FOUR-DIMENSIONAL VARIATIONAL ASSIMILATION

O. Talagrand
Laboratoire de Météorologie Dynamique
Paris, France

1. INTRODUCTION

The problem of *four-dimensional assimilation* of meteorological observations is usually described as the problem of defining the initial conditions of a numerical weather forecast, using observations distributed over some period in time. Although this description does not really convey the full generality and significance of the underlying *estimation* problem, it obviously corresponds to the daily situation to which meteorologists in weather prediction centres are confronted, and we will use it as a useful starting point for our study of *variational assimilation.*

All assimilation methods which are used (or have been used) in operational numerical weather prediction are *sequential,* in that they consist of a sequence of corrections performed on *one* temporal integration of the assimilating model over the assimilation period. The method of correction which is usually considered as providing the best results is a form of statistical linear estimation, called *optimal interpolation,* described in this volume by Hollingsworth. In a sequential assimilation process, each individual observation is used once and only once, without feedback to anterior times. It is not clear whether this is a real weakness in the case of the definition of the initial conditions of a numerical forecast (it is certainly one when assimilation is used for *a posteriori* determination of the state of the atmospheric flow, with observations available both *before* and *after* the time at which the state of the flow is to be estimated). But it is tempting to look for methods in which *one* model solution over the assimilation period is *globally* adjusted to the available observations, with propagation of the information contained in the observations both forward and backward in time. This is the basic simple idea which underlies variational assimilation. It turns out that, with present numerical models and computers, the only practical way to implement variational assimilation is through an appropriate use of the so-called *adjoint* of the assimilating model. The adjoint of a numerical model basically consists of the equations which govern the temporal evolution of a small perturbation imposed on a model solution, written in a form particularly appropriate for the computation of the sensitivities of output parameters of the model with respect to input parameters. The use of adjoint models is an application of the theory of *optimization* and *optimal control,* which has been developed in the last twenty years by mathematicians (see, *e.g.*, Lions, 1971), and which is progressively propagating in various fields.

This paper describes the principle of variational assimilation and deals in detail with the question of how the adjoint of a numerical model can be used for variational assimilation. It does not describe the numerical results which have already been obtained with variational assimilation. Such results are described in other articles, in particular in the contributions by Courtier and Lorenc included in this volume. On the other hand, emphasis is put here on the theory of adjoint models.

The principle of variational assimilation is stated in precise mathematical terms in Section 2, from which it appears that variational assimilation is a *constrained minimization problem.* It is shown in Section 3 how the adjoint of a model can generally be used for numerically determining the *gradient* of an output parameter of the model with respect to its input parameters. The practical use of adjoint models in the context of variational assimilation is described in Section 4. An important theoretical element is the close link which exists between the method of adjoint models and a classical method for solving constrained minimization problems, namely the method of *Lagrange multipliers.* This particular point is addressed in Section 5. Finally, a number of theoretical and practical problems are discussed in Section 6.

2. VARIATIONAL ASSIMILATION. THE BASIC PRINCIPLE

2.1 Statement of the problem

The two basic ingredients of an assimilation process, whether variational or not, are a set of observations, distributed over a time interval $[t_0,t_1]$, and a numerical model describing the temporal evolution of the atmospheric flow.

We will assume that the observations have been performed at n+1 successive instants τ_i $(i=0,1,...,n)$ which, for the sake of simplicity, will be supposed to be equally distributed over $[t_0, t_1]$, *i.e.* $\tau_i= t_0 + i\Delta t$, with $\Delta t=(t_1-t_0)/n$. The vector of observations at time τ_i will be denoted Z_i. We stress that there is no need for the nature or the number of the individual scalar observations making up the vector Z_i to be independent of i. We will only assume, again for the sake of simplicity, that the atmospheric parameters observed at time τ_i depend only on the state of the atmosphere at that time, so that we exclude observations which could bear, *e.g.*, on temporal integrals or averages.

The numerical model will be written in synthetic form

$$X_{i+1} = X_i + \Delta t\, G(X_i) \qquad i = 0, \ldots, n-1 \tag{2.1}$$

where X_i is an N-dimensional vector describing the state of the model at time τ_i. The N-dimensional phase space of all possible model states will be denoted $\mathcal{E}$. Inner products on $\mathcal{E}$ (and on other spaces when necessary) will be denoted $<\,,\,>$. For large scale models used in present operational numerical weather prediction, N is typically of the order 10^6.

Eq. (2.1) is a discretized form of a set of partial differential equations which can be formally written as

$$\frac{dY}{dt} = F(Y) \tag{2.2}$$

where Y(t) is an infinite dimensional vector describing the state of the atmospheric flow at time t. Passage from (2.2) to (2.1) requires the definition of both a spatial and a temporal discretization schemes, the details of which are not important at this stage.

The specification of an initial condition X_0 at time τ_0 defines a unique solution to eq. (2.1) (we ignore here the additional problems raised by the necessity of specifying appropriate lateral boundary conditions for limited area models, although we stress that the general principles of variational assimilation and of adjoint models apply to limited area models as well as to global models).

The model and the observations being available, it will be possible to assimilate the observations only if it is possible to compare them with the model fields. To this end, we will assume that, given a model state X_i at time τ_i, we can define a "model analogue" $\mathbf{Z}i$ of the observation vector Z_i, and that this model analogue can be computed as an explicit function of X_i, *viz.*

$$\mathbf{Z}i = H_i(X_i)$$

where the subscript i in H_i refers to the fact that the number and nature of the parameters observed at time τ_i may vary with i. H_i will be called the *observation operator* corresponding to the observation vector Z_i. If the model is a gridpoint model, and if the components of Z_i are point values of meteorological fields which are explicitly described in the state vector X of the model (such as horizontal wind components or temperature, which are usually described as such in meteorological models), H_i will then consist of spatial interpolation performed from the model gridpoints to the observation points. A sligthly more complicated situation will occur in the case of a spectral model, where the operator H_i will contain a transformation from spectral to physical space. But still more complicated situations could be imagined, as for instance if the observation vector contained satellite-measured radiances, which one wanted to assimilate directly without first "inverting" them to estimates of temperatures and humidities. The corresponding observation operator would then include an explicit integration of the radiative transfer equation, leading from the model temperature and humidity fields to estimates of the emitted radiances.

Two remarks must be made about the notion of an observation operator. Firstly, an observation operator is defined in terms of the state vector of a particular model, and is therefore "model dependent", even though the observing system will have been defined and implemented, in most cases, independently of any particular model. This is so, of course, because assimilation has to be performed with a particular model, but this fact has far-reaching implications on the way assimilation must be implemented. The second remark is relative to the fact that the observation operator may itself not be very accurate, be that only because a large proportion of meteorological observations are performed at spatial scales which are much smaller than the scales resolved by numerical models. We will come back briefly in Section 6 to the problem of the accuracy of observation operators, and will only mention at this stage that the question of the accuracy of the observation operators is (as the question of the accuracy of the observations themselves) fundamental in the implementation of an assimilation procedure.

An observation operator having been defined at time τ_i, the "misfit", or "distance" between the model state X_i and the observation Z_i will be defined as the scalar product

$$\mathfrak{J}_i = \frac{1}{2} < Z_i - H_i(X_i)\, ,\, Z_i - H_i(X_i) > \tag{2.3}$$

More generally, we consider a sequence $X_0, X_1, \ldots, X_n$ of model states, *i.e.* of vectors of the phase space $\mathfrak{E}$ (which may or may not satisfy the model equation (2.1)). We then define the misfit or distance between that sequence and the sequence of observations $Z_0, Z_1, \ldots, Z_n$ as the sum of the individual misfits

$$\mathcal{J} = \sum_{i=0}^{n} \frac{1}{2} < Z_i - H_i(X_i) \, , \, Z_i - H_i(X_i) > \tag{2.4}$$

The principle of variational assimilation can now be stated in its generality : we shall look for the model solution which minimizes $\mathcal{J}$, *i.e.* for the sequence of states $X_0, X_1, \ldots, X_n$ which minimizes $\mathcal{J}$, while at the same time verifying the model equation (2.1).

