A POSTERIORI EVALUATION AND VERIFICATION OF ANALYSIS AND ASSIMILATION ALGORITHMS

O Talagrand
Laboratoire de Météorologie Dynamique du CNRS
Paris, France

Summary. *A posteriori* evaluation and verification of analysis and assimilation algorithms is discussed in the framework of statistical linear estimation. It is shown that it is essentially equivalent to perform diagnostics on the innovation vector or on the *information-minus-analysis* (ImA) difference, *i. e.* the difference between the estimated fields and the information (including observations) used in the estimation process. Interpretation of the statistics of either quantity requires *a priori* hypotheses which cannot as such be objectively validated. In a variational estimation process, one simple significant diagnostic is the minimum of the objective function. In strong-constraint variational assimilation, useful information on the model error must be contained in the observations-minus-minimizing-solution (OmMS) difference. It is suggested that a positive convexity in the temporal varaition of the squared OmMS difference is a sign of model error. The adjoint solution, which consists of the Lagrange multipliers associated with the (strong-constraint) model equations, must also contain useful information.

1. INTRODUCTION

Analysis and assimilation of observations can be described as processes intended at estimating as accurately as possible the state of the atmospheric flow, using all available relevant information. The available relevant information essentially consists, in addition to the observations proper, of a prior *background* estimate of the state of the flow, of the physical laws which govern the evolution of the flow (available in practice under the form of a discretized numerical model), and of statistical or dynamical properties of the flow, such as for instance the approximate geostrophic balance of middle latitudes. The data (individual pieces of information) are combined together in the estimation process, each individual datum being given a weight which is meant to reflect the confidence granted to that datum. The weights are therefore essentially inversely proportional to the errors which statistically affect the various data. Much uncertainty remains however as to what these errors are (very little is quantitatively known for instance on the errors affecting numerical models used in weather prediction), and there potentially exists much room for improvement of analysis and assimilation methods through simply a better specification of the *a priori* assumed statistics of the errors affecting the various sources of information.

These notes discuss a few aspects of *a posteriori* evaluation and validation of analysis and assimilation algorithms. The discussion essentially revolves around the following question. Is it possible, by *a posteriori* objective verification, to identify 'errors', or imperfections of some sort, in the *a priori* specification of the weights used in the estimation algorithm, and to correct for these imperfections ? What is meant here by '*a posteriori* objective verification' is comparison between the estimated fields and whatever relevant quantitative information may be available. If it.is clear that objective assessment of the quality of an estimation procedure can be made only against observations which are independent of the information used in the estimation process itself, the fit of the estimated fields to that information can nevertheless provide useful diagnostics. An important fact stressed below is that objective validation of an estimation process requires *a priori* hypotheses which cannot as such be objectively validated.

Most of the algorithms used at present for analysis and assimilation of meteorological observations can be described as particular applications of the theory of *statistical linear estimation*, and the discussion below lies entirely within the range of that theory. Section 2 is a reminder of a few basics facts about statistical linear estimation, including facts which have so far been little used in meteorological applications. Section 3 describes a number of diagnostics which can be performed on the difference between estimated fields and the information used in the estimation process. Section 4 is more specifically devoted to diagnostics, bearing in particular on the errors in the assimilating model, which can be made from the results of strong-constraint four-dimensional variational assimilation. Some additional comments are given in Section 5.

2. A FEW BASIC FACTS ABOUT STATISTICAL LINEAR ESTIMATION

A fairly general presentation of the theory of statistical linear estimation is as as follows. A vector x, describing the state of the system under study, is to be estimated. The vector x belongs to the *state space* S, with dimension n. The quantitative information available for estimating x makes up a vector z, belonging to the *information space* I, with dimension m. The information vector is related to the state vector through the relation

$$z = \Gamma x + \zeta \qquad (2.1)$$

where Γ is a known (and possibly approximate) operator, represented by an m x n matrix, and ζ is a residual representing the various 'errors' in z. The vector ζ is assumed to be known only through its statistical properties.

