AN OVERVIEW OF METEOROLOGICAL DATA ASSIMILATION

P. Morel
Centre National d'Études Spatiales, Paris, France

1. INTRODUCTION. -

The process of numerical weather prediction is classically viewed as an initial value problem whereby the governing equations of geophysical fluid dynamics are integrated forward from fully determined initial values of the meteorological fields at some initial time. Given the peculiar mathematical properties of the Navier-Stokes equations applied to geophysical fluids and the complexity of energetic processes in the atmosphere, solving this "initial value problem" is by itself a tremendous task : many a "model er" is currently engrossed in it. Still, providing the ultimate scheme for integrating the equations of atmospheric dynamics would only be half the answer. For, one must also attend to the problem of determining, from observations of the real atmosphere, the initial values of the many time-dependent parameters which constitute one state of an atmospheric model.

In the ideal physical situation, one would expect that observations would specify unambiguously one value for each and every state parameter at the selected initial time t = 0. Needless to say, this ideal case never materializes in actual meteorological forecasting for a variety of reasons vz :

(i) conventional pressure, temperature and wind observations are inadequately distributed around the planet and leave severe geographical gaps where no data are available.

.../...

(ii) conventional observations are point measurements which do not provide a correct sampling of the highly variable meteorological fields ; such measurements are not representative of true volume averages as required by numerical models.

(iii) in addition, conventional observations are subject to significant random instrumental errors.

(iv) remote observations from space leading to an indirect determination of vertical temperature profiles, go a long way toward providing an homogeneous global coverage leaving no significant data gap over a 6 to 12 hours period of time. Such observations are not synoptic, however, but distributed in space and time following the trajectory of sunsynchronous orbiting satellites. The input from such observing systems must therefore be very incomplete at any one time.

(v) remote observations of cloud motions from geostationary satellites may be available at synoptic hours and provide essentially a proper horizontal sampling of the wind field. But the vertical resolution, limited to one or two levels, is grossly inadequate.

(vi) finally, one will note that such indirect observations from space always involve physical assumptions and fairly sophisticated data processing for reconstructing the meteorological parameters from remotely measured physical quantities. These procedures have their own deficiencies, causing significant random and systematic errors.

.../...

One must regretfully conclude from this brief discussion that single time data sets available to numerical forecasters for updating their computations are likely to remain incomplete and inaccurate, i-e by no mean sufficient to provide by themselves, an adequate description of the global atmosphere. Any forecasting scheme must therefore be initiated or "initialized" at time t = 0 (say) by merging the new observations with the currently estimated meteorological fields computed on the basis of earlier observations collected at times $t < 0$. Formally, the problem consists in optimizing the generalized N - dimensional "trajectory" of the model (considered as a mechanical system with N degrees of freedom) while taking into account all available information at times $t \leqslant 0$ as well as the dynamical constraints between successive (model) states, specified by the governing dynamical equations. This process of merging new observational data with the ongoing integration of a numerical forecasting model is known as "data assimilation" or equivalently "4-dimensional data assimilation" in consideration of the time-space distribution of the data base.

Whether they would recognize it or not, all numerical forecasters resort to some form of 4-dimensional data assimilation but not all forecasters can claim having achieved, or nearly enough, an optimal 4-dimensional data assimilation scheme. On the contrary, most forecasting teams will readily admit that a large part of forecasting deficiencies are rooted in the imperfect assimilation of available data in their numerical prediction process. During the recent years, the development of the Global Atmospheric Research Programme (GARP) and the increasing availability of non-synoptic remote observations from earth orbiting satellites, have fostered a growing interest in this matter of meteorological data analysis and trigerred very significant advances towards developing effective "optimized" data assimilation methods. It is the purpose of this paper, to review briefly the underlying physical principles which are operative in this problem.

.../...

2. STATISTICAL GRID-POINT ANALYSIS. -

Since observations seldom occur at the precise locations represented by the grid-points of one particular model, some 3-dimensional interpolation scheme is a necessary step whenever new data must be merged with predicted fields, whether they are represented by grid-point values or spectral expansion coefficients.

