Tellus (2005), 57A, 21–34
Printed in UK. All rights reserved

TELLUS

A posteriori validation applied to the 3D-VAR Arpège and Aladin data assimilation systems

By WAFAA SADIKI[1] and CLAUDE FISCHER[2*], [1]*Direction de la Météorologie Nationale, Casablanca, Morocco;* [2]*Météo-France/CNRM/GMAP, 42 avenue Coriolis, 31057 Toulouse, France*

(Manuscript received 13 November 2003; in final form 16 July 2004)

ABSTRACT

In this paper we present results from the application of a posteriori diagnostics on existing data assimilation systems. The systems of interest are the global data assimilation Arpège 3D-VAR of Météo-France, and the limited-area 3D-VAR analysis of Aladin. First, we discuss how the diagnostics can be ported from theory to practical applications, using aggregates of analysis information over time. We then compare the behaviour of the diagnostics in the two different data assimilation systems, with a focus on the properties of the background and observational error variances. Secondly, the a posteriori validation is used for the off-line tuning of two scalar error parameters in the Aladin 3D-VAR. The tuning provides error variances that better fit the statistics of the innovation vector, without loss of quality in the analyses. Finally, the link of this approach with ensemble-based techniques is made. Especially, we propose to couple the a posteriori diagnostics directly with the usual output of ensemble samples, which could be done at no extra cost. If the results are then used to tune error parameters, then an on-line adaptive assimilation system, based on a posteriori considerations, is obtained.

1. Introduction

The evaluation of the quality of an operational type of data assimilation system, using real data as opposed to simulated data, is a matter of permanent and various efforts in the numerical weather prediction community. On the one hand, classical monitoring, cross-validation and inter-comparisons between centres give either general or detailed overviews of the performances of the systems. On the other hand, it is difficult to verify the initial assumptions or simplifications needed to install the system in practice. One typical example is the specification of the error statistics that appear in the optimal control problems. From optimal interpolation (OI; see Gandin, 1963), to variational techniques (3D-VAR or 4D-VAR; see, for instance, Courtier et al., 1994, 1998) and Kalman filtering (Kalman and Bucy, 1961), we require realistic definitions for the error covariance matrices.

A number of approaches are available for estimating the major error covariance matrices, namely those associated with the model forecast information (the background) and those associated with the observations. These methods either rely strongly on the dynamical construction of the errors or on the direct comparison with validating observations. In the first category, we can find the so-called National Meteorological Center (NMC) – now the National Center for Environmental Prediction (NCEP) – method, where the background errors are derived from sets of forecast differences, valid for the same date (Parrish and Derber, 1992). In the second category, the sampling of statistics from sets of binned observation minus forecast differences, and their fit to pre-specified analytical functions can be cited (Hollingsworth and Lönnberg, 1986; Lönnberg and Hollingsworth, 1986). With the increasing computer power, ensemble estimates start to be used, either statically (fixed) inside the assimilation system (Houtekamer et al., 1996; Belo Pereira et al., 2002) or in a cycled mode in the frame of ensemble Kalman filters (Evensen, 1994; Houtekamer and Mitchell, 1998).

Another original approach for error covariance estimation consists of maximum-likelihood methods. One usually starts with an a priori construction of the covariance matrices. Using extra assumptions, the statistical behaviour of a given diagnostic functional, which depends on a set of parameters that control the shape of the covariance model, is known. The covariance parameters are then retrieved as those values that give the maximum likelihood for the functional (Dee, 1995; Wahba et al., 1995).

As described in Desroziers and Ivanov (2001) and Chapnik et al. (2004), a posteriori validation of the assimilation system provides diagnostics that can also be iteratively used to provide covariance parameter estimates, equivalent to maximum-likelihood estimates. A posteriori validation consists of deriving some diagnostic properties, which are verified by the assimilation system, provided the initial assumptions to derive the diagnostic properties are valid. As shown in the core of the paper,

*Corresponding author.
e-mail: claude.fischer@meteo.fr

one such property is the statistical behaviour of the total cost function of a variational assimilation system, after minimization is completed. Provided that the modelled background and observational error covariances correspond to those of the actual sequences of observation minus forecast vectors, the statistical expectation of the minimum of the cost function is proportional to the number of observations (Bennett et al., 1993; Talagrand, 1998). This specific property also has a counterpart in optimal interpolation in the form of a χ^2 law, when error distributions are assumed to be Gaussian (Ménard et al., 2000). Thus, in terms of a posteriori validation, any departure of the diagnostic from its theoretical behaviour indicates a misspecified initial assumption, probably a misspecification in the covariance model. Note that a good match of the a posteriori result with theory does not, reversely, imply correct a priori specifications. Furthermore, as we will show in this paper, it is not in general straightforward to perfectly pinpoint what exactly the initial misspecification is, in a real data environment.

The goal of our work is threefold. First, we derive a set of a posteriori diagnostic functionals (later, $J_{\min}$, S_{b} and S_{o}) along with practical methods for their computation in a real data environment (Sections 2.1–2.4). Mostly, we concentrate on the ergodicity of the diagnostic terms and their evaluation by a Monte Carlo method. These developments partially rely on previous works by Girard (1987), Talagrand (1998) and Desroziers and Ivanov (2001). We then present results for two operational types of data assimilation systems, namely the global Arpège 3D-VAR [formerly Météo-France's operational system, close to the European Centre for Medium-Range Weather Forecasts (ECMWF) version presented in Courtier et al. (1998)] and its limited-area, biperiodic counterpart in the Aladin model [see Radnóti et al. (1995) and Horányi et al. (1996) for a presentation of Aladin, and Široka et al. (2003) for information on the 3D-VAR in Aladin]. The parallel between two distinct, although closely intertwined, systems such as Arpège and Aladin gives an extra, original, view of the results, as shown in Section 2.5.

Secondly, we address the tuning of error covariances using the output a posteriori diagnostics in Aladin (Section 3). We concentrate on two scalar error variance parameters, for background and observation error statistics, respectively. The initial incentive for this part of our study was the early recognition that the Aladin background error variances probably are not very accurately known. Indeed, presently, the background error statistics are obtained from an NMC method. In its traditional setting, this method leads to generally overestimated error variances, especially in datavoid areas (for a short discussion about the limitations of the NMC method, we refer to Sadiki et al., 2000). A similar result was noticed in the frame of Aladin-derived NMC statistics. We refer to the background error statistics derived from the traditional NMC technique as the 'standard NMC' statistics. A modification of the NMC method was however proposed in Široka et al. (2003), which consists of computing the forecast differences at constant lateral boundary conditions. The constant boundary conditions force the sample statistics to maximum variance in an intermediate, mesoscale range, while the largest model scales have smaller error variance. Thus, a scale selective band-pass filter is obtained for the background error term. However, this so-called 'lagged NMC' technique also reduces the amount of variance globally, at all scales, so that one faces the reverse problem from the standard NMC case. Lagged NMC statistics in Aladin very likely underestimate variances.

Thirdly, we show the relationship of the a posteriori approach with ensemble techniques (Section 4). The functionals considered in this paper are derived in practice using a Monte Carlo method, and thus they can be evaluated using the output samples of an ensemble of forecasts and analyses. Thus, a posteriori validation can be a candidate for on-line parameter estimation in an ensemble-based assimilation system. Limitations to our study are discussed in Section 5, while conclusions and operational considerations are drawn in Section 6.

2. Minimum of the cost function and gain matrix products

2.1. Ensemble relationships

We focus on the properties of a given linear estimation problem, such as the least-squares error estimation or the variational algorithm. Following Talagrand (1998), it is always possible to define a complete information vector z of dimension m by an appropriate change of variable in information space

$$z = \begin{pmatrix} x^{\mathrm{b}} \\ y \end{pmatrix} \tag{1}$$

where x^{b} denotes the background vector and y represents the observations. We note that n and p are the lengths of x^{b} and y, respectively, and $m = n + p$. The information vector z and the truth x^{t} are related by the formula

$$z = \Gamma x^{\mathrm{t}} + \epsilon \tag{2}$$

where

$$\Gamma = \begin{pmatrix} I_{n*n} \\ H \end{pmatrix} \tag{3}$$

denotes the complete information operator, I_{n*n} is the Identity operator of rank n, H is the observation operator, and

$$\epsilon = \begin{pmatrix} \epsilon_{\mathrm{b}} \\ \epsilon_{\mathrm{o}} \end{pmatrix} \tag{4}$$

is the vector of errors, known only through its statistical properties.

