Impact of observations and tuning of observational error statistics

Bernard CHAPNIK

CNRM/GMAP

1. Introduction

This short paper aims to sum up results obtained during a PhD supervised by Olivier Talagrand (CNRS/LMD) and closely advised by Gérald Desroziers and Florence Rabier (CNRM/GMAP) in Toulouse. Most of of the results were obtained in the French global ARPEGE 4D-Var system.

The ever growing amount of available observations (among others, satellite data) reinforces the necessity of efficient tools able to evaluate the impact of each of them on the analysis ; moreover, proper statistics must be specified in order to retrieve as much information as possible from those observations.

Most data assimilation scheme rely on linear estimation theory : the analysis (further denoted x_a) is fundamentally a linear combination of the background (x_b) and of observations (*y*). From this basis, it will be tried to answer to the following questions :

How to evaluate the impact of observations, or of a certain subset of the observations?

How to use this impact in order to tune a data assimilation system?

2. Theory

A short theoretical part will be useful to introduce the notations and concepts required in order to answer those questions.

Let us first define an information vector *z*, $z^{\mathrm{T}} = \begin{pmatrix} x_b^{\mathrm{T}} & y^{\mathrm{T}} \end{pmatrix}$, the vertical concatenation of the background and the observation vectors. This vector is linked to the truth (*x*) by means of a linear operator Γ, $\Gamma^{\mathrm{T}} = \begin{pmatrix} I_n^{\mathrm{T}} & H^{\mathrm{T}} \end{pmatrix}$: $z = \Gamma x + \varepsilon$, where H is the observation operator and ε is the information error, concatenating background and observation errors $\varepsilon^{\mathrm{T}} = \begin{pmatrix} \varepsilon_b^{\mathrm{T}} & \varepsilon_o^{\mathrm{T}} \end{pmatrix}$ with covariance matrix $S = E(\varepsilon \varepsilon^{\mathrm{T}})$, where E is the expectation operator. In case observation and background errors are not correlated, *S* is equal to $\begin{pmatrix} B & 0 \\ 0 & R \end{pmatrix}$, where *B* and *R* are the background and observations error covariance matrices.

The analysis x_a is equal to : $x_a = x_b + K\,(y - H\, x_b)$, where *K*, the "gain matrix" can be written $K = P_a H^{\mathrm{T}} R^{-1}$, with $P_a = (B^{-1} + H^{\mathrm{T}} R^{-1} H)^{-1}$ which, in case the *B* and *R* matrices used in the system are the optimal matrices, is also the analysis error covariance matrix.

In variational data assimilation this analysis is obtained as the state vector minimizing an objective function *J* (often called "cost function") :

$$J(x) = (y - H\, x)^{\mathrm{T}} R^{-1} (y - H\, x) + (x - x_b)^{\mathrm{T}} B^{-1} (x - x_b) ,$$

which, using the information vector can be rewritten : $J(x) = (z - \Gamma x)^{\mathrm{T}} S^{-1} (z - \Gamma x)$.

If one supposes that this cost function can be split into several parts : $J = \sum J_i$, with each J_i written as : $J_i(x) = (z_i - \Gamma_i x)^{\mathrm{T}} S_i^{-1} (z_i - \Gamma_i x)$, z_i is the i^{th} subpart with dimension n_i extracted from the information vector *z*, associated with the Γ_i observation operator and the S_i covariance matrix of its associated errors.

Then, if the specified covariances of the assimilation system really are the optimal matrices, an important result provided by Talagrand (1999) applies. The expectation of the i^{th} subpart of the objective function at the minimum is :

$$E(J_i(x_a)) = n_i - Trace(\Gamma_i P_a \Gamma_i^T S_i^{-1})$$

Moreover, $Trace(\Gamma_i P_a \Gamma_i^T S_i^{-1})$ is a measurement of the contribution of $\mathbf{z}_i$ to the overall precision of the assimilation system.

A more explicit signification of these values can be obtained when focusing on the background/observations splitting of the objective function :

- when $z_i = x_b$, $Trace(\Gamma_i P_a \Gamma_i^T S_i^{-1}) = n - Trace(KH)$, then $E(J_b(xa)) = Trace(KH)$
- when $z_i = y$, , $Trace(\Gamma_i P_a \Gamma_i^T S_i^{-1}) = Trace(HK)$, , then $E(J_o(xa)) = Trace(I_p - HK)$. (I_p is the identity with order p)

It must be remarked that the use of the Trace of HK as quantification of the impact of observations had already be introduced by Wahba (1995) in a meteorological context. This quantity is called DFS, for Degrees of Freedom for Signal, see also Cardinali et al. (2004) for another example of implementation of this diagnostic in a global data assimilation system.