The problem just stated is a *constrained minimization problem* : the distance $\mathcal{J}$ (2.4) is to be minimized under the constraint defined by the model equation (2.1). It is known that there are basically two methods for solving a constrained minimization problem : the first one is to *eliminate*, or *reduce*, the constraint, the second one is to use *Lagrange multipliers* associated with the constraint. Interestingly enough, the adjoint of the model equation (2.1), which we will use for solving the above minimization problem, can be obtained from either of these two methods. We will first describe these methods on an elementary example, before coming back to the general problem of variational assimilation.

2.2 An elementary constrained minimization problem

Let us consider the following problem : find the minimum outer area of a right circular cylinder with given volume V. Denoting by h the height of the cylinder, and by R the radius of its basis, this amounts to minimizing the outer area

$$A = 2\pi R \, (R+h) \tag{2.5a}$$

under the constraint that the volume is equal to the known quantity V, *viz.*

$$\pi R^2 h = V \qquad\qquad R > 0 \, , \, h > 0 \tag{2.5b}$$

Elimination of the constraint is in this case absolutely straightforward. One can for instance solve eq. (2.5b) for h, and carry the corresponding expression into (2.5a). The outer area A is thus expressed as a function of the only variable R, and its minimum is obtained by imposing that the derivative dA/dR be equal to zero. This leads to

$$R = \left(\frac{V}{2\pi}\right)^{\frac{1}{3}} \quad \text{and} \quad h = \left(\frac{4V}{\pi}\right)^{\frac{1}{3}} \tag{2.6}$$

It is seen that h = 2R . The cylinder with minimum outer area is the cylinder with square lateral cross-section.

The method of Lagrange multipliers applies as follows in the present case. One scalar additional variable λ, which is the (unique) Lagrange multiplier of the problem, is associated with the scalar constraint (2.5b). One then defines the *Lagrangian* of the problem, *i.e.* the following scalar function of the three variables (R,h, λ)

$$\mathcal{L}(R,h,\lambda) = 2\pi R\,(R+h) - \lambda(\pi R^2 h - V)$$

where the minus sign in front of λ is arbitrary, but convenient. $\mathcal{L}$ is the sum of two terms : the first is the function of R and h to be minimized, and the second is the product of the Lagrange multiplier by the quantity which is constrained to be equal to zero.

A classical result states that the values (R,h) for which the area A (2.5a) is stationary under the constraint (2.5b) are the first two coordinates of the triplets (R,h, λ) for which the Lagrangian, considered as an unconstrained function of its three arguments, is itself stationary.These triplets are obtained by setting to 0 the partial derivatives of $\mathcal{L}$,*viz.*

$$\frac{\partial \mathcal{L}}{\partial R} \equiv 2\pi\,[\,2R + h - \lambda\,Rh\,] = 0 \tag{2.7a}$$

$$\frac{\partial \mathcal{L}}{\partial h} \equiv \pi R\,[\,2 - \lambda\,R] = 0 \tag{2.7b}$$

$$\frac{\partial \mathcal{L}}{\partial \lambda} \equiv -\,[\,\pi\,R^2 h - V] = 0 \tag{2.7c}$$

This system of three equations to the three unknowns (R,h,λ) is easily seen to possess a unique solution defined by (2.6) and by $\lambda = 2/R$.

The method of Lagrange multipliers can be readily generalized to the case of any number N_C of scalar constraints. Let us assume the constraints to be written as

$$f_j(\underline{x}) = 0 \qquad\qquad j = 1, \dots , N_C$$

where $\underline{x} = (x_1, \dots , x_M)$ is the vector of scalar variables with respect to which the constrained minimization is to be performed (the problem makes sense only if $N_C \leq M$). N_C scalar Lagrange multipliers λ_j must then be introduced and the Lagrangian is formed by adding to the scalar function to be minimized the quantity

$$- \sum_{j=1}^{N_C} \lambda_j f_j (\underline{x}) \qquad (2.8)$$

When writing that the Lagrangian is stationary with respect to λ_j, one will recover the constraint $f_j(\underline{x}) = 0$, just as eq. (2.7c) above is identical with the original constraint (2.5b). The quantity (2.8) can often be advantageously be written as the scalar product $- < \Lambda , F(\underline{x}) >$ of the vector $\Lambda = (\lambda_j)$ $(j = 1, \dots , N_C)$ of the Lagrange multipliers by the vector $F(\underline{x}) = (f_j(\underline{x}))$ $(j = 1, \dots , N_C)$ of the constraints.

The relative advantages and disadvantages of reducing the constraint(s) or of introducing Lagrange multipliers(s) depend very much on the particular minimization problem at hand, and will not be discussed here. We will only mention that, in the case of a problem of minimization with respect to M scalar variables constrained by N_C scalar constraints (in the above example, M= 2 and N_C =1), reduction of the constraints leads to a system of M-N_C algebraic equations, while introduction of Lagrange multipliers leads to a system of M+N_C algebraic equations.

Remark. The two methods which have just been described are not the only ones for *numerically* solving constrained minimization problems. Two classes of numerical algorithms for solving such problems are the so-called *penalty* and *augmented Lagrangian* algorithms (see, *e.g.* , Le Dimet and Talagrand, 1986). But these algorithms do not seem appropriate for variational assimilation, and will not be discussed here.

2.3 Minimization. Numerical aspects

We now come back to the constrained minimization problem (2.1-2.4) corresponding to variational assimilation. The dimension of the model phase space $\mathcal{E}$ is N, and n+1 model states are to be determined, so that the number M of scalar quantities which respect to which constrained minimization is to performed is equal to $M=(n+1)N$. As for the number of scalar constraints, which are expressed by the model equation (2.1), it is equal to $N_C = nN$. Now, a model state X_0 at time τ_0 defines a unique model solution, so that the constraints can be formally reduced by considering the distance function $\mathcal{J}$ as a function of X_0 only. This reduces the problem to an unconstrained minimization problem with respect to N scalar variables. But of course an obvious difficulty is now to relate, in a practically usable form, the variations of X_0 with the corresponding variations of $\mathcal{J}$. It would be totally impossible to do the equivalent of what can be trivially done in the elementary example (2.5), namely to derive analytical expressions for the partial derivatives of $\mathcal{J}$ with respect to the components of X_0 (not to mention the problem of determining the values of these components for which the partial derivatives are equal to zero).The approach to the problem must therefore be entirely *numerical.*

It will be convenient to make at this stage a slight change of notation. The vector of initial condition at time τ_0 will be denoted U instead of X_0, and its components will be denoted u_j , ($j = 1, \dots , N$). A basic quantity is the vector of the partial derivatives $\partial\mathcal{J}/\partial u_j$, or *gradient vector*, of the distance function with respect to the components u_j. This vector will be denoted $\nabla_U\mathcal{J}$. It is directed, in the phase space $\mathcal{E}$, along the direction of local fastest variation of $\mathcal{J}$. Given a point U in $\mathcal{E}$ at which $\nabla_U\mathcal{J} \neq 0$, the distance function decreases, for small positive ρ, along the straight line defined by

$$U' = U - \rho \nabla_U\mathcal{J}$$

On the basis of this fact, and provided the local gradient $\nabla_U\mathcal{J}$ can be numerically determined at every point in $\mathcal{E}$, the point in $\mathcal{E}$ at which $\mathcal{J}$ is minimum can be determined as the limit of a sequence of the form

$$U_{p+1} = U_p - \rho_p D_p \qquad p = 0,1,\dots \qquad (2.9)$$

where, for each p, D_p is a vector in $\mathcal{E}$ directed along the direction of the local gradient $\nabla_U \mathcal{J}(U_p)$ or (more efficiently) along a direction which is itself determined from the successive gradients $\nabla_U \mathcal{J}(U_p)$, $\nabla_U \mathcal{J}(U_{p-1})$, ... previously computed. As for ρ_p, it is an appropriately chosen scalar. Numerous *descent algorithms*, which all determine the point at which $\mathcal{J}$ is minimum through a sequence of form (2.9), have been defined. We will mention the *steepest descent* algorithm (in which $D_p = \nabla_U \mathcal{J}(U_p)$ for all p's), the *conjugate gradient* algorithm, and a class of algorithms known under the generic appellation of *quasi-Newton* (or *variable metric*) algorithms. For further information on descent algorithms, we refer the interested reader to the book by Gill *et al.* (1982) and to the article by Navon and Legler (1987). The latter discuss the use of descent algorithms in meteorological problems. The important point here is that, once a method is available for numerically computing the local gradient $\nabla_U \mathcal{J}$, the point at which $\mathcal{J}$ is minimum can then be determined through a descent algorithm.