The simplest approach (very bad) consists in substituting the observed value for the corresponding meteorological parameter at the nearest grid-point. The next simplest and most common approach consists in deriving the correction of some computed parameter at one particular location as a linear combination of the departures of actual observations obtained in the vicinity from the (interpolated) computed field values. This linear combination is usually made on the basis of appropriate coefficients or "weights" attributed to individual observations according to the radial distance of their location and their assumed "quality". GANDIN (1963) has originally described how these weights may be optimally determined in order to minimize in a statistical sense, the resulting interpolation error at each grid-point, taking into account the known spatial structure of the meteorological field. This so-called "optimal analysis scheme" was originally applicable to scalar fields only, i-e to one variable at a time. A fairly recent and interesting generalisation has been promoted by SCHLATTER whereby different meteorological parameters may be simultaneously analysed in term of grid-point values in a single multi-variate statistical interpolation process. Interpolation coefficients in such multivariate analysis schemes are computed on the basis of cross-correlation coefficients between different meteorological variables such as geopotential height and one component of the wind velocity, for example. Second-order correlations could in principle be derived directly from observations of the real meteorological fields. This is not usually done, however, because the autocorrelation statistics of the geopotential field are much better

.../...

established than any other second-order moment of the multivariate field. It is current practice therefore, to compute the various statistical moments from the geopotential autocorrelation function alone, using the geostrophic approximation to link the wind variables with the geopotential field. Through this device, the multivariate analysis of wind and geopotential data forces a degree of geostrophic balance which turns out to be beneficial generally, although certainly not justified by any careful consideration of the dynamical effects. The student of 4-dimensional data assimilation should, however, be alert to the fact that excessive filtering of the ageostrophic component of the real atmospheric flow could follow when multivariate interpolation is applied in data sparse regions and is too efficient at "reconstructing" the missing observed field values.

A further remark could be made regarding the spatial resolution of the grid (or equivalent spectral expansion) used for this analysis. The principle of multivariate interpolation is to induce a degree of geostrophic balance in the vicinity of data points. This objective however is attainable only if the model provides a large enough member of degrees of freedom within the influence area around the particular data point. Now, observations of meteorological fields indicate an effective correlation range of the order of 1000 to 1500 km. This means that a multivariate analysis scheme could be effective only when applied to a model with significantly better spatial resolution, i-e no less than 300-500 km.

This condition is normally fulfilled by modern general circulation models. But multivariate interpolation schemes may be, for this reason, ineffective when applied to the analysis of mesoscale features for which a much smaller correlation range is indicated.

.../...

3. THE DYNAMICS OF ADJUSTMENT. -

The very first attempt at numerical weather prediction by RICHARDSON, called attention to the need to eliminate the spurious high frequency oscillations which inevitably result from inserting data taken from an external reference (in this instance, the real atmosphere), in the course of the computation. These oscillations, also referred to as "meteorological noise", have the character of gravity waves superimposed on the quasi-geostrophic flow which constitutes the "meteorological signal". The existence of these waves would not be too serious if they were essentially independent modes, uncoupled to the main flow. But this is not so in the fully (non-linear) interactive case of a primitive equations model : gravity modes can draw energy from the main flow and grow rapidly in the course of time integration. The problem of dealing with high frequency meteorological noise has thus become a central issue in data assimilation.

It is permissible, as a convenient educational device, to rely on the simple model of the linearized shallow-water equations describing an incompressible fluid in a rotating frame for the purpose of discussing the basic concepts of adjustment dynamics. We shall follow this traditional approach and write down the familiar set of governing equations for one component of the FOURIER expansion corresponding to wavevector $\vec{k}$.

$$\frac{\partial \psi}{\partial t} = - f\chi \qquad (1)$$

$$\frac{\partial \chi}{\partial t} = f\psi - gh \qquad (2)$$

$$\frac{\partial h}{\partial t} = Hk^2\chi \qquad (3)$$

Here, the two scalar fields : stream function ψ and velocity potential χ, fully determine the 2-dimensional velocity field, while h is the departure of the free surface altitude from the mean level H. All three variables ψ, χ and h are first order quantities for a small perturbation of the

.../...

state of rest of the fluid. One recognizes above, the vorticity equation (1), the divergence equation (2) and the continuity equation (3). This set of equations admits of three linearly independent solutions or normal modes for each wavevector $\vec{k}$ namely :

- one geostrophically balanced mode which turns out to be stationary in this simple model.

- two high frequency modes corresponding to propagating gravity waves along the forward and backward directions, respectively.

