATION AT HIGH RESOLUTION

The assimilations performed with the model truncated at degree N=42 have produced results which are on the whole very similar to the result obtained with truncation at degree 21. No

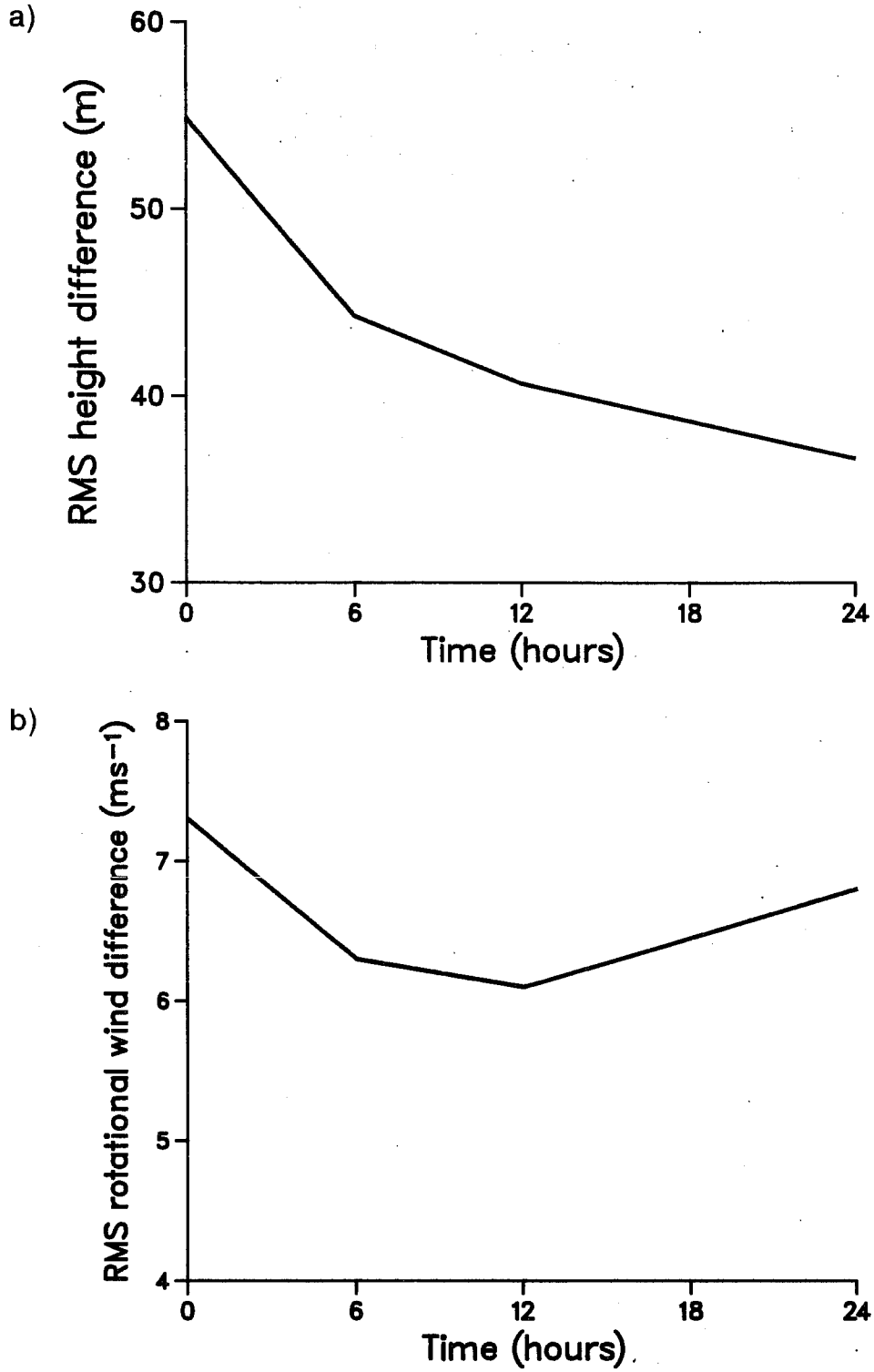

Fig. 13 As Fig. 12, but for the assimilation period only, and for assimilation of observations at the 300 mb level.

difference was observed in the assimilated fields which could be considered as meteorologically significant. The authors interpret this as meaning that the meteorological qualities of a shallow-water equation model are already present with truncation at degree 21, and that no gain can in effect be achieved by increasing the resolution. Another noteworthy fact is that the rate of convergence of the minimization, as measured by the number of descent steps necessary to reach convergence, was not modified by the change of resolution. A four-fold increase in the numbers of degrees of freedom of the model has therefore no significant impact on the efficiency of the minimization.