In a variational formulation, the best linear unbiased estimate (BLUE) is given by the value of the minimum of the cost function defined in information space:

$$J(x) = \frac{1}{2}(\Gamma x - z)^{\mathrm{T}} S^{-1}(\Gamma x - z). \tag{5}$$

Following the developments of Desroziers and Ivanov (2001), we now consider a subpart J_i of the total cost function, $J = \sum_i J_i$, computed over a subpart z_i of the information vector with m_i elements, $m = \sum_i m_i$. The subpart is obtained by projection of the total information vector, $z_i = \Pi_i z$, where Π_i is a projection operator. We also assume that these components have uncorrelated errors. We can then develop J_i

$$J_i(x^{\mathrm{a}}) = \frac{1}{2}\left(\Gamma_i x^{\mathrm{a}} - z_i\right)^{\mathrm{T}} S_i^{-1}\left(\Gamma_i x^{\mathrm{a}} - z_i\right), \tag{6}$$

where $\Gamma_i = \Pi_i \Gamma$ and S_i are the linearized information operator and the error covariance matrix associated with z_i, respectively. x^{a} denotes the analysis, taken as the solution of the minimization. Desroziers and Ivanov have shown that the expectation of $J_i(x^{\mathrm{a}})$ reads

$$E\left[J_i(x^{\mathrm{a}})\right] = \frac{1}{2}\left[m_i - \mathrm{Tr}\left(\Gamma_i P_{\mathrm{a}} \Gamma_i^{\mathrm{T}} S_i^{-1}\right)\right]. \tag{7}$$

A first application of eq. (7) is to compute the expectation of the weighted departure to the observations, which corresponds to the total observation cost function J_{o}, taking $z_i = y$, $m_i = p$, $\Gamma_i = H$ and $S_i = R$. We then have

$$E\left[J_{\mathrm{o}}(x^{\mathrm{a}})\right] = \frac{1}{2}\left[p - \mathrm{Tr}\left(H^{\mathrm{T}} R^{-1} H P_{\mathrm{a}}\right)\right]. \tag{8}$$

Because the gain K verifies $K = P_{\mathrm{a}} H^{\mathrm{T}} R^{-1}$ and $K^{\mathrm{T}} = R^{-1} H P_{\mathrm{a}}$ (P_{a} and R are symmetric), we also have

$$E\left[J_{\mathrm{o}}(x^{\mathrm{a}})\right] = \frac{1}{2}\mathrm{Tr}(I_{p*p} - HK). \tag{9}$$

In a similar way, we can derive an equivalent formula for the weighted departure to the background (the background cost function J_{b}), taking $z = x^{\mathrm{b}}$, $m_i = n$, $\Gamma_i = I_{n*n}$ and $S_i = P_{\mathrm{b}}$. Using the relationship $P_{\mathrm{a}} = P_{\mathrm{b}} - KHP_{\mathrm{b}}$, we have

$$E\left[J_{\mathrm{b}}(x^{\mathrm{a}})\right] = \frac{1}{2}\mathrm{Tr}(KH). \tag{10}$$

As $\mathrm{Tr}(HK) = \mathrm{Tr}(KH)$, we finally write

$$E[J(x^{\mathrm{a}})] = E\left[J_{\mathrm{o}}(x^{\mathrm{a}})\right] + E\left[J_{\mathrm{b}}(x^{\mathrm{a}})\right] = \frac{1}{2}\mathrm{Tr}(I_{p*p}) = \frac{p}{2}. \tag{11}$$

Thus, under the assumption that all the a priori error covariances are properly specified, the statistical expectation of the a posteriori total cost function is simply proportional to the number of observations. In contrast, any deviation from the theoretical value of any a posteriori formula indicates a failure in the a priori error specification. Note that in the case where an a posteriori relationship is found to be verified in practice, no firm conclusion can be drawn as concerns the reliability of the a priori specifications. In the following, we will also refer to eq. (11) in the form $E(J_{\mathrm{min}})/p = 1/2$, and we identify the value of the cost function at the analysis with its minimum value $[J(x^{\mathrm{a}}) = J_{\mathrm{min}}]$.

In Sections 3.1 and 3.2, we will go one step further and try to use the a posteriori diagnostics to calibrate two scalar error variance parameters. Following Desroziers and Ivanov, these two scalar parameters are defined from the previous relationships as

$$S_{\mathrm{o}} = \frac{2E\left[J_{\mathrm{o}}(x^{\mathrm{a}})\right]}{\mathrm{Tr}(I_{p*p} - HK)} \tag{12}$$

$$S_{\mathrm{b}} = \frac{2E\left[J_{\mathrm{b}}(x^{\mathrm{a}})\right]}{\mathrm{Tr}(KH)}. \tag{13}$$

If the assimilation systems were consistent as regards a priori versus a posteriori error statistics, then one would expect

$$S_{\mathrm{o}} = S_{\mathrm{b}} = 1. \tag{14}$$

A first illustration of the behaviour of the Aladin 3D-VAR as regards the diagnostics J_{min}/p, S_{o} and S_{b} is shown in Fig. 1. The diagnostics were computed for a one-month period (November 2000). The complete ratio J_{min}/p is stable, just slightly smaller than its theoretical expectation of 1/2. The two individual coefficients S_{o} and S_{b} however exhibit very different behaviours: S_{o} is slightly below its theoretical value (1.0), with a time mean of 0.86, while S_{b} is clearly above 1.0, with a mean of about 1.67. J_{min}/p and S_{o} are very stable with time. S_{b} has much more variability, as it is more sensitive to small daily changes in the estimate of Tr(*KH*) (see also Section 2.2). The a posteriori validation applied to S_{o} and S_{b} then tells us that the observation errors are slightly overestimated, while the background errors are underestimated. In its present setting, the Aladin 3D-VAR gives too much confidence to the background information according to the a posteriori diagnostics. As explained in the introduction, the underestimation of the variances in the Aladin J_{b} was anticipated, and therefore we believe this result from the a posteriori diagnostic to be valid.

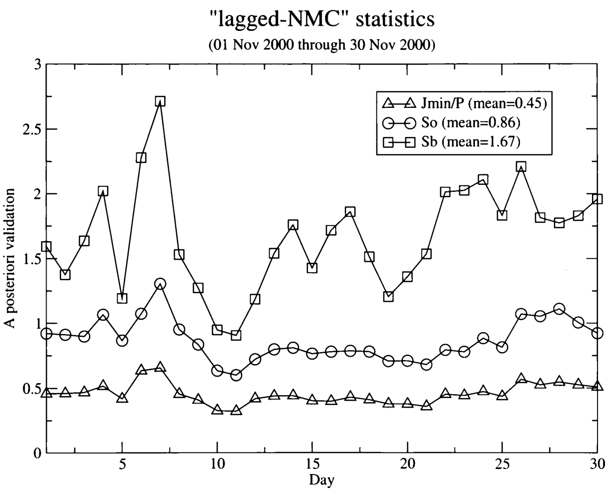


Fig 1. Time series of the total cost function at the minimum, weighted by the number of observations, and the ratios S_{o} and S_{b}, in the limited-area Aladin 3D-VAR over a one-month period. Note that J_{min}/p should be compared to its theoretical value 1/2, while S_{o} and S_{b} have to be compared with 1, by definition from Section 2.1.

2.2. *Practical evaluations*

We now briefly discuss the practical evaluation of the trace operators involved in expressions (9) and (10): $\mathrm{Tr}(I_{p*p} - HK)$ and Tr(*KH*). This estimation is made difficult because the explicit evaluation of the analysis error covariances P_{a} is not available in a variational scheme. A practical method to approximate the trace of the operators is therefore used. It is based on a Monte Carlo method developed by Girard (1987) and requires a random perturbation of the vector of observations. Let $\widetilde{\delta y}$ be a random vector of size *p*, following a normalized Gaussian probability density function, $\widetilde{\delta y} = N(0, I_{p\times p})$. If *p* is sufficiently large, the following relationships hold:

$$\widetilde{\delta y}^{\mathrm{T}} HK\widetilde{\delta y} = \mathrm{Tr}(HK\widetilde{\delta y}\widetilde{\delta y}^{\mathrm{T}}) \approx \mathrm{Tr}(HK). \tag{15}$$

Moreover, the gain matrix verifies

$$H\delta x^{\mathrm{a}}_{(y+\delta y)} - H\delta x^{\mathrm{a}}_{(y)} = HK\delta y = HKR^{1/2}\widetilde{\delta y} \tag{16}$$

where $\delta x^{\mathrm{a}}_{(y+\delta y)}$ and $\delta x^{\mathrm{a}}_{(y)}$ are the analysis increments obtained from perturbed and unperturbed observations, respectively. The perturbed observations are defined by adding the random noise to the original set of observations, with an extra normalization by $R^{-(1/2)}$ needed for the random trace estimate to hold. The approximate estimation of Tr(*HK*) is now obtained by

$$\mathrm{Rand\,Tr}(HK) = (\widetilde{\delta y})^{\mathrm{T}} R^{-(1/2)}\big[H\delta x^{\mathrm{a}}_{(y+\delta y)} - H\delta x^{\mathrm{a}}_{(y)}\big]. \tag{17}$$

Note that a similar approach can be followed when considering the background vector of information, by perturbing x^{b} with a random vector $\widetilde{\delta x^{\mathrm{b}}} = N(0, I_{n\times n})$ of size *n*. Knowing that

$$\delta x^{\mathrm{a}}_{(x^{\mathrm{b}}+\delta x^{\mathrm{b}})} - \delta x^{\mathrm{a}}_{(x^{\mathrm{b}})} = -KH\delta x^{\mathrm{b}} = -KHP_{\mathrm{b}}^{1/2}\widetilde{\delta x^{\mathrm{b}}} \tag{18}$$

we obtain the probabilistic estimate of Tr(*KH*)

$$\mathrm{Rand\,Tr}(KH) = -(\widetilde{\delta x^{\mathrm{b}}})^{\mathrm{T}} P_{\mathrm{b}}^{-(1/2)}\big[\delta x^{\mathrm{a}}_{(x^{\mathrm{b}}+\delta x^{\mathrm{b}})} - \delta x^{\mathrm{a}}_{(x^{\mathrm{b}})}\big], \tag{19}$$

where $\delta x^{\mathrm{a}}_{(x^{\mathrm{b}}+\delta x^{\mathrm{b}})}$ and $\delta x^{\mathrm{a}}_{(x^{\mathrm{b}})}$ are the analysis increments obtained from the perturbed and unperturbed backgrounds, respectively. Thus, the method requires us to run two analyses: the original analysis with the available observations, and a second analysis with the very same set of real observations but randomly perturbed. In this paper, we focus on the use of the observation perturbation, which directly provides an estimate for Tr(*HK*) and $E[J_{\mathrm{o}}(x^{\mathrm{a}})]$ using eq. (9). Because Tr(*HK*) = Tr(*KH*), we also obtain an information on $E[J_{\mathrm{b}}(x^{\mathrm{a}})]$ using eq. (10).