Several points must be stated at this stage about DFS. DFS quantifies how the system uses the observations to pull the signal from the background; in the optimal case (i.e K operationally specified = true K), this is also the relative reduction of variance. Used on its own, DFS says what the system does, without any other criterion it cannot say what it should do in order to improve the analysis.

A first clear problem appears : How do we compute DFS when K generally does not even explicitly exist in a variational scheme ?

3. Practical Computation of *Trace (HK)*

Two methods have been implemented in order to compute estimates of this Trace.

3.1 Girard's method

The first method was proposed Girard (1987), it was introduced in the field of meteorological data assimilation by Wahba (1995), and by Desroziers and Ivanov (2001).

The method is based on the following mathematical identity : considering a random vector ε with 0 mean and the identity covariance matrix, and an operator *A*, the expectation of the quadratic form $\epsilon^T A \epsilon$ is :

$$E(\epsilon^T A \epsilon) = Trace(A)$$

This mathematical property is used as following :

- make a first "normal" analysis x_a using the usual background and observations,
- make a perturbed analysis x_a^* using the same background and perturbed observations, $y^* = y + R^{0.5}\zeta$

It can easily be verified that the following scalar product approximates the wished quantity :

$$(y^* - y) R^{-1} (H x_a^* - H x_a) \tilde{} \; Trace(HK)$$

3.2 The Simulated Optimal Innovations (SOI) method

This method is introduced in Chapnik et al. (2005). It is based on the properties of subparts of the optimal objective function at the minimum described in section 2.

The algorithm consists in generating a situation, the errors of which are consistent with the specified covariance matrices.

A state vector x (for example a background vector) is considered as the "truth". Adding some noise, consistent with the specified statistics, a simulated background $x_b^* = x + B^{0.5}\zeta_b$ and observations $y^* = y + R^{0.5}\zeta_o$ are generated. The variational analysis of this simulated situation naturally leads to the computation of $J_b(x_a^{(*)})$ and of $J_o(x_a^*)$ and of possible subparts of them. One then has :

$$J_b(x_a^*) \simeq Trace(KH);$$
$$J_o(x_a^*) \simeq Trace(I_p - HK) \ ;.$$

The following equality : $(H(x_b^* - x_a^*))^T R^{-1}(y_o^* - H x_a^*) \simeq Trace(HK)$ can also be applied in order to compute subparts of $Trace(HK)$ (this can only be applied if this subpart corresponds to a diagonal block of R).

3.3 Comparison of the two methods

Figure 1 compares the DFS computed for several upper atmosphere observation types on 4/02/2004 at 00 UTC within ARPEGE 4D-Var system. One may see that the results of the two methods compare quite well. The small discrepancies observed, at least those larger than what can be expected from randomized estimation methods (see the AMSU observations DFS, for example) may be explained by the non-linearities of the multi-incremental 4D-Var scheme used here. It can be shown that even in this case, Girard's method still evaluates a good estimate of the sensitivity of the analysis to the observations while the SOI method may be less accurate.

4. The tuning of variances

A way to use DFS (or a very similar quantity) to tune the specified statistics was provided by Desroziers and Ivanov (2001).

An hypothesis is made that the true optimal matrices can be obtained from the specified matrices, just by multiplying them by multiplicative coefficients, the tuning coefficients; for example, it may be supposed that $\mathbf{B}_t$ and $\mathbf{R}_t$ (the optimal background and observation error covariances) may be deduced from the specified **B** and **R** as:

$$B_t = s_b B$$
$$R_t = s_o R$$

Supposing the system is variational, if J_o and J_b are the subparts of the objective function related to $\mathbf{x}_b$ and **y** respectively then $J_{opt} = J_o/s_o + J_b/s_b$ is the optimal objective function. The criterion of the expectation of subparts of the objective function at the minimum must then apply. Let $\mathbf{x}_a$ be the minimizer of J_{opt} ; replacing the expectation operator by one realization (which in case there are enough observations is justified) yields the following criterion to determine s_o and s_b:

$$J_b(x_a(s_o, s_b))/s_b = Trace(K(s_o, s_b)H);$$
$$J_o(x_a(s_o, s_b))¿/s_o = Trace(I_p - HK(s_o, s_b))..$$

The notations used here are to emphasize the fact that x_a and K are functions of s_o and s_b. Those equalities are easily transformed into:

$$s_o = 2J_o(x_a(s_o, s_b))/Tr(I_p - HK(s_o, s_b))$$
$$s_b = 2J_b(x_a(s_o, s_b))/Tr(K(s_o, s_b)H)$$