We now look for an estimate x^a of x of the form

$$x^a = \alpha + Az \qquad (2.2)$$

where α and A are respectively an n-vector and an n x m matrix to be determined under the following conditions

- the estimate x^a is invariant in a change of origin in state space
- the statistical variance $E[(x^a-x)^T(x^a-x)]$ of the estimation error x^a-x (where E denotes statistical expectation and ^{T} transposition) is minimum.

The solution to that problem is

$$x^a = [\Gamma^T S^{-1}\Gamma]^{-1}\, \Gamma^T S^{-1}[z - E(\zeta)] \qquad (2.3)$$

i. e.

$$A = [\Gamma^T S^{-1}\Gamma]^{-1}\, \Gamma^T S^{-1} \qquad (2.4a)$$

and

$$\alpha = - A\, E(\zeta) \qquad (2.4b)$$

In these equations, S is the covariance matrix of ζ, *i. e.*

$$S = E\{[\zeta - E(\zeta)]\,[\zeta - E(\zeta)]^T\} \qquad (2.5)$$

The corresponding estimation error $x^a - x = A[\zeta - E(\zeta)]$ is unbiased

$$E(x^a - x) = 0 \qquad (2.6a)$$

while its covariance matrix reads

$$P^a \equiv E[(x^a - x)\,(x^a - x)^T] = [\Gamma^T S^{-1}\Gamma]^{-1} \qquad (2.6b)$$

The estimate x^a is called the *Best Linear Unbiased Estimate*, or *BLUE*, of x from z. It is unambiguously defined if and only if the operator Γ is one-to-one, *i. e.* if and only if

$$\Gamma x_1 \neq \Gamma x_2 \text{ for any two distinct vectors } x_1 \neq x_2 \text{ in state space} \qquad (2.7)$$

This is an *observability condition* (*informativity* might be here a more appropriate word) which expresses that the vector z contains information, either directly or indirectly, on every component of x. It implies $m \geq n$. We will denote $m = n + p$.

We mention two interesting properties of the *BLUE*.

1. The matrix A (2.4a) is a left-inverse of Γ, *i. e.* $A\Gamma = I_n$, where I_n is the unit matrix of order n. This means that the BLUE obtained from a non-noisy (ζ=0) information vector is the exact state vector of the system.

2. The *BLUE* is invariant in any linear change of coordinates in either state or information space.

This means in particular that the variance E $[(x^a - x)^T \mathrm{B}\, (x^a - x)]$ is minimum for any n x n symmetric positive-definite matrix B, *i. e.* the estimation error is statistically minimum for any (quadratic) norm in state space.

Unless otherwise specified, it will be assumed in the following that the information vector is (or has been) unbiased, *i. e.* $\mathrm{E}(\zeta) = 0$.

When the observability condition (2.7) is verified, it is always possible, through an appropriate change of coordinates in information space, to put the information vector under the form

$$z = \begin{pmatrix} x^b = x + \zeta^b \\ y = \mathrm{H}x + \varepsilon \end{pmatrix} \qquad (2.8)$$

where x^b is called a *background* estimate of x, and y is a complementary vector of information, with dimension p, associated with an information operator H. The change of coordinates can always be defined in such a way that the errors ζ^b and ε are uncorrrelated

$$\mathrm{E}(\zeta^b \varepsilon^T) = 0$$

Introducing the covariance matrices of the errors ζ^b and ε, *viz.*,

$$\mathrm{P}^b \equiv \mathrm{E}(\zeta^b \zeta^{bT}) \quad \text{and} \quad \mathrm{R} \equiv \mathrm{E}(\varepsilon\varepsilon^T) \qquad (2.9)$$

equation (2.3) then takes the familiar form

$$x^a = x^b + \mathrm{K}d \qquad (2.10a)$$

where d is the *innovation vector*

$$d = y - \mathrm{H}x^b \qquad (2.10b)$$

and K is the *gain matrix*

$$\mathrm{K} = \mathrm{P}^b\mathrm{H}^T\, [\mathrm{H}\mathrm{P}^b\mathrm{H}^T + \mathrm{R}]^{-1} \qquad (2.11)$$