We also stress some practical hints that have arisen in the implementation of the random trace estimation in the real data systems. First, some physical fields, which do have threshold values, had to be filtered out of the procedure. This was the case for relative humidity which is bound by [0, 1]. In general, data that are constrained by physical bounds can slightly corrupt the results, because the Gaussian perturbation can lead to unphysical values. Secondly, Wahba et al. (1995) provide more numerical criteria for the Monte Carlo estimate to be valuable. Let *C* be a symmetric operator of rank *N*, depending on some tunable parameter α and let $\xi = N(0, I_{N\times N})$ be a random vector of size *N*. Then the estimate of the trace of $N^{-1}\,\mathrm{Tr}[C(\alpha)]$ is given by $[N^{-1}\xi^{\mathrm{T}}C(\alpha)\xi]$. The standard deviation of the estimate is

$$\sqrt{\frac{2}{N}}\sqrt{\frac{\mathrm{Tr}(C^2)}{N}}, \tag{20}$$

which indicates that the larger the size of the problem, the better the estimation. As we show below, in practice, the numerical convergence as stated by Wahba et al. was always achieved in our analyses, both in Arpège and Aladin.

Figures 2 and 3 illustrate the results for the trace operators, along with other products output from the variational algorithms (minima of cost functions, number of observations), for the Arpège and Aladin 3D-VAR analyses, respectively. The analyses use all the available sets of observations of these periods: radiosondes, surface observations, aircraft messages, 1D-VAR inverse brightness temperature profiles, a few cloud track derived winds from geostationary satellites. Obviously, the size of the problem is larger in the global 3D-VAR (about 114 000 observations) than in the limited-area version (about 7000 observations). However, we obtain reasonably stable estimates for any analysis date, as regards the Wahba et al. criterion. A slight discrepancy is observed between *p* and its random estimate $\mathrm{Rand}[\mathrm{Tr}(I_{p*p})]$, which is due to some remaining threshold observations.

In both systems, the Tr(*HK*) = Tr(*KH*) trace is roughly one order of magnitude smaller than the number of observations *p*,

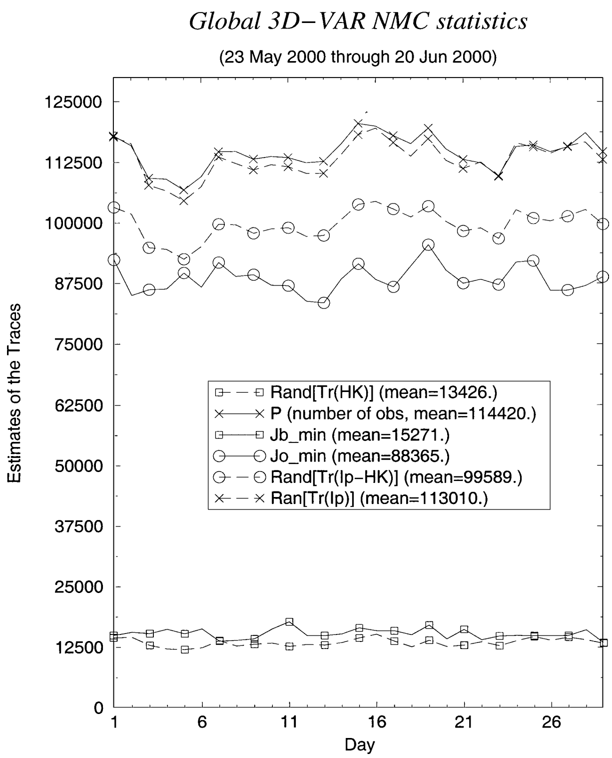


Fig 2. Time series of randomly estimated trace operators, J_{b} and J_{o} at the minimum and total number of observations, in the global Arpège 3D-VAR over a one-month period.

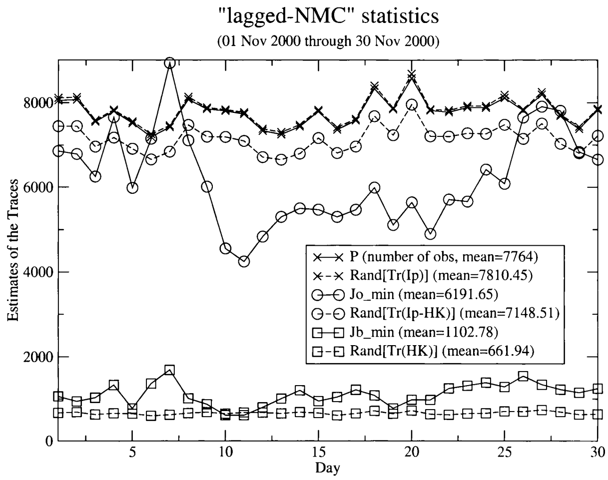


Fig 3. Time series of randomly estimated trace operators, J_{b} and J_{o} at the minimum and total number of observations, in the limited-area Aladin 3D-VAR over a one-month period.

and even smaller when compared to the size of the minimization problem, $n \approx 10^5$. Using again the relation $P_{\mathrm{a}} = P_{\mathrm{b}} - KHP_{\mathrm{b}}$, we can write

$$n = \mathrm{Tr}\big(P_{\mathrm{a}} P_{\mathrm{b}}^{-1}\big) + \mathrm{Tr}(KH). \tag{21}$$

Thus, we see that the information brought by the observations at a given analysis time remains fairly modest, and does not contribute to a large reduction of the analysis error. In addition, Figs. 2 and 3 also provide quick estimates for the S_{o} and S_{b} scalar variance parameters. In Arpège, we obtain $S_{\mathrm{o}} \approx 0.88$ and $S_{\mathrm{b}} \approx 1.13$. In Aladin, we have $S_{\mathrm{o}} \approx 0.86$ and $S_{\mathrm{b}} \approx 1.66$, as already shown in Section 2.1. Thus, the observation term S_{o} appears remarkably consistent between the global and the limited-area systems. This is a reason for confidence because we have indeed used in the limited-area analyses the geographical subset of those observations used in the global system. Thus, the similarity of S_{o} indicates that this term is controlled by the actual properties of the observation statistical parameters (such as their standard deviation) and not by other system-related parameters (resolution of the analysis grid, geometry). In contrast, S_{b} appears more system-dependent. However, while in the Aladin case we can guess the reason for the large value of S_{b}, it is not obvious what causes the $S_{\mathrm{b}} > 1$ value in Arpège. The deviation in Arpège might point to the inherent limitations of the NMC method.

2.3. Ergodic assumption and time-average estimates

In this section, we focus on two very specific problems that occur when the a posteriori probabilistic diagnostics are evaluated in a real data assimilation system. First, the ensemble statistics introduced in Section 2.1 are purely mathematical objects. In practice, we do neither possess the actual 'truth' of the atmosphere, with which to compare the information bits, nor are we able to rerun a large enough number of times exactly the same analysis to produce the true ensemble. Secondly, the evaluation of the traces by the Monte Carlo method requires a large number of observations, as stated in Section 2.2. In real situations, the amount of data is limited for obvious reasons of availability, transmissions, data quality control. Furthermore, it will be interesting to stratify the observational a posteriori information (as shown in Section 2.5) by observation type, observation field or atmospheric layer. This stratification reduces even more the size of the estimation problem. Thus, we face a difficulty both in going from the mathematical to the sampling estimate of the expectation, and in creating a large enough sample to derive fine grain diagnostics.

Classically, in statistical estimation, the mathematical expectation is replaced by a finite estimate on a sample of real data. To make sure that the real estimate converges to the theoretical value, assumptions of stationarity and ergodicity of the sampled signal are needed, and should be verified. In meteorological signals, the sample is either a time series of data or a spatial distribution, or a set of spatial distributions with time. Ergodicity of the signal requires the data to be uncorrelated after a certain finite time range, so that we can safely estimate the mathematical expectation by a time mean. We have therefore defined the autocovariance with time for a given signal, for instance the ratio $J_{\min}/p$

$$\mathrm{autocov}(j) = \frac{1}{N_j}\sum_{i=1}^{N_j}\left[\left(\frac{J_{\min}}{p}\right)_i - \overline{\left(\frac{J_{\min}}{p}\right)}\right] \times \left[\left(\frac{J_{\min}}{p}\right)_{i+j} - \overline{\left(\frac{J_{\min}}{p}\right)}\right], \tag{22}$$

where $\overline{(J_{\min}/p)}$ denotes the time mean. In a similar way, we define the autocovariances with time for the signals S_{o} and S_{b}, as defined by eqs. (12) and (13). We have computed the autocovariances of the three signals over a two-month period (November/December 2000), so that in the above expression j can take values in the range $j = 1, 59$ and each time gap N_j is in the range $N_j = 60 - j$. Figure 4 shows the associated time correlations (autocov(j)/autocov(0)) for S_{o} and S_{b}. The time correlations quickly drop to zero, with no correlation after 4/5 d. After 10/12 d, the curves become noisy and uncertain due to the small number of values in the sample. For time delays of 4–10 d, the time correlations are virtually zero. Similar results were obtained for the ratio $J_{\min}/p$ (not shown). These figures are obtained from the limited-area Aladin 3D-VAR, so that we suspect the 5-d decorrelation time to be a result both from the meteorological variability over Western Europe and from the propagation time of a synoptic system inside the Aladin/France domain (whose size is about 3000×3000 km^2).