This set of equation is a fixed-point relation $((s_o, s_b) = f(s_o, s_b))$. A fixed-point algorithm is then applied to compute (s_o, s_b) . The Trace term can be computed with Girard's method, but with the SOI method the previous relations become

$$s_o = J_o(x_a(s_o, s_b)) / J_o(x_a^*(s_o, s_b));$$
$$s_b = J_b(x_a(s_o, s_b)) / J_b(x_a^*(s_o, s_b)).$$

xa^* is the analysis made from the simulated situation defined in the presentation of the SOI method. These expressions outline that the method compares "true" and "simulated" statistics. An advantage of the SOI method is that the numerator and the denominator of these expressions can be obtained in the same way.

A nice property of the algorithm is that the first iteration of the fixed point used here converges very quickly. The first iteration generally provides a good estimate of the result and convergence is generally reached after two or three iterations, except for cases that are going to be defined in the next section.

5. Properties of the method

5.1 Equivalence to Maximum Likelihood Tuning

It may be shown (Chapnik et al 2004) that the tuning performs a Maximum Likelihood tuning of variances (see Dee and da Silva 1998 for meteorological use of Maximum Likelihood in data assimilation), meaning that the tuning coefficients are the most probable coefficients, considering an a priori model of covariance (the specified B and R matrix) and the data (in our case the innovation, obs-guess difference). This has important consequences for the tuning :

Since the method performs a statistic over he innovation, a large innovation vector, therefore a large observation vector, is needed.

An a priori hypothesis about the structure of correlations, allowing to split the innovation into observation and background errors is necessary (like in Hollingsworth and Lönnberg, 1986). This hypothesis is in HBH^T and R (the *a priori* specified matrices) which must be different (e.g. no spatial correlation in R , spatial correlation in B) to allow a useful tuning.

If this hypothesis is not (even roughly) respected by the true ϵ_o and ϵ_b , a poor tuning may be expected. In particular, performing the method on an observation type with spatially correlated errors represented with a diagonal R yields a very weak, possibly null s_o , which is the opposite of what should be done in this case (inflate the optimal σ_o). Note that a non optimal correlation length in B does not have such bad consequences.

5.2 A first try with real data

A first try is made to check in a first time if the tuning coefficients have *a priori* desirable properties they should have. Desroziers and Ivanov (2001) had already shown the ability of the method to retrieve the tuning coefficients for a simulated case in a comprehensive data assimilation system. The consistency between the tuning coefficients and the known quality of the observations is tested here.

Figure 2 shows the tuning coefficients computed for satellite borne instruments channels, in 1997 and in 2001. One may clearly see that small coefficients remain small and that large coefficients remain large over four years. Moreover, the variability between the two dates is of the same order as the one encountered when comparing the tuning coefficients computed at different dates of the same month (not shown). Such a behaviour is a positive point if we suppose that there were no major evolution of the quality of the observation between the two dates; yet, evidence that the result is not an artifact is still needed, evidence that the same result will not be obtained independently of the σ_o of the observation errors. A known dysfunction of NOAA 15 instruments was the occasion to document this point. Figure 3 shows the tuning coefficients obtained for three channels on dates when there were no problems (dashed black and white bars), on the day before the problem begins (green bars), and during the incident (the other bars). It is clearly seen for two of the channels that the tuning coefficients are multiplied by two (and even more on the last date) during the incident. The tuning coefficients are clearly related to the quality of the observations.

6. Impact of the tuning of the variance

The final point of this study is the assessment of the tuning of the specified observational error variances on the analysis and on the forecasts.

In a first time the tuning was performed for the assimilated observation types. For observations known to have very correlated observation errors (like SATOB observations), the tuned values were taken similar to the specified values (the true tuned values dropping to 0 along the tuning as noticed in section 5). A single coefficient was applied to tune B, and was found equal to 1, meaning that, as a whole, B was approximately correct.

An experiment was carried out, performing an analysis cycle for 20 days with the "tuned" analysis system. Figures 4a-c compare the rms differences between geopotential observations and forecast, for "tuned" forecasts and operational forecasts. The green lines denote improvements of the rms, the red ones show a deterioration. It can be seen that for this parameter, the impact is positive for all forecast ranges. For other parameters, the impact, though positive, may be less spectacular.

7. Conclusion

Techniques to evaluate the quantification of the impact of the observations, known as DFS, have been implemented for the French ARPEGE data assimilation system. These techniques can also be used for Desroziers and Ivanov's tuning of the variances. This tuning has been shown to have some positive impact on the analysis and on the forecasts.