One could also view each normal mode as an eigenvector X_i corresponding to each of the three eigenvalues $\lambda_1 = 0$, $\lambda_2 = \sqrt{gHk^2 + f^2}$ and $\lambda_3 = - \lambda_2$. If one now projects the equations (1) through (3) onto these normal modes, one obtains a set of ordinary differential equations which may be written symbolically as :

$$\dot{X} = - i\Lambda X \qquad (4)$$

where X is a column vector of normal mode expansion coefficients and Λ is a diagonal matrix, the elements of which are the three eigenvalues λ_i.

Initializing this simple dynamical model with an arbitrary state vector X at time t = 0 or equivalently, an arbitrary selection of the fields ψ, χ and h, will normally result in exciting all three normal modes about equally. In the simple linear model, the initial distribution of perturbation energy will simply be invariant at all times t > 0 without amplification nor damping.

In the more realistic non-linear dynamical models, however, the non-linear advection terms and the forcing terms have a non-zero projection on the

.../...

set of normal modes X, so that equation (4) would now read :

$$\dot{X} = - i\Lambda X + N\ (X,\ X) \qquad (5)$$

and allow modal coupling, which in turn induces a tendency towards equipartition of energy among the three competing modes. It is a fact that, barring ad hoc precautions, numerical models develop a much larger amount of gravity wave energy than actually found in the real atmosphere. Conversely, one could well take the view that the absence of significant gravity oscillations in free atmospheric flows, is a remarkable and yet not fully explained property of the Earth atmosphere. One possible reason for it might be found in the major energy conversion process associated with baroclinic instability, which favours the generation of eddy mechanical energy in the form of quasi-geostrophic perturbations of the mean flow. Alternately, another cause for selective dissipation of gravity oscillations might be looked for in the basic non-linear dynamics of spectral energy transfer which favours the concentration of quasi two-dimensional motions in the larger scales, while the energy of 3-dimensional gravity waves is allowed to cascade towards smaller scales and dissipate by friction. Yet another possible explanation could be seen in the dynamics of small scale energetic processes (like moist convection) which could be selectively triggered by gravity oscillations.

Whatever the main cause for the damping of gravity modes in the real atmosphere, it is imperative to develop filtering techniques for removing the spuriously large gravity oscillations induced in numerical models by non-linear interaction, as well as the insertion of heterogeneous data.

The simplest and historically the first approach was simply to exclude the possibility of such high frequency oscillations by reducing the model

.../...

atmosphere dynamics to the only quasi-geostrophic response allowed by a "filtered equation " like the balance equation. This approach so severely restricts the model dynamics, however, that corresponding forecasts are not useful beyond a few hours. We must therefore concern ourselves with the so-called primitive equation models which do allow the unrestricted propagation and eventual amplification of gravity waves and, lacking a fundamental understanding of the real processes which limit gravity oscillations in the atmosphere, we must provide an artificial damping scheme suitable for smothering the unwanted waves and restoring the quasi-geostrophic balance of the model flow.

Many ingenious damping schemes have been proposed in the literature and found to reduce significantly the level of "meteorological noise" in the models. But taking the large view, one may say that no such scheme could provide a satisfactory answer to the problem of data assimilation, for one basic dynamical reason. Artificial damping is needed in the normal course of numerical modelling to replace the small scale dissipation processes which must occur in the real atmosphere, but which lie beyond the range of scales explicitely represented in the model flow. In other words, the artificial damping must somehow stand for sub-gridscale processes which are, by necessity, not included in the finite number of degrees of freedom of the model flow. For dynamical consistency, the dissipation rate introduced by artificial damping must be commensurate with the spectral transfer rate which would develop naturally within the spectral range of motions explicitely represented in the model. These spectral transfers are determined by the laws of geophysical fluid dynamics and cannot be artificially increased without causing severe discrepancies with regard to the real atmosphere dynamics. The consequence is an unbreakable dilemma : the rate of artificial damping may be chosen either small enough to match the

.../...

real atmosphere sub-gridscale dissipation rate or large enough to reduce effectively the meteorological noise but both desirable conditions cannot be met simultaneously. Increasing the artificial damping rate is inevitably detrimental to the ability of the forecasting model to simulate the real atmosphere and results in serious deviations from the reality in data sparse regions where an accurate simulation is needed. On the other hand, decreasing the artificial damping rate results in excessive meteorological noise induced by new data inputs. A consequence of this situation is the unexpected but theoretically obvious fact that adding more noisy data into a well-tuned meteorological analysis routine results in a worse forecasting skill even though the new data do include additional information which could serve to improve the definition of the real atmosphere initial state. The most spectacular example of that was afforded by the protracted argument regarding the degree of usefulness of satellite temperature profile data. For the lack of a suitable optimization of their analysis scheme, many operational forecasting services were led to developing a "Dont bother us with new data " attitude when they found that satellite temperature data inserted in their normal procedure would not improve and could even degrade short and medium-term predictions.