We note that

$$HP^bH^T + R = E(dd^T)$$

Equation (2.6b) takes the form

$$P^a = P^b - P^bH^T [HP^bH^T + R]^{-1} HP^b \qquad (2.12)$$

Alternatively, the BLUE can be defined through a variational approach as the minimizer of the following scalar *objective function*, defined on information space

$$\xi \rightarrow J(\xi) \equiv (1/2) [\Gamma\xi - z]^T S^{-1} [\Gamma\xi - z] \qquad (2.13)$$

Minimization of (2.13) has a simple interpretation. For any two vectors η_1 and η_2 in information space, the quantity

$$(1/2)\ \eta_1{}^T S^{-1}\ \eta_2 \qquad (2.14)$$

is a proper scalar product, often called the *Mahalanobis scalar product* associated with the covariance matrix S. Minimizing (2.13) therefore amounts to looking for the point in the image $\Gamma(S)$ of the state space which lies closest to z in the sense of the scalar product (2.14). That point is the orthogonal projection, in the sense (2.14), of z onto $\Gamma(S)$. The *BLUE* x^a is therefore the inverse, through Γ, of the orthogonal projection of z onto the image space $\Gamma(S)$.

As already said, the matrix A defined by (2.4a) is a left-inverse of Γ. Conversely, any left-inverse of Γ is of the form (2.4a) for some symmetric positive-definite S. Any estimation scheme of the form (2.2), where A is a left-inverse of Γ, can therefore be considered as the *BLUE* of x for some expectation and covariance matrix of the information error ζ. In the background-innovation representation (2.8), saying that A is a left-inverse of Γ is equivalent to saying that x^a is of the form (2.10a-b), where K is any n x p matrix.

Most analysis and assimilation schemes used in meteorological applications can be decribed as determining the *BLUE* of x corresponding to some (often implicitly) prespecified $E(\zeta)$ and S. This is true of any scheme of the form (2.10a-b), and in particular of Kalman filtering, of three-dimensional variational analysis, and of four-dimensional variational assimilation (under both its weak- and strong-constraint formulations).

3. THE INFORMATION-MINUS-ANALYSIS DIFFERENCE

We now discuss the problem of *a posteriori* evaluation of analysis and assimilation schemes. We will assume that the information operator Γ is known, and consider schemes of the form (2.2), where A is a left-inverse of Γ. Questions that naturally arise in this context are for instance the following. Is it possible to objectively determine if a given scheme is optimal, in the sense of the *BLUE* ? Assuming a scheme has been shown not to be optimal, how is it possible to improve it ?

A readily available quantity is the difference between the information vector and the corresponding values in the analyzed fields, *viz.*,

$$\delta = z - \Gamma x^{\mathrm{a}}$$

δ belongs to information space. We will call it the *information-minus-analysis* (ImA) vector. The magnitude of that vector is essentially determined by the *a priori* specified expectation vector E(ζ) and covariance matrix S. That magnitude cannot therefore be used as a measure of the quality of the estimation, and cannot in particular be used for comparing the quality of two different estimation processes. This can be done only by comparing the analyzed fields with observations which have not been used in the estimation process, and more precisely with observations which, in addition to being non-biased, are affected by errors which are themselves uncorrelated with the error vector ζ.

The ImA difference can nevertheless be a powerful diagnostic tool, as we will discuss now. In the background-innovation description (2.8-2.10), the ImA difference reads

$$\delta = \begin{pmatrix} x^{\mathrm{b}} - x^{\mathrm{a}} = -\mathrm{K}d \\ y - \mathrm{H}x^{\mathrm{a}} = (\mathrm{I}_p - \mathrm{HK})\, d \end{pmatrix} \tag{3.1}$$

where I_p is the unit matrix of order p. Eq. (3.1) shows that δ is a linear transform of the innovation vector d. The corresponding transformation is invertible since

$$d = y - \mathrm{H}x^{\mathrm{a}} - \mathrm{H}(x^{\mathrm{b}} - x^{\mathrm{a}})$$

The innovation and ImA vectors therefore contain the same information, but many features may be more clearly apparent on the latter. Looking first at the expectation of the ImA difference, it must be zero for an optimal system.

$$E(z - \Gamma x^a) = 0 \tag{3.2}$$

If, as required above, the matrix A is a left-inverse of Γ, a non-zero expectation for the ImA difference is necessarily the sign that a bias has not been properly taken into account in the error ζ.