Furthermore, we have verified that the estimates of S_{o} and S_{b} are very stable, for any starting date (separated by 20 d) and any time length (5, 10, 15 and 20 d). Thus, the signals for S_{o} and S_{b} are stationary and we have decided to compute the ensemble estimates over a 30-d period, which is significantly larger than the time decorrelation interval (5 d).

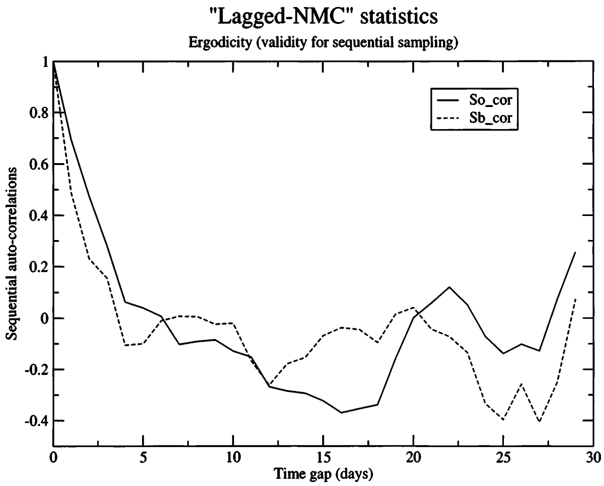


Fig 4. Sequential autocorrelations of the ratios S_{o} and S_{b} computed in the limited-area Aladin 3D-VAR over a two-month period, as a function of the time gap T.

2.4. Observation aggregate method

The second practical problem encountered concerns the size of the samples. First, although the number of observations p in Aladin is large enough to provide a reliable estimate with the Monte Carlo method, this number drops dangerously when we try to apply the projection operator Π_i from Section 2.1. In observation space, this projection is simple and means to retain only specific contributions in the Monte Carlo estimates of the traces (for instance, only temperatures from radiosondes in a given atmospheric layer). The stratification of the diagnostics reduces the size of the information vector. In Aladin, only a few tens or hundreds of data bits remain for radiosonde data. Even in Arpège, this number can become small especially when both vertical and horizontal selections are performed (for instance, to separate the Northern and Southern Hemispheres and the tropical latitudes).

However, we have already noticed that the signals of main concern, i.e. the a posteriori diagnostics given by eqs. (11)–(13), verify the ergodic assumption with a time decorrelation scale of about 5 d. Furthermore, the Monte Carlo estimates give directly the estimates of the traces, in a fully additive way.

Thus, if we consider a series of operators of the type H_jK_j, valid for different analysis dates j, we may generalize the Monte Carlo estimate to the associated series of trace operators, and estimate the total trace (defined as the total sum of all traces) by simply adding all the individual random estimates to one another. This result strictly only demands that the random input vectors for each analysis date be independent from one another.

Taking advantage of the good stability with time of all the a posteriori functions at stake [ergodic behaviour of the ratios and reasonable time invariance of the individual $\mathrm{Tr}(H_jK_j)$ values], we can assume that for any subpart z_i of the information vector, as defined in Section 2.1, the a posteriori ratio S_{o}^i can be estimated by an aggregate of the form

$$S_{\mathrm{o}}^i = \frac{\sum_{j=1}^{M} E\left[2J_{\mathrm{o}}^i(j)\right]}{\left(\sum_{j=1}^{M} p_j^i\right) - \mathrm{Rand}\,\mathrm{Tr}(D)}, \tag{23}$$

where D is the block-diagonal operator with all the individual H_jK_j aggregated, and j is the analysis date index. In eq. (23), we consider a set of M distinct analyses, for example 30 successive 3D-VAR analyses. We assume the searched ratio S_{o}^i to be stable during this period, and we obviously require the information subpart z_i under concern to be present with approximately the same properties for all the M analysis dates. In particular, the stability of the estimate in eq. (23) requires the observation set to be present with similar density and positions. As a counter-example, one can clearly not consider a specific satellite beam, over a precise geographical band. These data would be very irregularly present. An appropriate example is provided by a fixed network of radiosonde data, which is regularly available for all synoptic dates. Again, eq. (23) is not exact in theory. We have further verified the good approximation of the aggregate estimate of eq. (23) by recomputing the total ratios S_{b} and S_{o}, when all the observations are retained. The aggregate approach, compared with the exact technique suggested by eq. (12), implies the reversion of the order of the operations. First, we sum the individual terms over all analysis dates j [traces, $J_{\mathrm{b}}^i(x^{\mathrm{a}})$, $J_{\mathrm{o}}^i(x^{\mathrm{a}})$, p_i], and only then is the full ratio computed. We obtain $\widetilde{S_{\mathrm{o}}} = 0.86$ and $\widetilde{S_{\mathrm{b}}} = 1.66$, to compare with 0.86 and 1.67 obtained by the time average from Fig. 1.

The practical implementation of eq. (23) is straightforward, once the general tools for the estimation of the traces and other analysis products have been prepared. We refer to this technique as the 'observation aggregate' method.

2.5. Properties with respect to the observations

In this section, we present the a posteriori S_{o} ratios, as obtained with the aggregate observation technique. Both the global and the limited-area 3D-VAR systems are discussed. In the global system, the evaluation was performed over a one-month period at a time when the 3D-VAR Arpège analysis actually was operationally run at Météo-France. The Aladin statistics were obtained from a two-month period (November/December 2000), in order to increase the size of the sample.

First, we discuss the Arpège results. Table 1 shows the ratios S_{o} for 10-m wind components extracted from the SYNOP reports. These values are shown for a more general purpose, because the surface fields themselves are not variationally assimilated. Surface data are fitted after the variational analysis of the three-dimensional atmospheric fields, in a step of optimal interpolation which uses the three-dimensional analysis as first guess. Therefore, at the time of the 3D-VAR analysis, the surface data are not yet fitted, and they only enter the system as constraints

Table 1. Ratios S_o for SYNOP surface observations estimated by the observation aggregate method. The values are computed from the Arpège 3D-VAR over the period 23 May 2000 to 20 June 2000 (29 d)

Field	Number of observations	Ratio S_o
Zonal wind	16 179	1.00
Meridional wind	16 179	0.98

Table 2. Ratios S_o for AIREP, AMDAR, ACARS temperature aircraft data estimated by the observation aggregate method. The values are computed from the Arpège 3D-VAR over the period 23 May 2000 to 20 June 2000 (29 d)

Pressure level (hPa)	Number of observations	Ratio S_o
1000–950	5336	0.72
950–900	7427	0.58
900–800	10 803	0.64
800–700	9169	0.69
700–500	21 683	0.56
500–400	14 651	0.53
400–300	29 929	0.46
300–250	48 817	0.49
250–200	67 974	0.56
200–150	21 086	0.72

Table 3. Ratios S_o for AIREP, AMDAR, ACARS zonal and meridional wind aircraft data estimated by the observation aggregate method. The values are computed from the Arpège 3D-VAR over the period 23 May 2000 to 20 June 2000 (29 d)

Pressure level (hPa)	Number of observations	Ratio S_o for U	Ratio S_o for V
1000–950	4661	0.90	0.92
950–900	6449	1.08	1.06
900–800	9318	1.06	1.04
800–700	8035	0.84	0.90
700–500	18 124	0.74	0.79
500–400	12 735	0.66	0.71
400–300	26 052	0.67	0.74
300–250	44 262	0.76	0.83
250–200	64 217	0.76	0.81
200–150	20 336	0.83	0.85

in the observation operators: 10-m wind can constrain the wind profile in the planetary boundary layer observation operator and thus affect the lowest model levels. The S_o ratios are fairly high, compared to other data, and indicate that these data indeed are not analysed during 3D-VAR, although S_o are 'by chance' close to 1.

Tables 2 and 3 give the S_o ratios for data extracted from aircraft flight track samples, for temperature and wind components, respectively. The denomination 'AIREP' covers both actual AIREP, AMDAR and ACARS reports, whenever they are retained after the quality control step. These data cover the atmosphere in a rather heterogeneous manner, with many points in the low atmosphere, concentrated around airport locations, and along the typical mid- or long-distance flight tracks. However, with the high number of measurements captured by the aggregate technique, stable and 'realistic' profiles of S_o ratios are retrieved. The temperature S_o profiles in Table 2 indicate rather small values, between 0.5 and 0.7, which could indicate that temperature standard deviations from AIREP reports could be overestimated. Wind S_o profiles in Table 3 show that the corresponding standard deviations are fairly well prescribed, in the sense of the a posteriori diagnostic. There is however a slight bias between the zonal and meridional wind component S_o sets: the meridional estimates are larger for all layers above 800 hPa. This indicates that the error standard deviations should be differentiated, with a slightly smaller value for the zonal component. It is otherwise quite difficult to trace back the possible reason for this differentiation.