The question is therefore : how to numerically determine the gradient $\nabla_U \mathcal{J}$? One theoretical possibility would be to evaluate the components of $\nabla_U \mathcal{J}$ through finite-difference approximations of the form $\Delta \mathcal{J} / \Delta u_j$, where $\Delta \mathcal{J}$ is the computed variation of $\mathcal{J}$ resulting from a given perturbation Δu_j of u_j. This method has effectively been used by Hoffman (1986) for performing variational assimilation. But it requires as many explicit computations of $\mathcal{J}$ (*i.e.* explicit integrations of the model over the time interval $[t_0,t_1]$) as there are components in U, *i.e.* N in our notations. Its numerical cost would obviously be totally prohibitive in most situations.

The only possible way to numerically determine $\nabla_U \mathcal{J}$ at an acceptable (although still high) cost is through use of the *adjoint* of the model equation (2.1). The principle of adjoint models is based on a systematic use of the chain rule for differentiation. We will first demonstate the principle of adjoint models in a very general, but conceptually very simple context, in which its significance will appear clearly. We will then describe in detail what this general principle leads to in the case of variational assimilation.

3. THE ADJOINT METHOD FOR COMPUTING GRADIENTS

3.1 Principle

Let us consider a numerical process which, starting from an *input vector* U, with components u_j ($j = 1, \dots, N$), leads to an *output vector* V, with components v_k ($k = 1, \dots, M$). The input vector can be thought of as representing, as in the preceding section, the initial state X_0 from which the model (2.1) is integrated, while the output vector can be thought of as representing, *e.g.*, the complete sequence of model-minus-observations differences $H_i(X_i) - Z_i$ ($i = 0, \dots, n$). But the developments of this subsection are much more general and only require that the output vector is a uniquely defined function of the input vector, *viz.*

$$V = K(U) \qquad (3.1)$$

Now, a perturbation $\delta U = (\delta u_j)$ imposed on U results in a perturbation on V which, to first order with respect to δU, reads

$$\delta v_k = \sum_{j=1}^{N} \frac{\partial v_k}{\partial u_j} \delta u_j \qquad (k = 1, \dots, M) \qquad (3.2)$$

or, in matrix form

$$\delta V = K' \delta U \qquad (3.3)$$

where K' is the *jacobian* matrix of V with respect to U, *i.e.* the matrix whose entries are the partial derivatives $\partial v_k/\partial u_j$. In the general case when the function K is nonlinear, K' will of course be a function of the input vector U. Equation (3.3) will be called the *linear perturbation equation* corresponding to (3.1).

We now consider a scalar function $\mathcal{J}$ of the output vector V. Through (3.1), $\mathcal{J}$ can be considered as well as a function of U. If one knows the gradient $\nabla_V \mathcal{J}$ of $\mathcal{J}$ with respect to V, *i.e.* the partial derivatives $\partial\mathcal{J}/\partial v_k$, the chain rule for differentiation leads for the gradient $\nabla_U \mathcal{J}$ of $\mathcal{J}$ with respect to U, *i.e.* for the partial derivatives $\partial\mathcal{J}/\partial u_j$, to the following expression

$$\frac{\partial \mathfrak{J}}{\partial u_j} = \sum_{k=1}^{M} \frac{\partial v_k}{\partial u_j} \frac{\partial \mathfrak{J}}{\partial v_k} \qquad (j = 1, \dots , N) \tag{3.4}$$

In this expression, the summation is performed on the index k, while it is performed on the index j in eq. (3.2). This means that, expressed in matrix form, eq. (3.4) reads

$$\nabla_U \mathfrak{J} = K'^{*} \nabla_V \mathfrak{J} \tag{3.5}$$

where K'^{*} is the *transpose* of the jacobian matrix K'. It is worth comparing eqs. (3.3) and (3.5). In (3.3), the jacobian matrix K' leads from a *perturbation* on the input vector to the corresponding perturbation on the output vector. In (3.5), the transpose jacobian leads from a *gradient* with respect to the output vector to the corresponding gradient with respect to the input vector. It must be noted that eq. (3.5) is valid for *any* differentiable function $\mathfrak{J}$.

The principle of the transpose method (we do not call it the "adjoint method" yet) is to determine the gradient $\nabla_U \mathfrak{J}$ by numerically performing the computations represented by eq.(3.5). If $\mathfrak{J}$ is a "simple " enough function of the output vector V (and $\mathfrak{J}$ will always be a simple function of some appropriately chosen "output vector", be that $\mathfrak{J}$ itself), $\nabla_V \mathfrak{J}$ can be determined analytically and introduced as input of the computations represented by (3.5). The important point is that these computations can be performed without having to explicitly determine the complete matrix K'^{*}. This fact , which may not seem *a priori* obvious, becomes clear if one realizes that the cost of one transpose computation (3.5) must be the same as the cost of one linear perturbation computation (3.3). The latter does not require the explicit determination of the jacobian K', but can be performed by formally differentiating each step of the computations represented by the basic equation (3.1), and then numerically implementing the corresponding linear computations. For instance, if the basic process (3.1) involves at some stage a computation of the form

$$r = \sqrt{(x^2 + y^2)} \tag{3.6}$$

the corresponding linear perturbation computations will read

$$dr = \frac{\partial r}{\partial x}\, dx + \frac{\partial r}{\partial y}\, dy = \frac{x}{r}\, dx + \frac{y}{r}\, dy$$

or, in standard matrix notation

$$dr = \begin{bmatrix} \frac{x}{r} & \frac{y}{r} \end{bmatrix} \begin{bmatrix} dx \\ dy \end{bmatrix} \tag{3.7}$$

As for the corresponding transpose computations, let us assume that we know the partial derivative $\partial\mathcal{J}/\partial r$ of $\mathcal{J}$ with respect to r (*i.e.* with respect to the output of the particular simple process represented by (3.6)). The derivatives of $\mathcal{J}$ with respect to the inputs x and y are given by the chain rule

$$\frac{\partial\mathcal{J}}{\partial x} = \frac{\partial r}{\partial x}\frac{\partial\mathcal{J}}{\partial r} = \frac{x}{r}\frac{\partial\mathcal{J}}{\partial r}$$

$$\frac{\partial\mathcal{J}}{\partial y} = \frac{\partial r}{\partial y}\frac{\partial\mathcal{J}}{\partial r} = \frac{y}{r}\frac{\partial\mathcal{J}}{\partial r}$$

or, in matrix form

$$\begin{bmatrix} \frac{\partial\mathcal{J}}{\partial x} \\ \\ \frac{\partial\mathcal{J}}{\partial y} \end{bmatrix} = \begin{bmatrix} \frac{x}{r} \\ \\ \frac{y}{r} \end{bmatrix} \frac{\partial\mathcal{J}}{\partial r} \tag{3.8}$$

It is seen that the matrix which appears on the right-hand side of this expression is the transpose of the matrix appearing on the right-hand side of (3.7).