The covariance matrix of the ImA difference reads

$$\Delta \equiv E[\delta\delta^T] = S - \Gamma\, [\Gamma^T S^{-1} \Gamma]^{-1}\, \Gamma^T \tag{3.3a}$$

or equivalently

$$E[(z - \Gamma x^a)\, (z - \Gamma x^a)^T] = E[(\Gamma x - z)\, (\Gamma x - z)^T] - E[(\Gamma x^a - \Gamma x)\, (\Gamma x^a - \Gamma x)^T] \tag{3.3b}$$

The term subtracted on the right-hand side of these equations is positive, which means that the *BLUE* will fit the information z to within the corresponding error ζ (*Hollingsworth and Lönnberg*, 1989, have called *efficient* a system which possesses that particular property). More precisely, eqs (3.3) show that, as the estimation error will decrease (as a consequence for instance of the use of an increasing amount of information), the magnitude of the ImA difference will increase to asymptotically saturate to the level of information error.

The Pythagorean form of eq. (3.3b) means that the triangle defined by the three 'points' z, Γx^a, Γx has a right angle (with respect to the orthogonality defined by statistical covariance) at point Γx^a. More generally, the ImA difference and the estimation error are statistically uncorrelated

$$E[(z - \Gamma x^a)\, (x^a - x)^T] = 0 \tag{3.4}$$

This important property characterizes the *BLUE*.

Any statistically significant discrepancy between the *a posteriori* observed statistics of the ImA difference and the 'predicted' statistics (3.2-3), such as for instance an ImA variance that is larger than the variance of the corresponding information error component, is necessarily the sign of a misspecification in the *a priori* statistics of the information error ζ. How to interpret such a discrepancy, and how to determine whether it can be used for improving the estimate x^a, or the estimation error covariance P^a, nevertheless require some care. It results from the variational formulation (2.13) of the *BLUE*, and from its interpretation in terms of the Mahalanobis norm (2.14), that the ImA difference is the component of ζ orthogonal (in the sense 2.14) to the image space $\Gamma(S)$. That component is discarded in the estimation process, and the *a priori* specified statistics for that component have no impact on either x^a or P^a. As a consequence, any inconsistency between the *a posteriori* observed and the predicted statistics of the ImA difference can always be explained out as being due to misspecification of quantities which have no influence on either the estimate or the estimated estimation error (what would change the analysis would be to modify the component of $E(\zeta)$ along $\Gamma(S)$, or to modify S in a way which would change the orthogonality and the corresponding projection Γx^a). Another consequence is that consistency between predicted and *a posteriori* statistics of the ImA difference is neither a necessary nor a sufficient condition for optimality of a linear estimation system, and that there is no way, on the basis of only statistics of the ImA difference (or for that matter of statistics of the innovation vector d), to objectively determine whether such a system is optimal or not.

Independent additional hypotheses are necessary in order to be able to draw conclusions from the ImA difference or from the innovation vector. An example is given by the result of *Mehra* (1970), which states that the covariance matrices of the model and observation errors can be determined in a Kalman filter from an infinite sequence of observations under a condition which is essentially a condition of observability. But that result is valid only under the *a priori* hypothesis that all errors are uncorrelated in time. That hypothesis cannot be checked independently. Another example is given by the approach followed by *Hoang et al.* (1997), who determine the gain matrix of a stationary Kalman filter, independently of any quantitative hypothesis on the model or observation errors, as the matrix which *a posteriori* minimizes the amplitude of the innovation vector. In addition to the fact that it is not clear what this process can optimize beyond the estimation of the observed parameters, that approach requires the various errors to be uncorrelated in time, at least over infinitely long time lags.