9. NUMERICAL ASPECTS

We will now discuss some of the numerical aspects of the experiments described in the previous sections. An important question is raised by the possibility that the distance function might have several distinct minima. As in CT, no indication was observed that this could be the case. More precisely, variations of the distance function along straight lines in phase space have been studied. As in CT (see Fig. 9 therein), these variations turn out to be parabolic. This is a strong indication not only that the distance function has only one minimum, but also that it varies quadratically with respect to the model initial state at time t_0, at least within the domain explored by the minimization process. This in turn is in agreement with results obtained with another shallow-water equation model by Lacarra and Talagrand (1988) (see also Urban, 1985), who have found that, over a 24 hour or 48 hour period and for initial perturbations whose amplitude is comparable to the uncertainty with which the state of the atmosphere is known in practice, the temporal evolution of a perturbation is in effect governed by the corresponding tangent linear equation. If this result is confirmed for more complicated models, such as multilevel primitive equation models, it will mean that, in the context of data assimilation, atmospheric dynamics can to a high degree of accuracy be considered as linear. The appropriate linear dynamics will of course depend on the current state of the flow, but the knowledge that the dynamics is linear will nevertheless be extremely useful for implementing assimilation schemes and interpreting the results they produce. It will also be extremely useful for assigning confidence limits to short-range numerical forecasts.

In all experiments described above, the descent algorithm was interrupted after 30 computations of the gradient of the distance function. That value was in fact larger than necessary, and a satisfactory convergence, as judged by all reasonable standards and in particular by the fact that the changes caused in the fields by additional descent steps were meteorologically negligible, was usually reached after 10 to 15 descent steps. We stress that these values were the same for both low resolution and high resolution assimilations, although the latter was performed in a phase space whose dimension was four times as large. In CT, 10 to 15 iterations were also necessary to obtain satisfactory convergence, in a phase space with dimension 231. This suggests that the number of

descent steps in a minimization is almost insensitive to the dimension of the phase space, a fact which certainly contradicts what could be a priori expected and which can be of extremely great importance for the implementation of variational assimilation with large dimension models. Determining why the number of descent steps is insensitive to the model dimension would be of extremely great interest.

In most experiments, the descent was started, as already mentioned above, from the EMERAUDE analysis at time t_o. In a few experiments, it was started from a state of uniform geopotential and zero velocity. It then lagged behind the descent started from the EMERAUDE analysis by typically 2 or 3 descent steps. Compared to the 10 to 15 steps necessary to reach convergence, that lag is rather small, and suggests that not much gain could be achieved in practice by starting the descent from a state which is known to be close to the required minimum (for instance, the previous day's analysis). Further work will be necessary on this point and on the question of how much accuracy is in practice necessary on the localization of the minimum.

It has already been said that the descent algorithm used in the experiments described in this article was of the quasi-Newton type. We will now briefly explain what a quasi-Newton algorithm is. We first recall that a descent algorithm determines the value of the control variable which minimizes the distance function as the limit of a recursive sequence (u_p) of the form:

$$u_{p+1} = u_p - \rho_p D_p \tag{9.1}$$

At each descent step p, D_p is the descent direction, determined from the local gradient $\nabla (u_p)$ and (in most descent algorithms) from the previously computed gradients $\nabla (u_{p-1})$, $\nabla (u_{p-2})$, ..., as well as from the previously determined points u_{p-1}, u_{p-2}, ... The scalar ρ_p is determined in such a way as to (approximately) minimize along the descent direction. One descent step therefore consists of two parts:

i) The computation of the gradient $\nabla (u_p)$ and the determination of the descent direction D_p.

ii) The determination of the scalar ρ_p. This second part, called line search, may involve some trial and error experimentation, under the form of explicit determination of the value of the distance function at one or several points along the descent direction.

A general reference on descent algorithm is Gill et al. (1982). Navon and Legler (1987) have given a detailed discussion of the use of descent algorithms in meteorological problems.