The essence of the transpose method is simply to systematically perform computations of the form (3.8) for *all* the steps of the basic computations (3.1). Indeed, three properties of the transpose method clearly appear on (3.8).

a) Because of their very nature, the transpose computations, which proceed from the gradient of $\mathcal{J}$ with respect to the output vector V to the gradient of $\mathcal{J}$ with respect to the input vector U, must be performed in *reversed order* of the basic computations (3.1). In particular, if, as such is the case in variational assimilation, the direct computations contain some form of temporal integration into the future, the corresponding transpose computations will contain some form of *backward* integration into the past. This particular aspect will be discussed in more detail in the next section.

b) The numerical cost of one transpose computation is the same as the cost of one linear perturbation computation (3.3). The equations gòverning fluid motions, on which numerical models of the atmospheric circulation are built, contain both linear and quadratic terms, and can be symbolically written as

$$\frac{dX}{dt} = L X + N(X,X) \tag{3.9}$$

where LX represents the linear terms, and N(X,X) the quadratic terms. Differentiation, analogous to (3.3), will lead to

$$\frac{d\,\delta X}{dt} = L\,\delta X + N(\delta X, X) + N(X, \delta X)$$

While linear terms are unaltered by differentiation, each quadratic term gives rise to *two* terms. In consequence, the cost of one transpose computation will be between once and twice the cost of one basic computation (3.1).

c) The transpose computations (3.8) require the preliminary knowledge of the partial derivatives $\partial r/\partial x = x/r$ and $\partial r/\partial y = y/r$. The same will be true whenever the basic computations will be nonlinear. As already mentioned, the jacobian K' and its transpose K'^* will in such a case depend on the particular value of the input vector for which one wants to determine the gradient $\nabla_U \mathcal{J}$. This means that, before the transpose computation (3.5) can be implemented, a direct computation (3.1) must have been performed and that all computed intermediary values (or at least all values used in nonlinear computations) must have been stored in memory in order to be available for the transpose computations. The corresponding storage requirements may of course be enormous, and that is the price to be paid for the basic advantage of the transpose method, namely the gain in computing time, which we now briefly discuss.

Taking into account the fact that a direct computation must necessarily be performed before the transpose computation, the cost of the explicit determination of one gradient $\nabla_U \mathfrak{J}$ will, in the context of variational assimilation, be comprised between two and three times the cost of one integration of the model over the assimilation period. Although that cost remains high, it nevertheless achieves an enormous gain over the N integrations which would be necessary if the gradient $\nabla_U \mathfrak{J}$ was to be determined by explicit perturbation of the components of U.

It must also be mentioned that if the linear perturbation equation (3.3) is a necessary theoretical step towards the definition of the transpose equation (3.5), the corresponding numerical computation will not have to be effectively performed, at least in the context considered here. For computing one gradient $\nabla_U \mathfrak{J}$, one integration of the basic equation (3.1), followed by one integration of the transpose equation (3.5), is sufficient.

3.2 The general notion of an adjoint operator

Rather than using the notion of a transpose matrix, it is much more convenient to use the more general notion of an *adjoint operator*. Let us consider two (possibly infinite-dimensional) linear spaces $\mathfrak{E}$ and $\mathfrak{F}$, on which inner products, denoted < , > as before, have been defined. Given a continuous linear operator L of $\mathfrak{E}$ into $\mathfrak{F}$, there exists a unique continuous linear operator L^* of $\mathfrak{F}$ into $\mathfrak{E}$ such that, for any two vectors U and V belonging to $\mathfrak{E}$ and $\mathfrak{F}$ respectively, the following equality between inner products holds

$$< LU , V > \; = \; < U , L^*V > \qquad (3.10)$$

L^* is called the *adjoint* of L. If $\mathfrak{E}$ and $\mathfrak{F}$ have finite dimensions N and M respectively, and are described by coordinates along orthogonal unitary vectors ($U = (u_j), j = 1, \ldots , N$; $V = (v_k), k = 1, \ldots , M$; $L = (l_{jk}), j = 1, \ldots , N , k = 1, \ldots , M$), so that inner products take the familiar form $<u,u'> = \sum_j u_j u'_j$, eq. (3.10) reduces to a change in the order of summation indices

$$\sum_k [\sum_j l_{jk} u_j]\, v_k = \sum_j u_j \,[\sum_k l_{jk} v_k]$$

which shows that the matrix representing L^* is the transpose of the matrix representing L.

In inner product notation, the first-order variation $\delta\mathcal{J}$ resulting from a perturbation δV of V can be written

$$\delta\mathcal{J} = < \nabla_V \mathcal{J} \,, \delta V > \tag{3.11}$$

We can note that this relationship, which generalizes the finite-dimension relationship $\delta\mathcal{J} = \sum_k(\partial\mathcal{J}/\partial v_k)\, \delta v_k$, characterizes the gradient, in the sense that, if the first-order variation $\delta\mathcal{J}$ of a scalar function $\mathcal{J}$ of some vector V can be written as the inner product of δV with some vector, then that vector is necessarily equal to $\nabla_V \mathcal{J}$.

With this more general system of notations, the relationship (3.5) between gradients can be established as follows. Through uses of eq. (3.3) and of the "adjointness" relationship (3.10), (3.11) becomes

$$\delta\mathcal{J} = <\nabla_V \mathcal{J} \,, K'\delta U> = < K'^* \,\nabla_V \mathcal{J} \,, \delta U >$$

The last equality, which expresses $\delta\mathcal{J}$ as the inner product of δU by the vector $K'^* \nabla_V \mathcal{J}$, shows that the gradient $\nabla_U \mathcal{J}$ is equal to the latter vector, which is what is stated by eq. (3.5).

Use of equations such that (3.10) or (3.11) has two distinct advantages : it avoids cumbersome manipulation of indices, and covers the case of infinite-dimensional function spaces, which may be very instructive to consider, even in the context of discretized models. In addition, the general definition (3.10) of an adjoint operator is sometimes very useful, even at the coding level, for developing the adjoint of a numerical model.

4. APPLICATION OF ADJOINT EQUATIONS TO VARIATIONAL ASSIMILATION

4.1 General description

We will now describe in detail what the general adjoint approach described in the previous Section becomes in the particular situation when one wants to determine the gradient of the distance function $\mathcal{J}$ (2.4) with respect to the initial condition X_0 of the model integration (2.1). As already mentioned, the input vector U will be the vector X_0 of initial condition, while the output vector V will be the vector made up of the model-minus-observations differences $H_i(X_i) - Z_i$ ($i = 0,1, ..., n$). In order to take advantage of the form of eqs. (2.1) and (2.4), we will not strictly follow the derivation of the preceding Section. In particular, we will make repeated use of the adjointness relationship (3.10).

The analogue of the basic computation (3.1) is now made up of the integration of the model equation (2.1), which we rewrite here

$$X_{i+1} = X_i + \Delta t\ G(X_i) \qquad i = 0, 1, ... , n-1 \qquad (2.1)$$

followed by computation of the differences $H_i(X_i) - Z_i$ ($i = 0,1, ... , n$). For a given variation δU of the model initial state at time τ_0, the corresponding first-order variation of the distance function $\mathcal{J}$ is

$$\delta\mathcal{J} = \Sigma_i < H_i(X_i) - Z_i\, ,\ H'_i\, \delta X_i > \qquad (4.1)$$

where, for each i, δX_i is the first-order variation of X_i resulting from δU, and H'_i is the jacobian of the operator H_i, taken at point X_i. For each i, δX_i is obtained from the initial perturbation $\delta X_0 = \delta U$ by integrating from τ_0 to τ_i the equation obtained by formally differentiating eq. (2.1), *viz.*

$$\delta X_{i+1} = \delta X_i + \Delta t\, G'_i\, \delta X_i \qquad i = 0,1, ... , n-1 \qquad (4.2)$$

where G'_i is the jacobian of G, taken at point X_i. Equation (4.2) is called the *tangent linear equation* to (2.1), for the solution determined by the particular initial condition U under consideration. The integration of (4.2) is to the integration of the basic equation (2.1) the same thing the linear perturbation computation (3.3) is to the basic computation (3.1). Except of course when the basic equation (2.1) is itself linear, there is one tangent linear equation (4.2) for each solution of (2.1). Also, in the same way eq. (2.1) is a discretized form of eq. (2.2), eq. (4.2) is a discretized form of the following equation

$$\frac{d\,\delta Y}{dt} = F'_t\,\delta Y \tag{4.3}$$

obtained, for a given solution Y(t) of (2.2), by formally differentiating (2.2) with respect to Y. At each time t, F'_t is the jacobian operator of F with respect to Y, taken at point Y(t). Eq. (4.3) is the *tangent linear equation* to (2.2) for the solution Y(t).