Tables 4 and 5 show the S_o profiles for data extracted from TEMP reports, for temperature and wind components, respectively. We have obtained fairly high values for the temperature S_o estimates, with many values around 0.8–1.2 and more. Compared to other S_o profiles, these values are significantly larger. In the sense of a posteriori validation, Table 4 indicates that the standard deviations of TEMP temperature profiles should be reviewed, with larger values in the lower layers and above the tropopause, and a decrease in the middle troposphere. Table 5 shows that the wind S_o profiles are closer to the generally obtained S_o values for other fields and reports, with values around 0.7–0.9. The estimates could indicate a slight differentiation of the lower atmospheric layers, where S_o are largest (perhaps an effect of the greater uncertainty of wind observations in the boundary layer, due to various sorts of surface forcings). Unlike for the AIREP data, there is no significant difference between zonal and meridional components for TEMP data.

Table 4. Ratios S_o for TEMP temperature radiosonde data estimated by the observation aggregate method. The values are computed from the Arpège 3D-VAR over the period 23 May 2000 to 20 June 2000 (29 d)

Pressure level (hPa)	Number of observations	Ratio S_o
1000	13 440	1.49
850	27 997	1.25
700	40 645	0.88
500	36 466	0.67
400	24 044	0.66
300	16 770	0.83
250	12 846	1.30
200	15 294	1.93
150	16 158	1.64

Table 5. Ratios S_o for TEMP zonal and meridional wind radiosonde data estimated by the observation aggregate method. The values are computed from the Arpège 3D-VAR over the period 23 May 2000 to 20 June 2000 (29 d)

Pressure level (hPa)	Number of observations	Ratio S_o for U	Ratio S_o for V
1000	26 174	1.00	1.00
850	33 778	0.98	0.88
700	35 478	0.94	0.88
500	29 041	0.76	0.76
400	25 114	0.76	0.71
300	24 486	0.74	0.69
250	24 728	0.72	0.74
200	28 618	0.71	0.72
150	31 597	0.72	0.74

From a more general perspective, the global 3D-VAR estimates show the following behaviours.

(1) For data actually used and fitted in the variational formulation, the S_o estimates usually are around 0.7–0.9. These values appear to be the most common, and, under the assumption that the Arpège system is reasonably well tuned by the experience of routine model and data monitoring, these values also can be considered as typical for our system. Therefore, the straight application of the a posteriori validation, which would force us to retune the whole system, has not been retained for Arpège.

(2) The observation aggregate method, along with the flexibility of the general a posteriori diagnostic formulation in observation space (with the possibility to perform spatial stratifications), leads to 'realistic' vertical profiles of S_o. The profiles are not noisy, but smooth, and could actually be supported by some meteorological understanding. However, although we have tentatively tried some interpretations, we believe that it is not worth going too far in the explanations. It is actually difficult to compare the results from a posteriori diagnostics with independent information, mostly provided by the daily monitoring of the operational assimilation suite.

(3) As expected from theory, S_o values for data that enter the variational analysis in a very marginal manner (only as constraints in observation operators of other data) do not follow the general behaviour because no direct use of the error statistics is made. This holds for surface data.

Finally, we present also S_o estimates from the limited-area 3D-VAR analysis. The results are displayed in Fig. 5, for SYNOP, AIREP and TEMP reports (over the Aladin/France domain). Note that these data were not stratified into vertical profiles, only overall values are available per field and observation type. First, it is striking how well these S_o values compare with those from the global system. Although the data are extracted from the same type of messages, one could expect different behaviours due to different model geometries and resolutions. The similarity of the limited-area and global S_o indicates that representativeness error is not a dominant feature in our diagnostic. Furthermore, the Aladin 3D-VAR analyses are not embedded in an assimilation cycle. Despite these structural differences, the S_o estimates are close to the global ones, slightly smaller in general with values rather around 0.6–0.75. We recover values around 0.75 for the TEMP wind S_o, and fairly large values for the TEMP temperature S_o (around 1.25). The AIREP S_o also are similar to the Arpège estimates: around 0.6 for temperature, around 0.75 for zonal wind and 0.8 for meridional wind. SYNOP wind S_o are however much smaller for Aladin than for Arpège, but again these values cannot be diagnosed under the assumptions for a posteriori validation. In Section 3.1, we use another S_o estimate, valid for all observations in general, which is then considered as a tuning factor for a scalar standard deviation correction in the 3D-VAR analyses. Following the results developed in Section 2.1, a finer retuning either by observation type (following the values displayed in Fig. 5) or by vertical levels could have been possible as well.

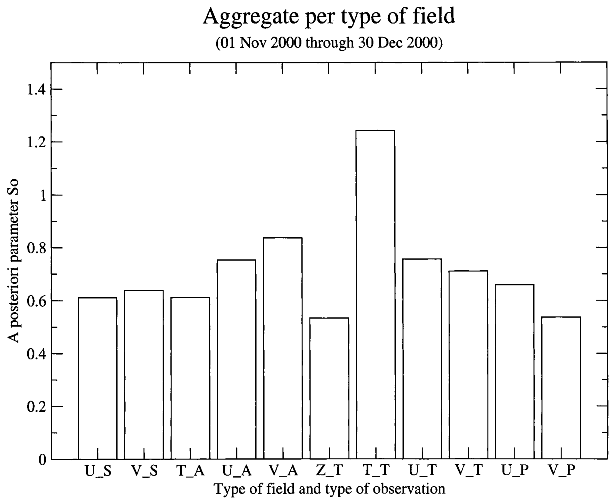


Fig 5. Evaluations of the ratio S_o on aggregated data sets, for different observational data types and fields (limited-area Aladin 3D-VAR). The acronyms denote: U_S, zonal wind of SYNOP; V_S, meridional wind of SYNOP; T_A, temperature of AIREP; U_A, zonal wind of AIREP; V_A, meridional wind of AIREP; Z_T, geopotential of TEMP; T_T, temperature of TEMP; U_T, zonal wind of TEMP; V_T, meridional wind of TEMP; U_P, zonal wind of PILOT; V_P, meridional wind of PILOT.

3. Application to the Aladin data assimilation system

3.1. Tuning of scalar error parameters

We have shown in Section 2 how to evaluate some a posteriori diagnostic outputs from an operational type of variational assimilation system. In Section 2.5, we have obtained very stable practical results, which give us confidence in the reliability of the diagnostics derived from aggregated analysis outputs as introduced in Section 2.4. A further step now consists of reintroducing the a posteriori results into the assimilation system, in order to modify its behaviour. To be specific, we follow the spirit of Desroziers and Ivanov (2001) and Chapnik et al. (2004) who propose an iterative method where the ratios S_o and S_b are reinjected into the analysis as scaling factors in front of the error covariance matrices. The new analysis then in turn produces a set of diagnostics, which are again reinjected in the analysis and so on. They show that this successive updating forces the analysis to converge towards the theoretical a posteriori result. Eventually, limit values for the ratios S_o and S_b are obtained. Chapnik et al. (2004) furthermore demonstrate that the iterative updates are equivalent to a maximum-likelihood estimate of S_o and S_b. However, they also give some limitations to their method, especially the need for a large enough amount of observations (an issue addressed here by the observation aggregate technique), the simultaneous evaluation of both ratios S_o and S_b (as opposed to a separate evaluation, with one ratio fixed and the other adaptive), and the spurious contamination of the tuned ratios by other shortcomings in the modellization of the error covariance matrices. In particular, the results of the tuning are less reliable when the modelled correlation lengths prescribed in the P_b matrix are shorter than the actually 'observed' correlation lengths from the innovation vector $y - H(x^b)$.

As described in the introduction, we strongly suspect the raw error variances specified in the Aladin 'lagged NMC' P_b matrix to be an underestimation of realistic values. In contrast, the 'standard NMC' evaluation probably overestimates the error variances. This behaviour is also suggested by the ratios S_b and S_o obtained in Aladin. Figure 1 indicates that the lagged background error variances are underestimated in the sense of the a posteriori diagnostics, because globally $S_b > 1$. In Fig. 6, however, we always have $S_b < 1$, which indicates overestimated variances by the standard NMC method. Note that the observation error variances behave in a more similar manner, whatever system is considered: as a mean, $S_o \approx 0.86$ when lagged statistics are used, while $S_o \approx 0.60$ for standard statistics. We retrieve a more stable behaviour of the observation term, and most of the variability by changing the background error matrix is supported by the S_b ratio. Therefore, we do have a natural testbed for the practical application of the variance parameter tuning by an a posteriori approach.

