Internal diagnostics of data assimilation systems

By **O. Talagrand**[1] **and F. Bouttier**[2]

[1] *Laboratoire de Météorologie Dynamique du CNRS, Paris, France*

[2] *European Centre for Medium-range Weather Forecasts, Reading, England*

Assimilation is intended at estimating the state of the atmospheric flow from all the available relevant data. In most present assimilation schemes, the data that are used consist, in addition to the observations proper, of a background estimate of the state of the flow at the beginning of the assimilation period, and of balance constraints describing the approximate geostrophic equilibrium between the mass and velocity fields. Weights are given to the various data which reflect the assumed accuracy of those data, well as the possible correlations between the corresponding errors. The theory of statistical linear estimation, of which most present assimilation algorithms can be described as being particular applications, requires *a priori* explicit specification of the first- and second-order statistical moments of the errors affecting the data. In that theory, the assimilated fields are simply described as a combination of the background on the one hand, and of the innovation vector, i. e. the deviation from the background of the data that come in addition to the background itself, on the other hand.

Two approaches at least can be followed for validating assimilation algorithms. One is to compare the assimilated fields with unbiased independent data, i.e., data which, in addition of not having been used in the assimilation process, are affected by errors that are statistically independent of the errors affecting the data used in the assimilation. That is the only way to objectively assess the quality of an assimilation algorithm, and in particular to objectively compare the performance of two different algorithms.

A second class of diagnostics bears on the data-minus-analysis (DmA) difference, i.e. the difference vector between the data that have been used in the assimilation and the assimilated fields themselves. The theory of statistical linear estimation provides, in addition to the assimilated fields, the first- and second-order statistical moments of the DmA difference. Any disagreement between those theoretical statistics and the *a posteriori* computed statistics of the DmA difference is the sign of some *a priori* mis specification in the statistics of the data errors. Once such a mis specification has been identified, it may be possible to correct it and to improve the quality of the assimilation.

The DmA difference is an invertible linear transform of the innovation vector, so that statistical diagnostics can equivalently be performed on either one of those two sets of values. Depending on the particular algorithm that is evaluated, and on which particular aspect one is more interested in studying, it might be more convenient to perform the diagnostics on the DmA difference or on the innovation vector.

Interpretation of such 'internal' diagnostics nevertheless requires some care. It can be shown (*Talagrand*, 1999) that, from a strict mathematical point of view, consistency between the theoretical and the *a posteriori* computed statistics is neither a necessary nor a sufficient condition for optimality of an assimilation algorithm. Independent hypotheses, which cannot be objectively verified (at least on the basis of either the innovation or DmA difference vector), are always necessary.

In spite of these caveats, internal diagnostics can be very instructive, and have been performed on the variational assimilation system of ECMWF over the period March-May 1999. In the present ECMWF system, the objective function J minimized by the assimilation is the sum of three terms defined by the three basic sets of data, namely the background, the observations proper and the balance constraints, and respectively denoted J_b, J_o and J_c.

A first test was to verify the rather obvious fact that the assimilated fields must fit the various data to within the assumed accuracy of the latter. That condition is verified, although only marginally for some types of observations, by the ECMWF system. A second test bore on the minimum J_{min} of the objective function J. That minimum is a norm of the DmA difference. Expressed as a function of the innovation vector, it is equal to that vector, normalized by its own covariance matrix. As a consequence, its expectation is equal to $p/2$, where p is the dimension of the innovation vector. For the ECMWF system, p varies between about 200,000 and 250,000 at the synoptic times 00:00Z and12:00Z, and between 150,000 and 200,000 at the subsynoptic times 06:00Z and18:00Z. The ratio $r = 2J_{min}/p$ is about .5 at 00:00Z and 12:00Z, and .3 at 06:00Z and 18:00Z. This means that the innovation vector is statistically overestimated (in quadratic norm) by a respective factor of 2 and 3. This result is paradoxical inasmuch the present assimilation system of ECMWF ignores model errors, which must necessarily increase the innovation vector, particularly at the end of the assimilation window. And it is indeed observed that the ratio r becomes closer to 1 as the length of the assimilation window is increased. It is also observed that r is not changed when satellite observations are not used in the assimilation, which means that the overestimation of the innovation vector cannot result from some form of error in the use of those observations.

The most reasonable explanation for the fact that the innovation vector is significantly overestimated is that it is the background error which is overestimated in the first place. It is safer to overestimate, rather than underestimate, the background error (underestimation may lead to a progressive drift of the assimilated fields from the observations). And background error estimates, which are still at present basically independent of the current state of the flow, result from a long process of empirical adjustment, and are likely to have been thus pragmatically defined so as to be safe for any situation. But the price for safety is of course here a degradation of the accuracy of the assimilated fields.

The only difference between synoptic and subsynoptic hours is that radiosonde observations are significantly more numerous at synoptic hours. The amplitude of the innovation vector has been verified to be statistically of the same amplitude at all hours, so that the difference observed in the ratio r between synoptic and subsynoptic hours necessarily results from a difference in the implicitly defined amplitude of the innovation vector. A simple analytical study suggests however that if the observations performed at synoptic hours, in addition to those already performed at subsynoptic hours, bear on different quantities, no impact must be visible on the ratio r. The fact that an impact is visible must therefore mean that the additional observations bear, at least to some extent, on quantities already measured at subsynoptic hours. Further work is clearly necessary on this point.

As said, the assimilation fits the various data to within their assumed accuracy. In the case of the balance constraint expressed by the J_c term, the fit is very close. The squared deviation of the assimilated fields from the balance constraint is typically two orders of magnitude less than the 'variance' implied by the coefficient in front of the J_c term. In addition, the J_c term increases in the course of the minimization, which is started from the background. Both facts together are compatible with the hypothesis that the variance of the deviation from the balance constraint is largely overestimated. Indeed, the coefficient in front of the J_c term has been determined, not on the basis of a realistic estimate of the amplitude of that deviation (which would be the 'error' affecting the balance data), but as the smallest coefficient ensuring that 'unrealistic oscillations' are absent from the assimilated fields. Since that coefficient is inversely proportional to the implied variance, it is not surprising to find that the latter is largely overestimated.

REFERENCE

Talagrand, O, 1999: *A posteriori* evaluation and verification of analysis and assimilation algorithms, Proceedings of Workshop on diagnosis of data assimilation systems, ECMWF, Reading, England, November 1998, 17-28.