The discretized tangent linear equation (4.2) integrates into

$$\delta X_i = (I + \Delta t\, G'_{i-1})\,(I + \Delta t\, G'_{i-2}) \dots (I + \Delta t\, G'_0)\, \delta X_0$$

After carrying this expression into (4.1) and making repeated use of the adjointness relationship (3.10), one obtains

$$\delta\mathfrak{J} = < \Sigma_i\, (I + \Delta t\, G'_0{}^*) \dots (I + \Delta t\, G'_{i-1}{}^*)\ H'_i{}^*\ (H_i(X_i) - Z_i)\, ,\, \delta X_0 > \quad (4.4)$$

in which the adjoint jacobians of the observation and integration operators H and G appear explicitly. By the general definition of a gradient, we see again that the gradient of $\mathfrak{J}$ with respect to the model initial condition is equal to the first factor in the inner product of eq. (4.4), *viz.*

$$\nabla_U\mathfrak{J} = \Sigma_i\, (I + \Delta t\, G'_0{}^*) \dots (I + \Delta t\, G'_{i-1}{}^*)\ H'_i{}^*\ (H_i(X_i) - Z_i) \tag{4.5}$$

This gradient is the sum of n+1 terms, which are the respective gradients of the n+1 terms making up the distance function (2.4). We are going to describe in some detail the exact significance of the term of index i in (4.5), which is the gradient of the individual misfit $\mathfrak{J}_i$ defined by eq. (2.3). To obtain this term, one first applies the adjoint $H'_i{}^*$ of the observation operator H_i on the difference $H_i(X_i) - Z_i$. In accordance with the general adjoint rule (3.5) for gradients, the result is the gradient of the misfit $\mathcal{J}_i$ with respect to the model state X_i at time τ_i. This gradient is then multiplied by the operator $I + \Delta t\, G'_{i-1}{}^*$, which leads to the gradient of $\mathcal{J}_i$ with respect to the model state at time τ_{i-1}. Successive multiplications by $I + \Delta t\, G'_{i-2}{}^*$, ... , $I + \Delta t\, G'_0{}^*$ finally lead to the gradient with respect to the initial state X_0. These successive multiplications can be interpreted as a "backward integration", from τ_i to τ_0, of the finite-difference equation

$$\delta' X_{j-1} = \delta' X_j + \Delta t\, G'_{j-1}{}^* \, \delta' X_j \qquad (4.6)$$

started at time τ_i from $\delta' X_i = H'_i{}^* \, (H_i(X_i) - Z_i)$. Equation (4.6) is called the *adjoint* of the tangent linear equation (4.2). Just as eq. (4.2) is a finite-difference approximation to eq. (4.3), the adjoint equation (4.6) is a finite-difference approximation to

$$- \frac{d\, \delta' Y}{dt} = F'_t{}^* \, \delta' Y \qquad (4.7)$$

One could think, in view of the expression (4.5) for the gradient $\nabla_U \mathcal{J}$, that n integrations of the adjoint equation are necessary, starting from the times τ_n, τ_{n-1}, ... , τ_1 and all ending at time τ_0. The corresponding individual gradients obtained at time τ_0 would then have to be added, together with the term $H'_0{}^* \, (H_0(X_0) - Z_0)$ corresponding to the misfit at time τ_0, in order to obtain the complete gradient $\nabla_U \mathcal{J}$. This is not so. Because of the linearity of the adjoint equation (4.6), *one* integration is sufficient, from τ_n to τ_0, during which the quantity $H'_i{}^* \, (H_i(X_i) - Z_i)$, which can be considered as a kind of "forcing term", is added at time τ_i to the current value of the adjoint variable $\delta' X$. This is in agreement with the already made remark that the numerical cost of an adjoint computation must be the same as the cost of one (hypothetical) tangent linear integration.

In summary, the gradient $\nabla_U \mathfrak{J}$ of the distance function $\mathfrak{J}$ with respect to the initial condition $U = X_0$ can be obtained, for given X_0, by performing the following operations

i) Starting from X_0, integrate the basic equation (2.1) from τ_0 to τ_n, and store in memory the corresponding sequence of states X_i $(i = 0, 1, \dots , n)$ produced by the integration.

ii) Starting from $\delta' X_n = H'_n{}^* \, (H_n(X_n) - Z_n)$, integrate the "forced" adjoint equation

$$\delta' X_i \ = \ \delta' X_{i+1} \ + \ \Delta t \, G'_i{}^* \ \delta' X_{i+1} \ + \ H'_i{}^* \ (H_i(X_i) - Z_i) \qquad i = n-1, \dots , 0 \qquad (4.8)$$

backward in time from τ_n to τ_0. The final result $\delta' X_0$ is the gradient $\nabla_U \mathfrak{J}$. The explicit knowledge of the basic solution X_i is necessary in the adjoint integration in order to compute the quantities $G'_{i-1}{}^*$ and $H'_i{}^* \, (H_i(X_i) - Z_i)$.

The exact nature of the adjoint equation (4.8) and the significance of the corresponding backward integration may arouse some curiosity. It must be clear from the foregoing developments that the adjoint integration is *not* a backward integration of either the basic equation (2.1) or the tangent linear equation (4.2). The fields $\delta' X_i$ produced at time τ_i by the adjoint integration are neither physical fields at time τ_i, nor perturbations on physical fields at time τ_i, but partial derivatives of the distance function (2.4) *with respect* to physical fields at time τ_i.

4.2 An example

In order to describe more precisely the nature and properties of the adjoint equation, we will study in some detail the example of the one-dimensional nonlinear advective-diffusive equation. We will place ourselves in a *non-discretized* setting. The theory of adjoint equations has been developed above in the case of a discretized equation (2.1), but can be developed as well in a non-discretized setting. This leads directly to the non-discretized tangent linear and adjoint equations (4.3) and (4.7), but also requires the use of the somewhat more abstract notions of inner products and gradients in infinite-dimensional spaces. On the other hand, the simplicity of notations makes a number of properties of adjoint equations more clearly apparent in the non-discretized case.

The nonlinear one-dimensional advective-diffusive equation reads, in non-discretized form

$$\frac{\partial u}{\partial t} + \frac{\partial}{\partial x}\left(\frac{u^2}{2}\right) = \nu \frac{\partial^2 u}{\partial x^2} \qquad \nu > 0 \tag{4.9}$$

In this equation, u(x,t) is the scalar velocity, function of time t and of a one-dimensional spatial coordinate x. The latter will be assumed to be periodic with period l. Eq. (4.9) is the analogue of the basic equation (2.2), the analogue of the state vector Y(t) being now, for given t, the one-dimensional field u(x,t), $0 \leq x < l$. The diffusion coefficient ν is positive and it is well known that integration of eq. (4.9) for increasing t constitutes a mathematically well-posed problem, while integration for decreasing t constitutes an ill-posed problem. This fact is of course the mathematical consequence of the irreversible character of the dissipative effects represented by the right-hand side term in eq. (4.9).