However, unlike Desroziers and Ivanov (2001) and Chapnik et al. (2004), we do not consider an iterative method, but we focus on a simpler 'one-step' procedure. The first set of ratios S_b and S_o obtained as time-averaged values are used to rescale the columns of the P_b and R matrices, respectively, and a second set of 'tuned' analyses is then performed, over the whole period of interest (November 2000). Because a single scalar rescaling is performed, the ratios S_b and S_o are called 'scalar error parameters'. The retained parameters are $S_b = 1.67$ and $S_o = 0.86$ for lagged NMC statistics, and $S_b = 0.53$ and $S_o = 0.60$ for standard NMC statistics. We then recompute the a posteriori diagnostics (J_{min} here) for the tuned analyses. Figure 7 shows the results for J_{min} for the original and new analyses, and for both lagged and standard statistics. For standard statistics, the tuned curve is clearly shifted towards the theoretical mean, starting from initially very low values. For lagged statistics, where the initial curve already seems reasonably close to the theoretical mean, the tuned curve is only slightly displaced. As a whole,

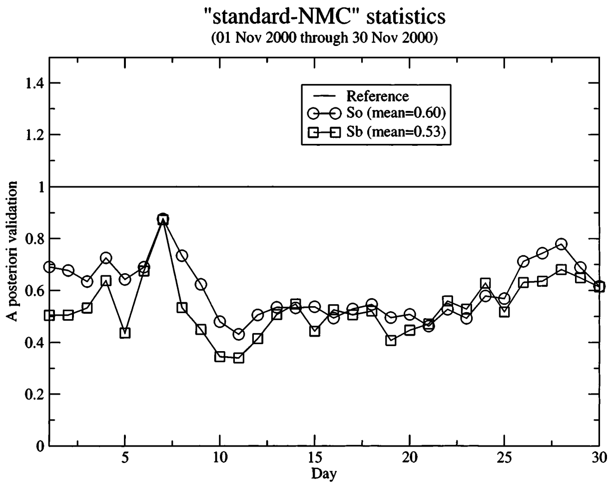


Fig 6. Time series of the total cost function at the minimum, weighted by the number of observations, and the ratios S_o and S_b, in the limited-area Aladin 3D-VAR but for a so-called 'standard NMC' J_b, over a one-month period. These curves have to be compared with those of Fig. 1 where the background error statistics are different (so-called 'lagged NMC' J_b).

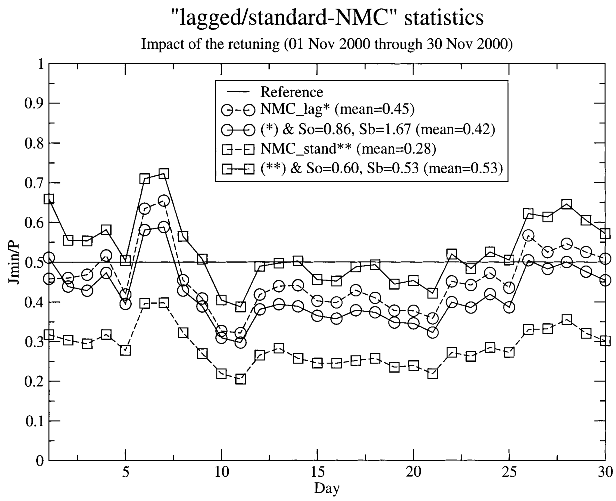


Fig 7. Time series of the total cost function at the minimum, weighted by the number of observations, after the limited-area 3D-VAR is rerun with tuned background and observational standard deviations. Both the 'lagged' and 'standard' NMC J_b cases are depicted.

we have verified that the rescaling by the scalar error parameters indeed forces the new analyses to better verify the a posteriori theoretical values, for the parameters $J_{\min}$, S_b and S_o.

3.2. *Impacts on the performances of the analysis*

We have first verified that the rescaled analyses produce solutions that are quite close to the original 3D-VAR increments. When looking at individual charts, the analysis solutions look very similar, in intensity and structure, which points to the dominating effect of the observations as strong constraints in the minimization problem (not shown). As a further validation, we have computed some classical monitoring scores for the different sets of Aladin 3D-VAR analyses: original 3D-VAR with lagged and standard P_b matrix formulations, and rescaled 3D-VAR with lagged and standard matrices. The period of validation is again November 2000. We stress that the analyses are not cycled: each analysis uses as first-guess field a 6-h Aladin forecast started from a global Arpège analysis, not a regional LAM analysis. Thus, we measure the impact of the rescaling and the individual 3D-VAR analyses, but not the cumulative effect inside an assimilation cycle. Figure 8 shows that, as expected from an optimal control system, the four analyses are closer to the observations (here, zonal wind profiles from radiosondes) than the first guesses (6-h forecast). For the lagged P_b matrix, the analyses are closer to the observations when the rescaled variances are used. This result is expected because the rescaling acts in the direction of amplifying the background error variances (multiplication by $S_b = 1.67$) while diminishing the observation error variances (multiplied by $S_o = 0.86$). For the standard P_b matrix, where both background and observation error variances are reduced, Fig. 8 indicates that the analyses are slightly further away from the radiosonde observations after rescaling.

We have furthermore computed 12-h forecasts started from each individual analysis, and computed the scores with respect to the radiosonde data for 6 and 12 h of integration. The forecasts are run without numerical initialization (by digital filters), but are otherwise close to operational Aladin forecasts. Figure 9 shows the results for the 12-h forecasts compared to zonal wind profiles. Figure 9 also includes the scores for the model integrations started from the first-guess data. There remains a clear benefit from first performing the 3D-VAR analysis in Aladin, and it is reassuring to observe that the benefit from the 3D-VAR still is perceptible at 12 h, and secondly rescaling in the case of the lagged statistics. At 6 h of integration, the beneficial signal from both 3D-VAR and rescaling is even larger (not shown).

Thus, the scalar error variance rescaling, based on the reuse of the a posteriori information, is consistent with the overall expected performance of a real assimilation device. In particular, the rescaling does not lead to non-meteorological or absurd solutions in the analysis, nor does it lead to deteriorated scores.

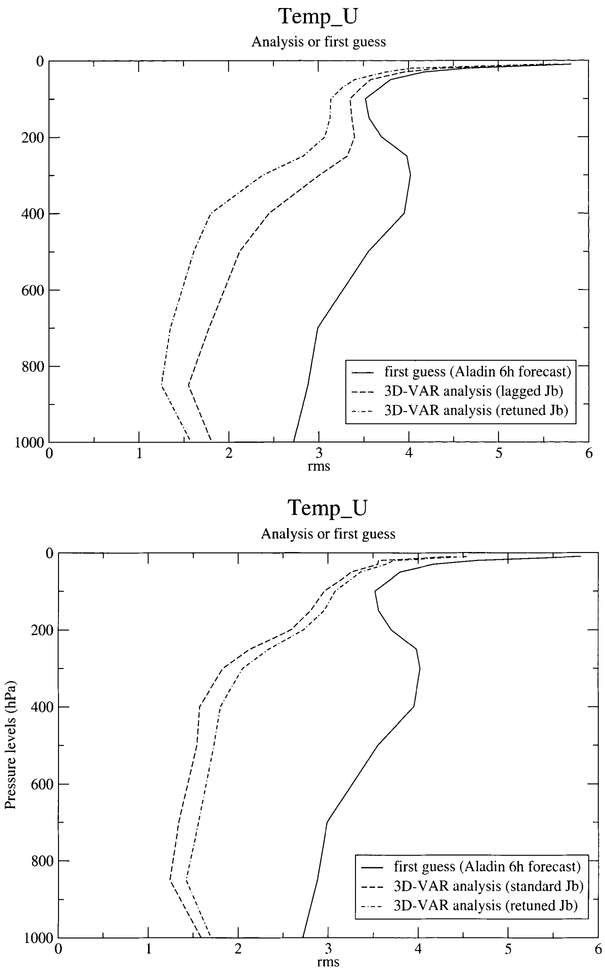
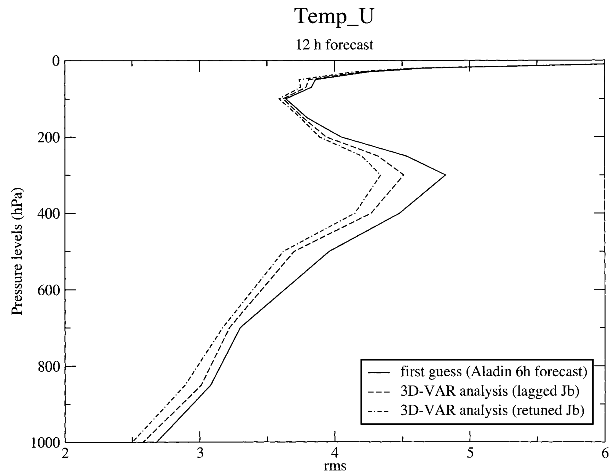


Fig 8. Root mean square of analysis zonal wind field minus TEMP radiosonde data, computed over the one-month period. The retuned analyses have been performed with $S_b = 1.67$ and $S_o = 0.86$ for lagged NMC statistics (top panel) in P_b, and $S_b = 0.53$ and $S_o = 0.60$ for standard NMC statistics (bottom panel), respectively.