We must conclude then, that the provision of artificial damping is not the proper way for reducing the amplitude of the high frequency oscillations induced by the insertion of heterogeneous (and possibly noisy) data in the course of numerical modelling of the atmosphere circulation. The only proper solution is thus to avoid triggering such oscillations by appropriate data conditioning before insertion. This is the topic of the next section on normal mode initialization procedures.

.../...

4. NORMAL MODE INITIALIZATION. -

This method is based on the obvious idea of separating from the outset the contribution to updating the balanced quasi-geostrophic motion of the model flow,from the contribution to the excitation of spurious gravity oscillations, by means of a suitable modal decomposition of the raw correction field resulting from the input of a new data set. Such an approach, based upon normal mode solutions to a linearized version of the forecasting equations, was first used by FLATTERY (1970) in his original analysis scheme featuring HOUGH functions which are the solutions of the linearized shallow water equations on a sphere, first introduced in tidal research. DICKINSON and WILLIAMSON (1972) later proposed a general method to find the normal modes for an arbitrary general circulation model and a corresponding linear mode filtering scheme for initializing numerical forecasts.

In the linear filtering procedure, the observational data are expanded in terms of the complete set of normal modes for the linearized governing equations, and then the amplitudes of the unwanted computational and gravity modes are set equal to zero. This method does reduce the unrealistic large high-frequency oscillations which occur during the initial period of the forecasts when no a-priori filtering is used. But the technique suffers from a deficiency which is apparent from equation (5) above : because of the presence of a non-linear coupling term, the unwanted high-frequency oscillations can be regenerated fairly rapidly in the course of the time integration. This deficiency may be eliminated to a large extent by the so-called non-linear normal mode initialization defined as a normal mode filtering procedure for which the time derivatives of the gravity modes coefficients (but not the initial values of these coefficients) are set equal to zero. Gravity mode coefficients are modified so that the

.../...

linear tendency corresponding to the diagonal part $-i\Lambda X$ of the evolution matrix, compensates for the contribution N (X,X) from the non-linear interactions between all modes. See the original discussions by MACHENHAUER (1977) and BAER (1977) who first proposed this refinement of the normal mode filtering technique.

This idea is equivalent to the concept developed by LEITH (1979), of projecting the initial (noisy) meteorological fields produced by straightforward interpolation of the observational data, onto the slow manifold constitued by all slowly varying dynamical solutions of the model equations while excluding all rapidly oscillating states (represented in configuration space by fast oscillations about the slow manifold).

Various versions of this refined filtering procedure with or without precautions for minimizing the rejection of original data inputs, have been discussed by DALEY and PURI (1980) and a procedure of this kind has been implemented with encouraging results by the European Center for Medium Range Weather Forecasts. It remains to be seen whether one standard modal expansion, based on the normal mode solutions of the linearized equations about one simple state of the model flow (i-e the state of rest) is adequate for all initial meteorological conditions, or whether a much more cumbersome diagonalization procedure for each one initial state is actually needed.

A very positive indication of the success of this method for filtering out the meteorological noise from the outset, is seen in the remarkable ability of the ECMRWF forecasting procedure, to accept large amounts of unconventional non-synoptic satellite data and produce a definite improvement over forecasts based on synoptic data only. In this instance, we find at last that more information, even if it is noisy information, yields a more accurate description of the real atmosphere.

5. THE MATHEMATICAL PROBLEM OF 4-DIMENSIONAL ASSIMILATION. -

The mathematical basis for understanding the continuous or discontinuous adjustment process involved in 4-dimensional data assimilation is not well established. Coming back to the introductory remarks, we can see that one reason for the unavailability of a sound theoretical background may be the lack of strong existence and unicity theorems for the solutions of the Navier-Stokes equations in the case of a fully developed turbulent flow (large REYNOLDS number). The fact is that forcing successive corrections onto the on-going time integration of a general circulation model is a purely numerical process which has no counterpart in the physical world. Consequently the eventual convergence of this process or the lack thereof, cannot be judged on the basis of qualitative physical arguments or limited numerical experimentation. A thorough theoretical approach is a very difficult, possibly intractable problem, however. A linear theory of the convergence of one particular dynamical initialization method based on repeated forcing the same set of field values on the computed fields obtained in successive forward and backward time-integrations, has been attempted by TALAGRAND (to be published). This iterative procedure lends itself readily to a proper definition of "convergence" while convincing numerical experiments have shown it to be essentially equivalent to successive forcing of an ongoing forecast with the same amount of external data as provided in the forward-backward assimilation process during a given time interval. This linear theory proved to be inconclusive, however.