For a given solution u(x,t) of eq. (4.9), the corresponding tangent linear equation, which is the analogue of (4.3), is obtained by formally differentiating (4.9) with respect to u

$$\frac{\partial\, \delta u}{\partial t} + \frac{\partial}{\partial x}(u\, \delta u) = \nu \frac{\partial^2 \delta u}{\partial x^2} \tag{4.10}$$

In order to define the adjoint of eq. (4.10), it is first necessary to define an inner product for velocity fields defined on the spatial domain $0 \leq x < l$. An obvious choice is the L^2 product

$$< u_1 , u_2 > = \int_0^l u_1 u_2 dx$$

With this definition, the adjoint of the operator ∂/∂x can be obtained easily, as seen from the following sequence of equalities

$$< u_1 , \frac{\partial u_2}{\partial x} > = \int_0^l u_1 \frac{\partial u_2}{\partial x} dx = \int_0^l -\frac{\partial u_1}{\partial x} u_2 \, dx = < -\frac{\partial u_1}{\partial x} , u_2 >$$

in which integration by parts has been used together with the periodicity condition. The result, according to the general definition (3.10) of an adjoint, is that the adjoint of $\partial/\partial x$ is $-\partial/\partial x$. The operator $\partial/\partial x$, which is equal to the opposite of its own adjoint, is *antisymmetric.* Similar computations show that the operator $\partial^2/\partial x^2$ is *self-adjoint, i.e.* equal to its own adjoint, while the adjoint of the operator $\delta u \rightarrow \partial(u\, \delta u)/\partial x$ is the operator $\delta' u \rightarrow -u\ \partial(\delta' u)/\partial x$. Carrying these results into (4.10) and using (4.7) lead to the adjoint equation of (4.10)

$$-\frac{\partial\, \delta' u}{\partial t} - u\,\frac{\partial\, \delta' u}{\partial x} = \nu\,\frac{\partial^2 \delta' u}{\partial x^2} \qquad (4.11)$$

It is seen that the dissipative term has not been modified but that, because of the minus sign in front of the time derivative, it is now dissipative for integration *into the past.* This fact is easily understandable. If a quantity is dissipated in time, its value at a given time t will become less and less sensitive, as t increases, to its value at an initial time t_0. The quantity produced by the corresponding adjoint integration must therefore *decrease* as the backward integration proceeds.

This fact is quite general, and provides the answer to a natural question about the effect of irreversible processes in adjoint equations. Whenever the basic direct equations contain terms representing irreversibilities which allow only integration into the future, the corresponding terms in the adjoint equation will also be "irreversible", but in the special sense that they will allow only integration into the past. No problem can therefore arise, as long as one is concerned with the existence and the integrability of the adjoint equation, because of the presence of irreversibilities in the basic equation.

In order to give a better understanding of how adjoint equations can be derived in practice, we will now assume that eq. (4.9) has been discretized in both space and time according to the scheme

$$\frac{u_{j,i+1} - u_{j,i-1}}{2\,\Delta t} + \frac{1}{3}\left(u_{j+1,i} + u_{j,i} + u_{j-1,i}\right)\frac{u_{j+1,i} - u_{j-1,i}}{2\,\Delta x} = \nu\,\frac{u_{j+1,i-1} - 2u_{j,i-1} + u_{j-1,i-1}}{\Delta x^2}. \qquad (4.12)$$

where i is, as before, the time index, while j and Δx are the spatial index and increment respectively ($u_{j,i} \approx u(j\Delta x, i\Delta t)$). The particular form chosen for the discretization of the advective term conserves the total kinetic energy $\sum_j (u_{j,i})^2$, while the temporal discretization (leapfrog for the advective term, Euler for the dissipative term) ensures stability of the integration. It is seen that the value $u_{j,i+1}$ at point j and time i+1 depends on the values at the three points j-1, j, j+1 and at both times i-1 and i.

For a given solution $u_{j,i}$ of (4.12), the corresponding tangent linear equation reads

$$\frac{\delta u_{j,i+1} - \delta u_{j,i-1}}{2\,\Delta t} + \frac{1}{3}\,(\delta u_{j+1,i} + \delta u_{j,i} + \delta u_{j-1,i})\,\frac{u_{j+1,i} - u_{j-1,i}}{2\,\Delta x}$$

$$+\frac{1}{3}\,(u_{j+1,i} + u_{j,i} + u_{j-1,i})\,\frac{\delta u_{j+1,i} - \delta u_{j-1,i}}{2\,\Delta x} = \nu\,\frac{\delta u_{j+1,i-1} - 2\,\delta u_{j,i-1} + \delta u_{j-1,i-1}}{\Delta x^2} \tag{4.13}$$

which is a finite-difference approximation of (4.10). Several possibilities are at this stage available for explicitly determining the adjoint equation of (4.13). We will take the one which most closely follows the general description of the adjoint method given in Section 3.

Eq. (4.13), which can be considered as "centred" at point (j,i), defines $\delta u_{j,i+1}$ as a linear combination of the six values $\delta u_{j-1,i-1}$, $\delta u_{j,i-1}$, $\delta u_{j+1,i-1}$, $\delta u_{j-1,i}$, $\delta u_{j,i}$, $\delta u_{j+1,i}$. We will look for the equation which defines the adjoint quantity $\delta' u_{j,i-1}$. The quantity $\delta u_{j,i-1}$ appears in the equations defining $\delta u_{j-1,i}$, $\delta u_{j,i}$, $\delta u_{j+1,i}$, $\delta u_{j-1,i+1}$, $\delta u_{j,i+1}$, $\delta u_{j+1,i+1}$, centred at points (j-1,i-1), (j,i-1), ... respectively. Because of the "transpose" character of the adjoint equation, each of the six adjoint quantities $\delta' u_{j-1,i}$, $\delta' u_{j,i}$, ... will appear in the equation defining $\delta' u_{j,i-1}$, each of them with the same coefficient given to $\delta u_{j,i-1}$ in the respective definitions of $\delta u_{j-1,i}$, $\delta u_{j,i}$, ... For instance, $\delta u_{j,i-1}$ appears in the definition of $\delta u_{j-1,i}$ (equation (4.13) centred at point (j-1, i-1)) with the coefficient

$$\frac{2\,\Delta t}{3}\left[\frac{u_{j,i-1} - u_{j-2,i-1}}{2\,\Delta x} + \frac{u_{j,i-1} + u_{j-1,i-1} + u_{j-2,i-1}}{2\,\Delta x}\right]$$

The quantity $\delta' u_{j-1,i}$ will therefore appear with that same coefficient in the definition of $\delta' u_{j,i-1}$. Identifying all coefficients and recombining terms finally lead to

$$\frac{\delta' u_{j,i-1} - \delta' u_{j,i+1}}{2\,\Delta t} - \frac{1}{6}\, u_{j-1,i-1} \frac{\delta' u_{j,i} - \delta' u_{j-1,i}}{\Delta x} - \frac{2}{3}\, u_{j,i-1} \frac{\delta' u_{j+1,i} - \delta' u_{j-1,i}}{2\,\Delta x}$$

$$- \frac{1}{6}\, u_{j+1,i-1} \frac{\delta' u_{j+1,i} - \delta' u_{j,i}}{\Delta x} = \nu\, \frac{\delta' u_{j-1,i+1} - 2\,\delta' u_{j,i+1} + \delta' u_{j+1,\,i+1}}{\Delta x^{2}} \qquad (4.14)$$

This equation is of course a finite-difference approximation of (4.11). It can be noted in particular that the dissipative term remains dissipative, as in (4.11), for backward integration. It can also be noted that (4.14) is by no means a "straightforward" finite-difference approximation to (4.11). In particular, the basic solution u appears in (4.14) through values at time i-1, while it appeared in (4.13) through values at time i. This fact shows that some care must be exercised when derivating the adjoint of a given direct discretized equation. Remark. We have ignored the (minor) additional difficulty that the three-level leapfrog time-differencing scheme (4.12) must be initialized by a two-level scheme. This means that the adjoint of that two-level scheme will have to be implemented at the *end* of the adjoint integration (4.14). For additional information on that particular point, see, *e.g.*, Appendix C of Talagrand and Courtier (1987).