Studies performed a few years ago on what was then the Optimal Interpolation system in operational use at ECMWF showed that retaining some data from the analysis statistically improved the subsequent forecast (Kelly, *pers. com.*). That was certainly a sign that the OI system was not optimal, under the hypothesis however that the verifying observations were both unbiased and affected with errors uncorrelated with errors affecting the information used in the analysis. It must be stressed that this kind of diagnostics can show an estimation system is not optimal, but cannot show it is optimal. More recent studies, performed on the 3D-Var analysis system of ECMWF (*Kelly*, 1997), have not shown any symptom of sub-optimality similar to the ones observed on the OI system.

The conclusion is that *a posteriori* validation of an estimation system is impossible without *a priori,* unverifiable, hypotheses. This raises the obvious following questions. In meteorological applications, what *a priori* hypotheses can legitimately be made, concerning for instance the absence of correlation between different types of errors ? And, once those hypotheses are made, what can be objectively determined, concerning the statistics of the error vector ζ, from the available observations ? These fundamental questions will not be discussed here.

A particular simple and significant diagnostic on the ImA difference is the value of the objective function (2.13) at its minimum, *viz.*,

$$J_{min} \equiv J(x^a) = (1/2)\,[\,\Gamma x^a - z\,]^T S^{-1} [\,\Gamma x^a - z\,]$$

J_{min} is the Mahalanobis norm (2.14) of the ImA difference. It also reads in the background-innovation representation (2.8-11) (*Bennett et al.*, 1993)

$$J_{min} = (1/2)\, d^T [HP^bH^T + R]^{-1} d$$

i. e., J_{min} is also the Mahalanobis norm of the innovation vector, but for the norm associated with its own covariance matrix. One consequence of this fact is that, on statistical average

$$E[J_{min}] = p\,/\,2 \qquad (3.5)$$

where p is, as before, the dimension of d. If the error ζ is in addition assumed to be gaussian, the variance of J_{min} is equal to

$$\mathrm{Var}[J_{min}] = p\,/\,2 \qquad (3.6)$$

In meteorological applications, p lies at the very least in the range 10^3-10^4 (it lies in the range 10^5-10^6 in the present 4D-Var system of ECMWF), so that the fluctuations of J_{min} will be negligible in comparison with its expectation $E[J_{min}]$. This is likely to remain true even if the information error is not gaussian. In the present ECMWF 4D-Var system, J_{min} is too small by a factor of about 2-3 (*Rabier, Järvinen, pers. com.*). This means that the amplitude of the innovation vector is *a priori* largely overestimated.

The objective function J will normally be the sum of a number of independent terms, such as the J_b, J_o and J_c terms of standard 4D-Var

$$J(\xi) = \sum_{j=1}^{k} J_j(\xi)$$

where

$$J_j(\xi) = (1/2)\ [\Gamma_j\xi - z_j]^T S_j^{-1} [\Gamma_j\xi - z_j]$$

In these equations, z_j is an m_j-dimensional component of z ($\Sigma_j m_j = m$), and the rest of the notations is obvious. It is possible to show that the expectation of the j-th term at the minimum is

$$E[J_j(x^a)] = (1/2)\ [m_j - \mathrm{tr}\,(\Gamma_j^T S_j^{-1} \Gamma_j P^a)] \tag{3.7}$$

Although the trace on the right-hand side may be difficult to calculate explicitly, especially for large values of m_j, this equation provides an additional potentially useful diagnostic.

4. THE CASE OF STRONG-CONSTRAINT VARIATIONAL ASSIMILATION

In the case of strong-constraint variational assimilation, a very useful diagnostic is the difference between the minimizing solution and the observations, which is a sub-component of the whole ImA difference. A perfect model will fit observations to within observational error. Therefore, if statistics of all errors other than model errors are properly specified, any observation-minus-minimizing-solution difference (or OmMS difference) that is statisticaly and significantly larger than the corresponding observation error will necessarily be the signature of model error. This can provide a very powerful quantitative diagnostic.