4. A posteriori diagnoses in an ensemble framework

The fact that the practical estimation of the (S_b, S_o) diagnostics is performed with a Monte Carlo estimate of some trace operators suggests a closer comparison with the general ensemble-based computations of error statistics. In particular, ensemble analyses are becoming common methods in order to evaluate background error covariances (Houtekamer et al., 1996; Belo Pereira et al., 2002; Mike Fisher, ECMWF, private communication). The ensemble of analyses is usually obtained by running a number of assimilation cycles which differ by the inclusion of perturbations in the first guesses, the observations and the model parameters. Typically, 30 d and 10 analyses per day are considered. The



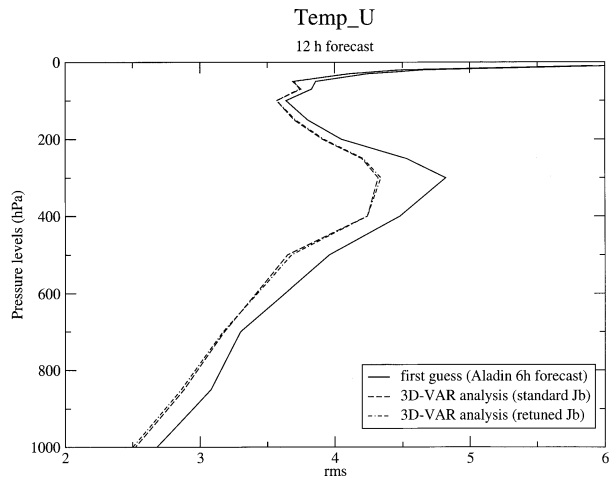


Fig 9. Root mean square of 12-h forecast zonal wind field minus TEMP radiosonde data, computed over the one-month period. The retuned analyses have been performed with $S_b = 1.67$ and $S_o = 0.86$ for lagged NMC statistics (top panel) in P_b, and $S_b = 0.53$ and $S_o = 0.60$ for standard NMC statistics (bottom panel), respectively.

analysis error covariances are then estimated as the mean overall differences:

$$P_a = E[\Delta x^a (\Delta x^a)^T]. \tag{24}$$

Δ represents here the difference between two members of the ensemble and the statistical expectation $E[\,]$ denotes actually sampled statistics, for simplicity. Each analysis departure is added to the control, and a set of forecasts is performed, possibly with some noise process on model parameters. The differences between the forecasts (Δx^b) are then sampled to provide an ensemble estimate of the a priori background error statistics P_b. Furthermore, the forecasts provide a set of background fields for the next set of analyses, possibly using also perturbed observations. If, at first, we assume no perturbation on the observations, and using the same notations as in Section 2.1, we have

$$\Delta x^a = x^a_{(x^b + \Delta x^b)} - x^a_{(x^b)} \tag{25}$$

with $\Delta x^b = N(0, P_b)$ by construction. Introducing the Kalman gain K yields

$$\Delta x^a = K[y - H(x^b) - H\Delta x^b] - K[y - H(x^b)] \tag{26}$$

$$= -KH\Delta x^b \tag{27}$$

$$= -KHP_b^{1/2}\widetilde{\Delta x^b} \tag{28}$$

where $\widetilde{\Delta x^b} = N(0, I_{n\times n})$ is normalized. Thus, any Δx^a from the ensemble can stand for the analysis difference formally needed to derive the probabilistic traces (compare the above formula with eq. 18). By substitution of Δx^a into eq. (19), we therefore have

$$\begin{aligned} \text{Rand Tr}(KH) &= -(\widetilde{\Delta x^b})^T P_b^{-(1/2)} \Delta x^a \\ &= -\Delta x^{b^T} P_b^{-1} \Delta x^a \end{aligned} \tag{29}$$

and eq. (13) yields

$$S_b = \frac{2E\left[J_b(x^a)\right]}{-\Delta x^{b^T} P_b^{-1} \Delta x^a} \tag{30}$$

which corresponds to an 'ensemble analysis' estimate of S_b. Thus, at the same time as the differences of ensemble members are sampled to compute the a priori error statistics for the next assimilation cycle, we could also randomly evaluate the traces. Then, an ensemble-derived estimate of S_b is available. From scratch, there is no reason for the a posteriori relationship to be valid (for instance, $S_b \approx 1$). The ensemble S_b could then indicate whether the ensemble spread should be inflated or deflated, from an a posteriori point of view. Further work will be devoted to the application of this proposal in the future.

5. Limitations

A specific example of the difficulties encountered with the (S_b, S_o) interpretation is given in the frame of our work with the Aladin 3D-VAR. Indeed, an interesting point there was to compare the parameters for two different specifications of the P_b matrix. However, these two statistics differ not only by their global error variance, but also by their typical horizontal correlation length-scales. We illustrate this aspect in Figs. 10 and 11 where we have plotted several horizontal autocovariance profiles: the lagged and standard P_b horizontal autocovariances, the autocovariance profile derived from a Lönnberg–Hollingsworth method (HL for short), and a simple Gaussian fit to the HL estimate. The HL curves are obtained by sampling the statistics of the departures $y - H(x^b)$ over horizontal bins of size 100 km and using temperature data from TEMP messages (see also Hollingsworth and Lönnberg, 1986; Lönnberg and Hollingsworth, 1986). Figure 10 is for 500 hPa, and Fig. 11 for 850 hPa. At both levels, it appears clearly that the HL autocovariance profiles are in reasonable accordance with the standard NMC profiles, while the lagged NMC profiles are shorter. This result is consistent with the

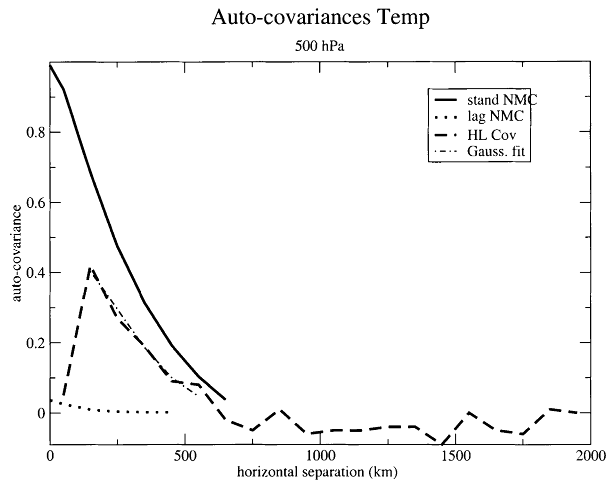


Fig 10. Horizontal error autocovariances of temperature at 500 hPa derived from different sets of data: standard NMC statistics over 90 d (solid); lagged (constant lateral boundary) NMC statistics (short-dashed); HL for TEMP data (long-dashed); HL curve fitted to a Gaussian curve (thin dash-dotted).

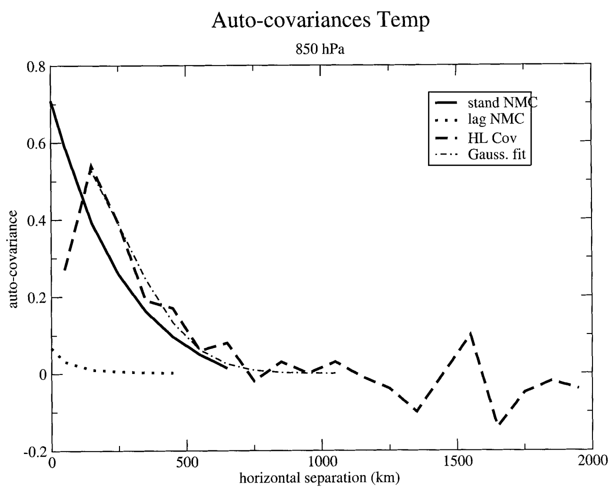


Fig 11. Horizontal error autocovariances of temperature at 850 hPa derived from different sets of data: standard NMC statistics over 90 d (solid); lagged (constant lateral boundary) NMC statistics (short-dashed); HL for TEMP data (long-dashed); HL curve fitted to a Gaussian curve (thin dash-dotted).

fact that when sampling in the HL method, all horizontal scales between TEMP station locations are considered. Thus, the HL statistics do not concentrate on a specific, mesoscale band of scales, as for the lagged NMC autocovariances, which were designed exactly for this purpose of scale selection. Therefore, in a strict application of the a posteriori diagnostics, the departures of S_b from 1 in the Aladin 3D-VAR can also partially be due to the different correlations.

More generally, it can sometimes be very difficult to prove that an observed departure of an a posteriori diagnostic is due to a misspecification of error variances only. The trace-based diagnostics used in this paper should not leave the misleading impression that only variances are at stake. Wrong error correlations can spoil the results, and in that case one should develop more refined diagnostics that actually can separate the information on correlations and length-scales from the information on variances.

Alternatively, an extension of a posteriori diagnostics to the estimation of correlation length-scales represents a challenging topic for future research. New a posteriori properties need to be defined which would point to these geographical parameters.

6. Conclusion

The a posteriori validation, and the trial to derive useful information from it with respect to the behaviour of the assimilation systems, certainly represents an elegant theoretical framework. The very specific methodology implied by a posteriori validation, based both on (sometimes strong) a priori assumptions and on the derivation of a posteriori properties of the assimilation system, makes it quite a powerful tool for general investigations and pedagogical tutorials.

However, the move from the theoretical properties to the actual evaluation in a real data assimilation system demands some preliminary verification, as we have shown in this paper. Typically, the conditions for which the a posteriori diagnostics are obtained generally are not met in reality. A striking example is the impossibility to repeat the very same analysis a number of times ‘in laboratory conditions’. Thus, ensemble statistics need to be replaced by sampling, and therefore basic properties of stationarity and ergodicity must be verified (Section 2.3). We have shown that these properties indeed hold for the diagnostics of interest in this paper, namely scalar error variance parameters for the background and the observation cost functions (so-called S_b and S_o). In addition, the practical evaluation of the diagnostics requires the use of Monte Carlo estimates of specific trace operators, following the suggestions of Girard (1987) and Desroziers and Ivanov (2001); see Section 2.2. An interesting aspect of the (S_b, S_o) type of diagnostics is that they can be stratified per pieces of the full information vector, as shown in Section 2.1. The more trivial splitting for the time being consists of stratifying S_o into observation data type, physical field, possibly geographical location. We have then obtained very stable and ‘realistic’ results, using observation aggregates as explained in Section 2.4. Mostly, when assuming no other shortcoming in the assimilation systems at stake, the results indicate that the observation error variances are generally slightly overestimated, because typical values for the S_o parameters are about 0.7–0.9 for the global 3D-VAR Arpège, and about 0.6–0.75 for the limited-area 3D-VAR Aladin system (Section 2.5). The similarity of the two S_o parameters is an indication that representativeness error is not a dominant feature, compared to the variability with respect to observation type or vertical levels. However, we cannot rule out the possibility that the S_o diagnostic can point to representativeness error if it were fed with a good quality, dense observation network. An apparently significant difference is obtained for temperature of TEMP messages, with S_o values larger than 1 both in Arpège and Aladin. Also, a difference for aircraft message zonal and meridional wind components was found in both global and limited-area systems. This latter point is quite difficult to explain, and a deeper investigation would be needed. We would recommend a more systematic evaluation of these diagnostics, in the frame of a daily or weekly monitoring over specific time intervals, in any operational or pre-operational assimilation system. Cross-comparisons can give useful indications on typical behaviours, as started in this study with both a global and a limited-area system.