Let us now assume that an observation, denoted $\upsilon_{k,i}$, has been performed on the u-field at point (k,i), and that one wants to determine the gradient, with respect to the initial conditions of (4.12), of the elementary misfit

$$\mathcal{J} = \frac{1}{2}\left(u_{k,i} - \upsilon_{k,i} \right)^{2}$$

According to what has been explained in the previous subsection, one will have to integrate the adjoint equation (4.14), starting from the "final" conditions $\delta' u_{j,\,i+1} = 0$ for all j's, $\delta' u_{j,i} = 0$ for $j \neq k$, and $\delta' u_{k,i} = \partial \mathcal{J} / \partial u_{k,i} = u_{k,i} - \upsilon_{k,i}$. If observations $\upsilon_{k,i}$ are available at several points in space and time, then, as already explained, one integration of (4.14) will have to performed, in the course of which, at each point (k,i) in the space-time domain, the quantity $u_{k,i} - \upsilon_{k,i}$ will have to be added to the currently computed quantity $\delta' u_{k,i}$. And, if the available observations bear, not on the quantities $u_{k,i}$ themselves, but on quantities Z which are functions of the $u_{k,i}$'s, then, as also explained in the previous subsection, it will be first necessary to apply on the model-minus-observations differences $\mathbf{Z}$ -Z the adjoint of the corresponding observation operators. This will lead to the partial derivatives $\partial\mathcal{J}/\partial u_{k,i}$, which will then have to be fed into the adjoint integration (4.14). The adjoints of the observation operators will have to be defined, from the equations defining those operators, through the same logic which, starting from the basic equation (4.12), has led to the corresponding adjoint equation (4.14). As a simple example, let us assume that an observation, which we will denote z, has been performed on the u-field at time i, and spatial location $(k+\alpha)\Delta x$, $0 < \alpha < 1$. We can take as the corresponding "observation operator" linear interpolation between the gridpoints k and k+1. This gives for the "model analogue" $\mathbf{z}$ of the observation z

$$\mathbf{z} = (1 - \alpha)\, u_{k,i} + \alpha\, u_{k+1,i} \qquad (4.15)$$

The corresponding misfit term is defined by $2\mathcal{J} = (\mathbf{z} - z\,)^2$, and the quantities to be introduced in the adjoint integration at points (k, i) and (k+1, i) are respectively $\partial\mathcal{J}/\partial u_{k,i} = (1 -\alpha)\,(\mathbf{z} - z\,)$ and $\partial\mathcal{J}/\partial u_{k+1,i} = \alpha(\mathbf{z} - z\,)$.

In this example, the observation operator (4.15) is so simple (in particular, it is linear with respect to the model variables) that applying its adjoint through the chain rule for differentiation is absolutely elementary. A much less simple example is treated in Courtier and Talagrand (1987). The model variables are there the spectral components of the vorticity field, while the observations are point values of the geopotential. The corresponding observation operator is the nonlinear balance equation.

5. ADJOINT EQUATIONS AND LAGRANGE MULTIPLIERS

As explained in Section 2, variational assimilation can be considered as a constrained minimization problem, in which the distance function (2.4)

$$\mathcal{J} = \sum_{i=0}^{n} \frac{1}{2} < Z_i - H_i(X_i)\,,\, Z_i - H_i(X_i) >$$

must be minimized under the constraint that the sequence of model states (X_i) $(i = 0,1, \ldots , n)$ satisfy the dynamical equation (2.1)

$$X_{i+1} = X_i + \Delta t \ G(X_i) \qquad i = 0, 1, \ldots, n-1 \qquad (2.1)$$

It has been mentioned that the adjoint equation can be obtained either (as has been done in Section 4) by reducing the constraint (2.1) so as to keep the initial state X_0 as the only independent variable, or by using the general method of Lagrange multipliers. It is the latter approach that we will describe now.

Eq. (2.1) expresses n vector constraints, each of dimension N. It is therefore convenient to define the Lagrange multipliers as making up n vectors Λ_i $(i = 1, \ldots, n)$, each of dimension N. The corresponding Lagrangian, which is a function of the n+1 model states X_i $(i = 0,1,\ldots, n)$ and of the n vector Lagrange multipliers Λ_i $(i = 1, \ldots, n)$ therefore reads

$$\mathcal{L} = \sum_{i=0}^{n} \frac{1}{2} < H_i(X_i) - Z_i \ , \ H_i(X_i) - Z_i >$$

$$- \sum_{i=1}^{n} < \Lambda_i \ , X_i - X_{i-1} - \Delta t \, G(X_{i-1}) >$$

The constrained minimum is obtained by imposing that $\mathcal{L}$ be stationary with respect to all X_i's and Λ_i's. The condition of stationarity with respect to the Λ_i's restores the model equation (2.1). As for the condition of stationarity with respect to the X_i's, let us first consider the case $0 < i < n$. The first -order variation $\delta\mathcal{L}$ of $\mathcal{L}$ resulting from a variation δX_i of X_i (i given) is equal to

$$\delta\mathcal{L} = < H_i(X_i) - Z_i \ , \ H'_i \, \delta X_i > \ - < \Lambda_i \, , \delta X_i > + < \Lambda_{i+1}, (I + \Delta t \, G'_i) \, \delta X_i >$$

which, after using the adjointness relationship (3.10) and reordering terms, becomes

$$\delta\mathcal{L} = < - \Lambda_i + \Lambda_{i+1} + \Delta t \, G'^{*}_i \, \Lambda_{i+1} + H'^{*}_i \ (H_i(X_i) - Z_i) \ , \delta X_i >$$

$\mathcal{L}$ will be stationary with respect to X_i if and only if $\delta\mathcal{L}$ is 0 for any δX_i, *i.e.* if and only if

$$\Lambda_i = \Lambda_{i+1} + \Delta t\, G'^{*}_i \Lambda_{i+1} + H'^{*}_i \left(H_i(X_i) - Z_i \right) \qquad (5.1)$$

In a similar way, stationarity with respect to X_0 leads to the condition

$$\Lambda_1 + \Delta t\, G'^{*}_0 \Lambda_1 + H'^{*}_0 \left(H_0(X_0) - Z_0 \right) = 0$$

which can be written

$$\Lambda_0 = 0$$

if Λ_0 is now defined by extending (5.1) to i = 0. As for the condition for stationarity with respect to Λ_n, it leads to

$$\Lambda_n = H'^{*}_n \left(H_n(X_n) - Z_n \right) \qquad (5.2)$$

The condition for stationarity of $\mathbf{L}$ with respect to all the X_i's (i = 0,1, ... , n) is therefore that the sequence Λ_i defined by (5.2) and by (5.1) for i = n-1, ... , 0 lead to $\Lambda_0 = 0$. Now, eqs. (5.2) and (5.1) are identical with the "forced" adjoint equation (4.8), Λ_i simply replacing $\delta' X_i$. The condition for stationarity is therefore that the adjoint solution take the value 0 at time τ_0. This should certainly not be surprising since we already know that the value taken by the adjoint solution at time τ_0 is the gradient of the distance function with respect to the model state at τ_0. This gradient must of necessity be equal to 0 at the minimum. The additional information here is that the adjoint solution *coincides at the minimum with the Lagrange multipliers.*

Lagrange multipliers have been used by several authors in order to derive the adjoint equations in the context of variational assimilation (Thacker and Long, 1988, O'Brien, pers. com.). Lagrange multipliers can be used for instance for deriving eq. (4.14). The fact that the Lagrange multipliers coincide with the adjoint solution at the minimum has interesting consequences, which will not however, for lack of place, be discussed here. On the other hand, it is not obvious from the Lagrange multiplier approach that the value taken by the adjoint solution at time τ_0 is *always* the gradient of the distance function with respect to the model initial condition, whether the distance function is at its minimum or not. The proof of this fact basically requires the development given here in Sections 3 and 4.