Finding a stationary point of the distance function is equivalent to solving the equation

$$\nabla (u) = 0 \tag{9.2}$$

A ready generalization of the classical Newton algorithm for solution of scalar equations leads to the following sequence of approximations to the solution of (9.2)

$$u_{p+1} = u_p - [\nabla^2 (u_p)]^{-1} \nabla (u_p) \tag{9.3}$$

where ∇^2 denotes the matrix of second derivatives, or hessian, of with respect to u. The particular descent algorithm (9.3) is extremely efficient in terms of the rate of convergence per descent step, but is nevertheless extremely costly in that each descent step requires the solution of a linear system whose dimension is the dimension of the phase-space. Quasi-Newton algorithms are algorithms of the general form (9.1) with descent direction

$$D_p = H_p^{-1} \nabla (u_p)$$

where H_p^{-1} is an approximation of the local inverse hessian at point u_p. H_p^{-1} is updated at each step by use of the local gradient. The various quasi-Newton algorithms differ by the updating procedure, and also by the line search procedure (see the above quoted references). Generally speaking, the accuracy of the approximate inverse hessian, as well as the corresponding storage requirements, increase as the descent proceeds. In the Buckley and Lenir (1983) algorithm used for the experiment described in this article, a given amount of core memory is a priori assigned to the storage of the approximate inverse hessian. If that amount happens to be filled in the course of the minimization, the programme then shifts to a less demanding, but also less efficient, conjugate-gradient algorithm. In order to take full advantage of the quasi-Newton phase, the conjugate-gradient algorithm is preconditioned by the last computed approximation of the inverse hessian (preconditioning can be described as a change of metric intended at making the iso- surfaces more spherical).

Tests have been performed in order to determine to which extent the efficiency of the minimization depended on the size of the core memory allocated to the storage of the

approximate hessian. It turns out that doubling the size of the allocated memory does not significantly reduce the number of descent steps necessary to reach convergence, but that each step is on the average more economical. It is so because the quasi-Newton algorithm requires less trial and error computations in the line search than the conjugate-gradient algorithm. On the whole, a 25% relative gain in computing time was achieved when doubling the core memory allocated to the inverse hessian. This result shows that, in agreement with general theory, a quasi-Newton algorithm is more efficient than a conjugate-gradient algorithm. It is also worth mentioning that theory says that the differences between various descent algorithms become negligible for spherical, or nearly spherical, iso-surfaces. Therefore, in spite of the presence of the Machenhauer algorithm the iso-surfaces remained rather strongly elongated for the matrix defined by (3.3).

Concerning (3.3), one can wonder if the use of a large numerical value (10^5) for the coefficient Φ_0 might not have had an unfavourable effect on the minimization, and might not have in particular contributed to the numerical difficulties observed in the presence of the penalty term (5.3). Actually it is easy to see that, in order to make the penalized distance function more spherical, it would be necessary to give a relatively larger weight to the divergence in the definition of the scalar product (Fig. 9). The observed difficulties therefore certainly did not result from assigning too large a value to Φ_0, at least as far as divergence is concerned. Moreover, the effect of a quasi-Newton algorithm can be described as a constant redefinition of the scalar product in order to make the distance function more spherical. The first thing a quasi-Newton algorithm will do in that respect is to renormalize the various components of the state vector in order to compensate for a possible inappropriate relative weighting of these components in the definition of the scalar product. Had the value of Φ_0 been too large, this would have accordingly been corrected in the first steps of the minimization. The difficulties described in Section 5 must therefore result from deeper causes, and from the fact that making the function spherical requires much more than a different weighting of the geopotential, vorticity and divergence, but an appropriate mutual recombination of these fields. As already said, this is achieved to some extent by the Machenhauer initialization algorithm.

10. CONCLUSIONS

We have not discussed in this paper all the aspects of our numerical experiments, but only those aspects which seem most instructive. For instance, as in CT, it was found necessary, in order to avoid the occurrence in the assimilated fields of small-scale noise (not necessarily gravity wave noise), to include an appropriate penalty term in the definition of the distance function. But our conclusion on this respect did not add anything new to what has already been said in CT, and this particular problem has accordingly not been discussed here.

Generally speaking, the conclusions of the present article confirm the conclusions of CT. Variational assimilation does numerically converge, and the model solution to which it converges looks, from a meteorological point of view, quite acceptable. In particular, the main features of the mid-latitude flow are reconstructed to a satisfactory degree of accuracy. Because the model used in the present experiments is still rather simple, these conclusions cannot at this stage be much more precise. But additional conclusions can also be drawn from the results presented above.

i) The quality of the forecasts produced from the result of a variational assimilation seems as good as can be expected from a shallow-water model.

ii) The problem of imposing a proper balance between the mass and velocity fields can be easily and satisfactorily solved by introducing in the variational process an appropriate constraint (based on Machenhauer's condition (5.2) in the present case). This is a new example of the power of variational assimilation and of its adaptability to new situations.

iii) Increasing by an order of magnitude the dimension of the model state vector does not decrease the efficiency of the minimization, as measured by the number of descent steps necessary to reach convergence.