In parallel to our study, more work is presently devoted to the evaluation of (S_b, S_o) type of diagnostics in the presently operational Arpège 4D-VAR, with the aim of obtaining a stratification into both background and observation bits of information, and the reuse of these parameters for a retuning of the error variances as proposed by Desroziers and Ivanov (2001). Chapnik et al. (2004) already have presented this technique in an academic framework. In our work, however, we slightly depart from their suggestion, and we only perform a 'first-order' retuning of the total error variances using the (S_b, S_o) parameters in the Aladin 3D-VAR (Section 3.1). By 'first order', we understand that we do not apply an iterative evaluation of (S_b, S_o), as suggested in the previously mentioned papers. We simply consider the first step (S_b, S_o). In theory, our estimates are therefore suboptimal, because only the fully converged iterative method leads to a maximum-likelihood estimate. Our procedure is however close to the tuning proposed in Ménard et al. (2000) or Ménard and Chang (2000) in the frame of a Kalman filter. We show that the correction is consistent with our a priori knowledge of the shortcomings of the system (especially with respect to the background error variance) and that it does not deteriorate the general quality of the assimilation. Because the correction implies a closer fit to the observations, the scores with respect to a validating observational data set are even better after retuning (Section 3.2). In general, however, we cannot state that departures observed in the (S_b, S_o) parameters systematically should lead to error variance retunings. Obviously, there are other shortcomings in the a priori modelled error statistics both for P_b and R matrices, such as misspecified or omitted error correlations (see Section 5). Furthermore, one often wants its operational system to fit more drastically to either the background (if the model is considered as reliable, or simply to preserve model balances in the analyses), or to closer fit to certain types of observations. In practice, the overall experience provided by daily monitoring, observational quality control and tools independent from a posteriori diagnostics will remain dominant features. In that sense, a posteriori diagnostics such as (S_b, S_o) should be understood as indicators, rather than systematic tuning parameters.

The scalar error parameters considered in our tuning obviously only represent a crude 'global' adaptation of the error variances. As we have shown in the Arpège 3D-VAR, the S_o term certainly changes from one observation type to another, with sometimes a significant vertical structure. Thus, an observation-dependent or vertically varying tuning would be meaningful. The main justification in the Aladin 3D-VAR case, for a couple of scalar parameters, is the early experimental recognition that the overall estimation of background error variances by the NMC method was not reliable. In that respect, the S_b values obtained in our study are in full accordance with our experimental knowledge of the system.

As discussed in Section 4, the Monte Carlo estimates presented in our study can be coupled with the output of an ensemble-based assimilation system. Thus, an interactive a posteriori diagnosis during the ensemble computations is feasible. The randomization process could then be used both for the a priori estimation of background and analysis error covariances, and indirectly in the on-line estimation of the trace operators and their associated a posteriori diagnostics. Finally, the ensemble (S_b, S_o) could be reinjected to inflate or deflate error variances, leading to an on-line tuning in the frame of an ensemble Kalman filter for instance.

7. Acknowledgments

This topic was initiated after discussions with Jean-Noël Thépaut and Olivier Talagrand, who encouraged and helped in the early stages of the work. The authors are also very grateful to Bernard Chapnik and Gérald Desroziers for fruitful discussions, and Jean-François Geleyn for an internal review. The work was partially supported by the CNRS/INSU Programme National Atmosphère et Océan à Multi-échelles (PATOM). WS benefited from stays on leave of absence from the DMN/Maroc-Météo. Her stays were supported by Météo-France grants as well as funding from the French Ministry of Foreign Affairs.

References

Belo Pereira, M., Berre, L. and Desroziers, G. 2002. Evaluation and study of the forecast error covariances of Arpège/Aladin. In: *Ateliers de Modélisation de l'atmosphére*, Météo-France/CNRM, pp. 17–20 (in French).

Bennett, A. F., Leslie, L. M., Hagelberg, C. R. and Powers, P. E. 1993. Tropical cyclone prediction using a barotropic model initialized by a generalized inverse method. *Mon. Wea. Rev.* **121**, 1714–1729.

Chapnik, B., Desroziers, G., Rabier, F. and Talagrand, O. 2004. Properties and first application of an error statistics tuning method in a variational assimilation. *Q. J. R. Meteorol. Soc.* **130**, 2253–2275.

Courtier, P., Thépaut, J.-N. and Hollingsworth, A. 1994. A strategy for operational implementation of 4D-VAR using an incremental approach. *Q. J. R. Meteorol. Soc.* **120**, 1367–1387.

Courtier, P., Andersson, E., Heckley, W., Pailleux, J., Vasiljevic, D., et al. 1998. The ECMWF implementation of three-dimensional variational assimilation (3D-VAR). Part I: formulation. *Q. J. R. Meteorol. Soc.* **124**, 1783–1807.

Dee, D. P. 1995. On-line estimation of error covariance parameters for atmospheric data assimilation. *Mon. Wea. Rev.* **123**, 1128–1145.

Desroziers, G. and Ivanov, S. 2001. Diagnosis and adaptive tuning of information-error parameters in a variational assimilation. *Q. J. R. Meteorol. Soc.* **127**, 1433–1452.

Evensen, G. 1994. Sequential data assimilation with a nonlinear quasi-geostrophic model using Monte Carlo methods to forecast error statistics. *J. Geophys. Res.* **99**(C5), 10 143–10 162.

Gandin, L. S. 1963. *Objective Analysis of Meteorological Fields.* Israel program for scientific translations, translated from Russian 1965 edition.

Girard, D. 1987. A fast Monte Carlo cross-validation procedure for large least square problems with noisy data. Technical report RR 687-M, IMAG, IMAG, Université de Grenoble, France. 22 pp.

Hollingsworth, A. and Lönnberg, P. 1986. The statistical structure of short-range forecast errors as determined from radiosonde data. Part I: the wind field. *Tellus* **38A**, 111–136.

Horányi, A., Ihász, I. and Radnóti, G. 1996. Arpège–Aladin: a numerical weather prediction model for Central Europe with the participation of the Hungarian Meteorological Service. *Idöjárás* **100**, 277–301.

Houtekamer, P. L. and Mitchell, H. L. 1998. Data assimilation using an ensemble Kalman filter technique. *Mon. Wea. Rev.* **126**, 796–811.

Houtekamer, P. L., Lefaivre, L., Derome, J., Ritchie, H. and Mitchell, H. L. 1996. A system simulation approach to ensemble prediction. *Mon. Wea. Rev.* **124**, 1225–1242.

Kalman, R. E. and Bucy, R. S. 1961. New results in linear filtering and prediction theory. *Trans. ASME,* **83D**, 95–108.

Lönnberg, P. and Hollingsworth, A., 1986. The statistical structure of short-range forecast errors as determined from radiosonde data. Part II: the covariance of height and wind errors. *Tellus* **38A**, 137–161.

Ménard, R. and Chang, L.-P. 2000. Assimilation of stratospheric chemical tracer observations using a Kalman filter. Part II: χ^2-validated results and analysis of variance and correlation dynamics. *Mon. Wea. Rev.* **128**, 2672–2686.

Ménard, R., Cohn, S. E., Chang, L.-P. and Lyster, P. M. 2000. Assimilation of stratospheric chemical tracer observations using a Kalman filter. Part I: formulation. *Mon. Wea. Rev.* **128**, 2654–2671.

Parrish, D. and Derber, J. 1992. The National Meteorological Center's spectral statistical interpolation analysis system. *Mon. Wea. Rev.* **120**, 1747–1763.

Radnóti, G., Ajjaji, R., Bubnová, R., Caian, M., Cordoneanu, E., et al. 1995. The spectral limited-area model Arpège–Aladin. In: *PWPR report series n. 7 - WMO TD n. 699*, pp. 111–118.

Sadiki, W., Fischer, C. and Geleyn, J.-F. 2000. Mesoscale background error covariances: recent results obtained with the limited-area model ALADIN over Morocco. *Mon. Wea. Rev.* **128**, 3927–3935.

Široka, M., Fischer, C., Cassé, V., Brožkova, R. and Geleyn, J.-F. 2003. The definition of mesoscale selective forecast error covariances for a limited-area variational analysis. *Meteorol. Atmos. Phys.* **82**, 227–244.

Talagrand, O. 1998. A posteriori evaluation and verification of analysis and assimilation algorithms. In: *Workshop on Diagnosis of Data Assimilation Systems*, 2–4 November 1998, ECMWF, pp. 17–28 (available from the ECMWF).

Wahba, G., Johnson, D. R., Gao, F. and Gong, J. 1995. Adaptive tuning of numerical weather prediction models: randomized GCV in three- and four-dimensional data assimilation. *Mon. Wea. Rev.* **123**, 3358–3369.