6. DISCUSSION AND CONCLUSION

Variational assimilation of meteorological observations, with explicit use of the adjoint of the assimilating model, has been studied and used in the last few years by a number of authors (see, *e.g.*, Lewis and Derber, 1985, Courtier, 1987, Derber, 1987, Talagrand and Courtier, 1987, Courtier and Talagrand, 1987, Lorenc, 1988, and also the contributions by Lorenc and Courtier in this volume). It has also been studied and used for the problem of assimilation of oceanographical observations, which will in the coming years become a problem of primary importance (see, *e.g.*, Schröter and Wunsch, 1986, Wunsch, 1987, Thacker and Long, 1988). We also mention again the work of Hoffman (1986), who has performed experiments of variational assimilation, but without using the adjoint model.

All experiments of variational assimilation have so far been performed on relatively simple, small dimension models (N of the order of at most a few thousands), such as barotropic or quasi-geostrophic models. The results obtained so far show that variational assimilation is numerically successful, in the sense that the minimization process does converge to a minimum of the distance function. As for the meteorological quality of the results, minimization of a distance function which contains only misfit terms between model values and observations *stricto sensu* leads to a model solution which contains an unrealistic amount of noise, in the form of small scale and/or gravity wave motions. This difficulty can be satisfactorily solved by adding to the distance function one or several terms which measure the amount of noise. The presence of such terms, called *penalty terms* in the general theory of optimization, imposes that the minimizing solution must contain only a small amount of noise. The results obtained by several authors with penalty terms are meteorologically realistic and seem by all standards to be as good as one can expect from the relatively simple models used so far. Lorenc (1988), who has performed variational asssimilation experiments with a one-dimensional model containing a hydraulic jump (a "front") has found that the jump was reconstructed by variational assimilation with a much better accuracy than by optimal interpolation. This is so because the model equation (2.1) is imposed in variational assimilation as a constraint which must be exactly satisfied by the fields produced by the assimilation.

The necessity of adding appropriate penalty terms to the distance function directly points to a basic question. How to exactly define the distance function to be minimized? The numerical results obtained so far suggest that the choices made in the definition of the distance function have been reasonable, but it is nonetheless necessary to try and identify rational grounds on which to define the distance function. This is a vast problem, for the solution of which one theoretical result can be extremely useful. This result is the basic equivalence between statistical linear estimation and one particular form of variational estimation. Statistical linear estimation, of which optimal interpolation is one example, takes as estimate of a given quantity x the linear combination of the observations which best fits x in the mean square statistical sense. In the case of a system whose temporal evolution is governed by a linear equation, statistical linear estimation extends to the temporal dimension under the form of *Kalman filtering* (see, *e.g.*, Ghil *et al.*, 1981). Kalman filtering is a sequential estimation process, whose output at the end of the assimilation is the best linear combination of all the available observations. Still in the case of linear dynamics, the forecast produced from the result of a Kalman assimilation will be the best statistical estimate of the future evolution of the system. The general equivalence between statistical and variational estimation is that statistical estimation produces the same result as the minimization of a distance function defined as the sum of model-minus-observations squared differences, weighted by the inverse of the statistical variances of the corresponding observational errors. The distance function defined by (2.4) is an example of such a sum of squared differences provided that the variances of the observational errors are included in the definition of the inner products appearing in (2.4). Indeed, all authors who have made experiments on variational assimilation have weighted the various terms making up the distance function which coefficients reflecting the estimated accuracy of the observations. The question of whether the linear approximation for the temporal evolution of the forecast error is legitimate, although of great interest, will not be discussed here (see Lacarra and Talagrand, 1988, and Lorenc, 1988, for various aspects of this question). But we will stress the fact that the general equivalence between statistical and variational estimation provides a systematic approach for the definition of the distance function. All observations must be introduced in a model-minus-observation squared difference weighted by the inverse variance of the corresponding observational error. If observational errors associated with different observations are statistically correlated, it is then the inverse of the variance-covariance matrix of observational errors which must be used in the definition of the distance function. This simple logic applies, not only to observations *stricto sensu*, but more generally to all the available information, provided this information is available in numerical form and with an estimate of the corresponding uncertainty. We will take as an example the penalty terms which must be introduced in order to ensure that the minimizing solution is free of unrealistic noise. Such terms, which measure the quadratic norm of the noise, can be described as quadratic misfits to fictitious "observations" that the amount of noise is equal to zero (see also Thacker, 1988). In this sense, variational assimilation provides a systematic and logical approach for using all the available information in a consistent way, each particular piece of information being weighted with its own intrinsic accuracy. The generality and consistency of this approach have far-reaching implications. It may happen for instance that the observation operators H in eq. (2.3) are themselves not very accurate. The corresponding uncertainty can be introduced in the definition of the distance function by noting from (2.3) that what really matters is not the accuracy with which the observation Z fits the the real value of the corresponding meteorological parameter, but the accuracy with which Z fits the model value H(X). The error on the observation operators must therefore be included in the weight given in the distance function to the square of the difference H(X) - Z. And there will be situations where the error on the observation operator will be much larger than the error on the observation itself.

Variational assimilation thus provides a very systematic and general approach for treating in a logical and consistent way all the available information. At the same time, it points out to what the problem of assimilation intrinsically is. It is the problem of *estimating as accurately as possible, using all the available information*, the state of the atmospheric flow at a given time (or at a succession of given times). Seen in this perspective, the problem of assimilation takes a new significance which will certainly be useful for future developments.

If variational assimilation possesses a number of attractive qualities, there nevertheles remain problems, both theoretical (how to introduce the fact that the model, which has here been implicitly assumed to be perfect, will in fact never be perfect?) and practical (the numerical cost of variational assimilation is at present too high for operational use). Also, a systematic comparison between variational asiimilation and squential statistical linear estimation remains to be done. All these problems, which will not be discussed here, are the subjects of active research.

7. REFERENCES

Courtier, P. 1987: *Application du contrôle optimal à la prévision numérique en Météorologie* (in French). Thèse de Doctorat de l'Université-Pierre-et-Marie-Curie, Paris.

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

Derber, J.C., 1987: Variational Four-Dimensional Analysis Using Quasi-Geostrophic Constraints. *Mon. Wea. Rev.*, 115, 998-1008.

Ghil, M., Cohn, S., Tavantzis, J., Bube, K., and E. Isaacson, 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.

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

Hoffman, R.N., 1986: A Four-Dimensional Analysis Exactly Satisfying Equations of Motion. *Mon. Wea. Rev.*, 114, 388-397.

Lacarra, J.F. and O. Talagrand, 1988: Short-range evolution of small perturbations in a barotropic model. *Tellus*, 40A, 81-95.

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

Lewis, J.M., and J.C. Derber, 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, New-York.

Lorenc, A.C., 1988: Optimal nonlinear objective analysis, *Q. J. R. Meteorol. Soc.*,14, 205-240.

Navon, I.M. and D.M. Legler, 1987: Conjugate-Gradient Methods for Large-Scale Minimization in Meteorology. *Mon. Wea. Rev.*, 115, 1479-1502.

Schröter, J. and C. Wunsch, 1986: Solution of Nonlinear Finite Difference Ocean Models by Optimization Methods with Sensitivity and Observational Strategy Analysis. *J. Phys. Oceanogr.*, 16, 1855-1874.

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

Thacker, W.C., 1988: Fitting Models to Inadequate Data by Enforcing Spatial and Temporal Smoothness. *J. Geophys. Res.*, 93, 10,655-10,665.

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

Wunsch, C., 1987: Using transient tracers : the regularization problem. *Tellus*, 39B, 477-492.