On the whole, the power, the generality and the versatility of variational assimilation are such as to make it a tool whose study and development is of the greatest interest. To that must be added the fact that, once the adjoint of a numerical weather prediction model is available, it can be applied to many other uses than assimilation stricto sensu. Most problems of sensitivity of a small number of output parameters of a model with respect to a large number of input parameters can be studied with the adjoint of the model (see e.g. Hall et al., 1982, or Courtier, 1987). Still other examples of uses of adjoints of numerical models can be found in e.g. Urban (1985), Vautard and Legras (1988) or Lacarra and Talagrand (1988). The development of the adjoint of a numerical weather prediction model is an investment which must in the long term become extremely profitable.

Another great advantage of variational assimilation is that its theoretical bases are perfectly rigorous and sound. This is extremely useful for the design and implementation of an assimilation scheme by providing safe and well-defined guidelines at every stage, and ensuring global consistency of the entire assimilation process. It also helps in anticipating what can or cannot be expected from variational assimilation, and in interpreting and understanding the results it produces. In addition, the links between variational assimilation and other assimilation techniques, such as optimal interpolation and Kalman filtering, are perfectly well understood at the theoretical level (for a reference, see, e.g.

Jazwinski, 1970, although that book does not deal explicitly with meteorological problems). This also is extremely important, in that it will allow to study the various advantages or disadvantages of different methods at the deepest theoretical level, and not only from the results of comparative numerical experiments.

For all these advantages, there certainly remain problems with variational assimilation, especially if one intends to use variational assimilation in operational numerical weather prediction. We will briefly mention and discuss what we think are the most important of these problems.

i) What is the "best" definition of the distance function? In particular, how can the fact that a model will always be imperfect be taken into account in the assimilation?

ii) Is it possible to obtain, in addition to the minimizing solution, estimates of the corresponding uncertainty on the actual state of the atmosphere?

iii) Is it possible to introduce some form of "quality control", i.e. of elimination of erroneous observations, in the assimilation process?

iv) Can the numerical cost of variational assimilation be reduced, and how?

As concerns the first two of these problems, the general result that there exists a variational formulation to any problem of statistical linear estimation is of extremely great importance (see, e.g. Jazwinski, 1970; Wahba, 1982; Lorenc, 1986, 1988a, for various aspects of that general result). In the context of variational assimilation, where statistical linear estimation takes the form of Kalman filtering (Ghil et al., 1981, 1982), that result means, as already mentioned in CT, that the distance function can be defined in such a way as to lead to the same fields at time t_1 as Kalman filtering. Theory of Kalman filtering allows for the explicit introduction of model errors, as described by their first and second order statistical moments. It must therefore be possible to introduce the same information in the definition of the distance function. Research work is being done on this particular problem, and should lead, as already discussed in Section 6, to a variable temporal weighting of the observations.

Statistical linear estimation, and Kalman filtering in particular, produces, in addition to estimated fields, the variance-covariance matrix of the corresponding estimation errors. This matrix, in the variational formulation of statistical estimation, is the inverse hessian of the distance function at the minimum. Quasi-Newton algorithms do compute successive approximations of the inverse hessian. This is one possibility for solving the second of the above mentioned problems, which is also being studied by Thacker (pers. com.). It must be

noted that this will lead to estimation errors at the initial time t_0 of the assimilation period, while it is estimation errors at the final time t_1 which are really useful. It will therefore be necessary to carry the estimation errors from t_0 to t_1. In the context of numerical forecasting, this is not a problem by itself in the sense that estimation errors at time t_1 will in any case be useful only if some means is already available for transforming these errors into estimates of forecast errors at later times. But it clearly shows that the problem of estimating assimilation errors is intrinsically linked to the problem of estimating forecast errors.

The problem of "quality control" has been studied by Lorenc (1988a), who has taken into account the possible presence of "gross" errors in the observations. In the variational statistical formulation of the assimilation, this leads to a distance function which, contrary to (5.1), is not quadratic with respect to the differences between model and observations. Instead, once one of these differences increases beyond a certain value, the corresponding term in the distance function remains constant. Lorenc has found that this modified distance function can have several distinct minima, corresponding to whether individual erroneous observations are "accepted" or "rejected". If this result is confirmed by future experiments, it will mean that a preliminary control of the quality of the observations will probably remain necessary before the variational assimilation itself.