A more demonstrative approach has also been tried by TALAGRAND when the flow can be legitimately approximated by solutions or any linear combination of solutions, of the linearized shallow water equations (1) to (3) ; a similar reasoning can also be developed in the case of a multilevel model. This approach is based on the consideration of the perturbation energy invariant :

$$E = \frac{1}{2} \int \left[gh^2 + HV^2 \right] ds \qquad (6)$$

.../...

where $\frac{1}{2} gh^2$ is the perturbation available potential energy and $\frac{1}{2} HV^2$ is the perturbation kinetic energy per unit area (assuming an incompressible fluid of unit density). Now, since a linear combination of two solutions of the linearized problem is also a solution, it follows that the quadratic différence :

$$D = \frac{1}{2} \int \left[g(h-h_0)^2 + H (\vec{V} - \vec{V}_0)^2 \right] ds \qquad (7)$$

is also invariant in the course of the unperturbed time integration of the initial value problem starting with the reference field $(h_0, \vec{V}_0)$ and with the "first-guess" field or trial field $(h, \vec{V})$. But D is obviously changed by forcing "observational"data taken from the reference model state onto the evolution of the trial field. Indeed, substituting values of parameters h_0 or $\vec{V}_0$ from the reference field at time t, for the corresponding parameters h or $\vec{V}$ in the trial field nullifies the contribution to expression (7) at this time and place. Thus, the straightforward substitution of any subset of reference field values for the corresponding subset of parameters in the trial field always reduces the quadratic difference D between the two model states. The resulting monotonous decrease of the quadratic difference does not, by itself, guarantee that the difference will eventually decrease to zero, i-e that the successive forcing process would converge. It can be shown in the case of backward-forward integration tion that the process does converge if and only if the "observational data" available during the time interval of one cycle specify a unique solution of the governing equations. This is obviously a necessary condition for convergence ; it is also sufficient in this simple case, irrespective of the nature of the reference state, i-e geostrophically balanced or not. The convergence of 4-dimensional data assimilation, towards one particular solution of the model equations, does not depend here upon geostrophic adjustment nor damping gravity waves generated by the straightforward substitution of heterogeneous field values. One must indeed consider that

.../...

the very special role played by the quasi-geostrophic modes in meteorological forecasting is a contingent and indeed very fortunate circumstance which reduces the 4-dimensional assimilation problem to the consideration of the "slow-manifold" instead of the full range of all possible solutions of the equations of geophysical fluid dynamics.

But one should not forget that this special circumstance is not central to the mathematical problem of the convergence of this rather exotic numerical procedure described as 4-dimensional data analysis.

R E F E R E N C E S

Baer, F. (1977) Adjustment of initial conditions required to suppress gravity oscillations in non-linear flows. Contrib. Atmos. Phys. 50, 350-366.

Daley, R. and Puri, K. (1980) Four-dimensional data assimilation and the slow manifold. Mon. Weather Rev., 108, 85-99.

Dickinson, R.E. and Williamson, D.L. (1972) Free oscillations of a discrete stratified fluid with application to numerical weather prediction. J. Atmos. Sci., 29, 623-640.

Flattery, T.W. (1980) Spectral models for global analysis and forecasting. Proc. Sixth AWS Tech. Exchange Conf., U.S. Naval Academy, 21-24 Sept. 1974. Air Weather Service Tech. Rept. 242.

Gandin (1963) Objective analysis of meteorological fields, Leningrad, Gidrometeoizdat.

Leith, C.E. (1979) Non-linear normal mode initialization and quasi-geostrophic theory, NCAR MS, 0901-79-01. Available from National Center for Atmospheric Research. Boulder, Col. USA.

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

.../...

Schlatter (1975) Some experiments with a multivariate statistical objective analysis scheme. Mon. Weather Rev., 103, 246-257.

Talagrand (1980) On the mathematics of Data Assimilation. (submitted to Tellus).