The temporal variation of the OmMS difference can also be very useful. If the model error, accumulated over the assimilation window, is significantly larger than the observation errors, one can expect the closest fit of a model solution to the observations (which is what strong-constraint variational assimilation produces) to be closest to the observations at about the mid-point of the assimilation window, and farthest from the observations at both ends of the window. More precisely, let us consider a one-dimensional perfect model of the form

$$\frac{dx}{dt} = \gamma x \tag{3.8}$$

where $\gamma = \alpha + i\beta$ is a complex constant. This model is used to variationally assimilate observations contaminated by noise with variance s. The variance of the estimation error, which is the statistical mean of the squared difference between two exact solutions of (3.8), will be of the form $a\exp(2\alpha t)$, where a is some positive constant. As for the OmMS variance, it will be equal, according to eqs (3.3), to

$$E[(OmMS)^2] = s - a \exp(2\alpha t)$$

Whatever the value of α, the second time derivative of E[(OmMS)^2] will be non-positive. *The fit of a perfect model to noisy observations has a non-positive convexity* (see also *Ménard and Daley*, 1996). It is not clear whether this result, which is exact for a constant-coefficient, one-dimensional model, generally extends to more complex models. But, together with the argument presented at the beginning of this paragraph, it strongly suggests that a *positive* convexity in the temporal variations of (OmMS)^2 is a sign of model error.

Still another potentially useful diagnostic of variational assimilation is provided by the adjoint solution corresponding to the minimizing solution. The adjoint solution, through the adjoint equation, is a linear transform (actually a temporal integral) of the OmMS difference. It does not contain as such anything that is not already contained in the ImA difference. But it consists of the Lagrange multipliers associated with the constraints expressed by the model equations (*Thacker and Long*, 1988, *Talagrand*, 1989). As such, it says how the model equations should be modified (by addition of a forcing term) in order to further decrease the objective function. It is not clear whether much useful information is contained in the Lagrange multipliers in an individual realization of the variational process. But the statistics of the Lagrange multipliers necessarily reflect the statistics of the various errors, and in particular of the model errors. As such, they presumably contain very useful information.

5. CONCLUSIONS

A number of *a posteriori* diagnostics of analysis and assimilation have been presented and discussed. All these diagnostics bear on quantities which are components, or linear transforms, of the information-minus-analysis difference. All these quantities are necessarily computed in a variational algorithm, so that performing the proposed diagnostics only requires accumulation of statistics on already available quantities. Some of the discussed diagnostics are already routinely performed in several places, while others, such as for instance diagnostics on the adjoint solution, are new.

One important conclusion is that interpretation of the proposed diagnostics requires *a priori* hypotheses which cannot as such be objectively validated. An obvious, and perfectly legitimate, hypothesis of this type is that errors in observations performed by different instruments are statistically uncorrelated. But even that simple hypothesis must be used with some care. If the observations are closely located in space or time, the representativeness parts of the corresponding errors will be correlated.

One major source of uncertainty in assimilation of meteorological observations lies in the errors in the assimilating model. It is probably for identifying and quantifying this type of errors that the diagnostics discussed here can be most useful.

6. REFERENCES

Bennett, A F, L M Leslie, C R Hagelberg and P E Powers, 1993: Tropical Cyclone Prediction Using a Barotropic Model Initialized by a Generalized Inverse Method, Mon. Wea. Rev., 121, 1714-1729.

Hoang S, R Baraille, O Talagrand, X Carton and P De Mey, 1997: Adaptive filtering: application to satellite data assimilation in oceanography, Dyn. Atmos. Oceans, 27, 257-281.

Hollingsworth, A, and P Lönnberg, 1989: The Verification of Objective Analyses: Diagnostic of Analysis System Performance, Meteorol. Atmos. Phys., 40, 3-27.

Kelly, G, 1997: Influence of observations on the operational ECMWF system, ECMWF Newsletter, No 76, 2-7.

Mehra, R K, 1970: On the Identification of Variances and Adaptive Kalman Filtering, IEEE Trans. Automat. Contr., AC-15, 175-184.

Ménard, R and R Daley, 1996: The application of Kalman smoother theory to the estimation of 4DVAR error statistics, Tellus, 48A, 221-237.

Talagrand, O., 1989: Four-dimensional variational assimilation, in Proceedings of Seminar on Data assimilation and the use of satellite data, ECMWF, Reading, England, September 1988, 2, 1-30.

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