The problem of the numerical cost of variational assimilation is of the utmost importance, especially of course in the context of operational numerical weather prediction. The present number of iterations (10 to 15) necessary to reach convergence would be absolutely prohibitive in today's practice of weather prediction. As already discussed in CT, many possibilities can be thought of for decreasing the computational cost of variational assimilation. In addition the available computing power will continue to increase in the coming year. It will certainly be possible to significantly reduce the cost of variational assimilation, but how exactly, and to what extent that will be possible remain important questions to be answered.

11. ACKNOWLEDGEMENTS

The authors are indebted to Claude Lemaréchal, from Institut National pour la Recherche en Informatique et Automatique (INRIA) who provided them with the computer programmes for implementating descents algorithms. Maxine Simpson found the Rosetta Stone for deciphering the esoteric hand-writing of the authors, and carefully typed the manuscript. The Conseil Scientifique of Centre de Calcul Vectoriel pour la Recherche allocated most of the computing means necessary for the experiments.

APPENDIX A - Adjoint of term - $\nabla.(\varphi\nabla\Delta^{-1}\delta\eta)$ in eqn. (5.2)

The term $-\nabla.(\varphi\nabla\Delta^{-1}\delta\eta)$ in eqn. (5.2) is the output of the operator L which, starting from a perturbation $\delta x = (\delta\xi, \delta\eta, \delta\varphi)^*$ (* meaning here transpose) leads to

$$L\,\delta x = [0, 0, -\nabla.(\varphi\nabla\Delta^{-1}\delta\eta)]^*$$

For any δx and $\delta' x = (\delta'\xi, \delta'\eta, \delta'\varphi)^*$, the inner product of $L\,\delta x$ and $\delta' x$ reads

$$< L\,\delta x, \delta' x > = \frac{1}{2a^2} \int_\Sigma [-\nabla.(\varphi\nabla\Delta^{-1}\delta\eta)]\, \delta'\varphi\, d\Sigma$$

Expression which we want to transform into the inner product of $\delta' x$ with some other factor. Successive uses of green formula lead to

$$< L\,\delta x, \delta' x > = \frac{1}{2a^2} \int_\Sigma \varphi\nabla\Delta^{-1}\delta\eta.\nabla\delta'\varphi\, d\Sigma$$

$$= \frac{1}{2a^2} \int_\Sigma [-\Delta^{-1}\delta\eta\nabla.(\varphi\nabla\delta'\varphi)]\, d\Sigma$$

$$< L\,\delta x, \delta' x > = \frac{1}{2a^2} \int_\Sigma [-\Delta^{-1}\delta\eta\ \nabla\Delta^{-1}\nabla.(\varphi\nabla\delta'\varphi)]\, d\Sigma$$

$$= \frac{1}{2a^2} \int_\Sigma \nabla\Delta^{-1}\delta\eta.\nabla\Delta^{-1}\nabla.(\varphi\nabla\delta'\varphi)\, d\Sigma$$

$$= \frac{\varphi_o}{2a^2} \int_\Sigma \nabla\Delta^{-1}\delta\eta.\nabla\Delta^{-1}\frac{1}{\varphi_o}\nabla.(\varphi\nabla\delta'\varphi)\, d\Sigma$$

which is the inner product (3.3) of δx by the $(0, \frac{1}{\varphi_o}\nabla.(\varphi\nabla\delta'\varphi), 0)$.

The "transpose" character of the adjoint appears very clearly in the same way that the output of $L\delta x$ of the direct operator depends only on the η-component of δx and that only its φ-component is non zero, the output $L^*\delta' x$ of the adjoint operator depends only on the φ-component of $\delta' x$ and only its η-component is non zero. The term $\frac{1}{\varphi_o}\nabla.(\varphi\nabla\delta'\varphi)$ accordingly appears, in the adjoint equation for $\delta'\eta$ (eqn. (3.4)).

References

Asselin R., 1972: Frequency filter for time integrations. Mon. Wea. Rev., 100, 487-490.

Buckley, A.A. and Lenir, A., 1983: QN-like Variable Storage Conjugate Gradients, Mathematical Programming, 27, 155-175.

Cats, G., Wergen, W., Daley, R., 1986: The objective analysis of planetary-scale flow. Mon. Wea. Rev., 114, 1892-1908.

Courtier, P., 1987: Application du contrôle optimal à la prévision numérique en météorologie. Thèse de doctorat de l'université Paris 6. (in French).

Courtier, P. and Talagrand, O., 1987; Variational assimilation of meteorological observations with the adjoint equation - Part II. Numerical results. Q. J. R. Meteorol. Soc., 113, 1329-1347.

Daley, R., 1978: Variational non-linear normal mode initialization. Tellus, 30, 201-218.

Derber, J.C., 1987: Variational four dimensional analysis using quasi-geostrophic constraints. Mon. Wea. Rev., 115, 998-1008.

Eliasen, E., Machenhauer, B. and Rasmussen, E., 1970: On a numerical method for integration of the hydrodynamical equations with a spectral representation of the horizontal fields. Report No. 2, Institute for Theoretical Meteorology, Copenhagen University, Copenhagen.

Ghil, M., Cohn, S., Tavantzis, J., Bube, K. and Isaacson, E., 1981: Applications of estimation theory to numerical weather prediction. In Dynamic Meteorology. Data Assimilation Methods (L. Bengtsson, M. Ghil and E. Källén, editors). 139-224. Springer-Verlag, New York.

Ghil, M., Cohn, S.E. and Dalcher, A., 1982: Sequential Estimation. Data Assimilation and Initialization. In The Interaction between Objective Analysis and Initialization (D. Williamson, editor). Proceedings of the Stanstead Seminar, McGill University, Montreal, Canada.

Gill, P.E., Murray, W. and Wright, M.H., 1982: Practical optimization. Academic Press, London.

Hall, M.C.G., Cacuci, D.G. and Schlesinger, M.E., 1982: Sensitivity analysis of a radiative-convective model by the adjoint method. J. Atmos. Sci., 39, 2038-2050.

Hoffmann, R.N., 1986: A four-dimensional analysis exactly satisfying equations of motion. Mon. Wea. Rev., 114, 388-397.

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

Lacarra, J.F. and Talagrand, O., 1988: Short range solution 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: theoretical aspects. Tellus, 38A, 97-110.

Leith, C.E., 1980: Nonlinear normal mode initialization and quasi-geostrophic theory. J. Atmos. Sci., 37, 954-964.

Lewis, J.M. and Derber, J.C., 1985: The use of adjoint equations to solve a variational adjustment problem with advective constraints. Tellus, 37A, 309-322.

Lions, J.L., 1971: Optimal control of systems governed by partial differential equations (English translation). Springer-Verlag, Berlin.

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

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

Lorenc, A.C., 1988b: A practical approximation to optimal four dimensional objective analysis. Mon. Wea. Rev., 116, 730-745.

Machenhauer, B., 1977: On the dynamics of gravity oscillations in a shallow water model with applications to normal mode initialization. Beitr. Phys. Atmos. 50, 253-271.

Navon, I.M. and Legler, D.M., 1987: Conjugate-gradient methods for larger scale minimization in meteorology. Mon. Wea. Rev., 115, 1479-1502.

Pailleux, J., Nuret, M., Durand, Y. and Betout, P., 1982: Développement et résultats de l'analyse objective opérationnelle SAPHIR. Working note no 26, Etablissement d'Études et de Recherches Météorologiques, Paris.

Robert, A.J., Henderson, J. and Turnbull, C., 1972: An implicit time integration scheme for baroclinic models of the atmosphere. Mon. Wea. Rev., 100, 329-335.

Sasaki, Y., 1970: Some basic formalisms in numerical variational analysis. Mon. Wea. Rev., 98, 875-883.

Talagrand, O. and Courtier, P., 1987: Variational assimilation of meteorological observations with the adjoint vorticity equation - Part I. Theory. Q. J. R. Meteorol. Soc., 113, 1311-1328.

Thacker, W.C. and Long, R.B., 1988: Fitting dynamics to data. J. Geophys. Res., 93, 1227-1240.

Urban, B., 1985: Croissance maximum des erreurs dans les modèles météorologiques simples, Working report, Ecole Nationale de ls Météorologie, Toulouse, France.

Vautard, R. and Legras, B., 1988: On the sources of the low frequency atmospheric variability. Nonlinear equilibration and weather regimes. To be published in J. Atmos. Sci.

Wahba, G., 1982: Variational methods in simultaneous optimum interpolation and initialization, in The interaction between objective analysis and initialization. (D. Williamson, editor). Proceedings of the Fourteenth Stanstead Seminar. McGill University, Montreal, Canada.