The first future direction which might be taken is the tuning of the B matrix. As stated before, only one global coefficient was applied to this matrix, which is certainly not enough. Several strategies for a finer tuning may be considered.

Another difficulty to be considered in the future is the tuning of observations with correlated errors (like SATOB). This case is more difficult since no known objective criterion allowing to tune it can apply.

8. References

C. Cardinali, S. Pezzuli and E. Andersson Influence Matrix diagnostic of a data assimilation system. *Proceedings of the Seminar on Recent developments in data assimilation for atmosphere and ocean*, ECMWF, Reading, 8-12 September 2003.

B. Chapnik, G. Desroziers, F.Rabier and O.Talagrand. 2004, Properties and first applications of an error statistic tuning method in variationnal assimilation. *Quart. J. Roy. Meteor. Soc.,* 130, no. 601, pp. 2253-2275

B. Chapnik, G. Desroziers, F. Rabier, and O. Talagrand. 2005, Diagnosis and tuning of observational error statistics in a quasi operational data assimilation setting. Accepted with minor revision by *Quart. J. Roy. Meteor. Soc.*

G. Desroziers and S. Ivanov. 2001, Diagnosis and adaptive tuning of information error parameters in a variational assimilation. *Quart. J. Roy. Meteor. Soc.,* 127, 1433-1452.

D. Dee and A. da Silva. 1998, Maximum likelihood estimation of forecast and observation error covariance parameters. part I: Methodology. *Mon. Wea. Rev.*, 124:1822--1834.

D. Girard. 1987, A fast Monte Carlo cross-validation procedure for large least squares problems with noisy data. *Technical Report* 687-M, IMAG, Grenoble, France.

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

O. Talagrand. 1999 A posteriori verification of analysis and assimilation algorithms. In *Proceedings of the ECMWF Workshop on Diagnosis of Data Assimilation Systems*, pages 17--28, Reading, November 1999.

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

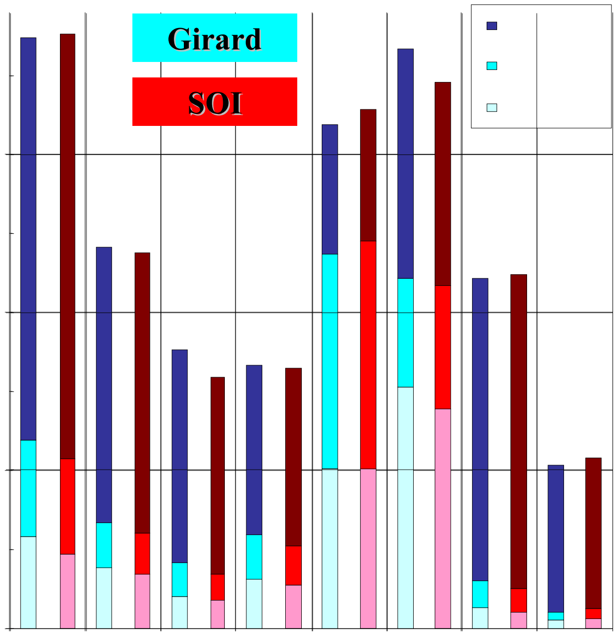

Figure 1 : DFS computed for different observation types. The bars in blue were computed with Girard's method, bars in red with the SOI method. Each bar is divided into three parts: the upper part is the contribution of observations from the northern hemisphere; the middle part is the contribution from subtropical observations ; the lower part, the contribution from the southern hemisphere.

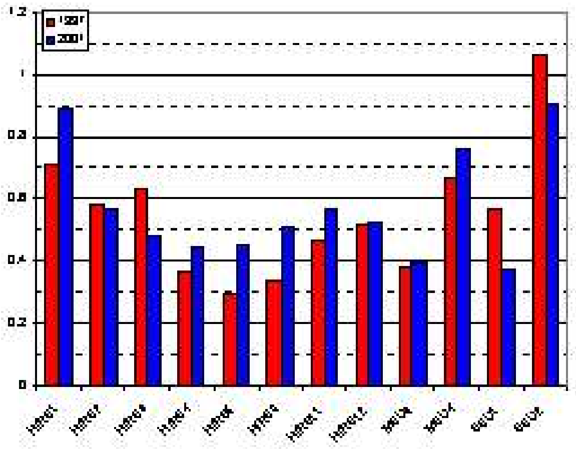

Figure 2 : Comparison between tuning coefficients computed on a 1997 date (in red) and a date in 2001 (in blue), for different satellite channels (along the x axis).

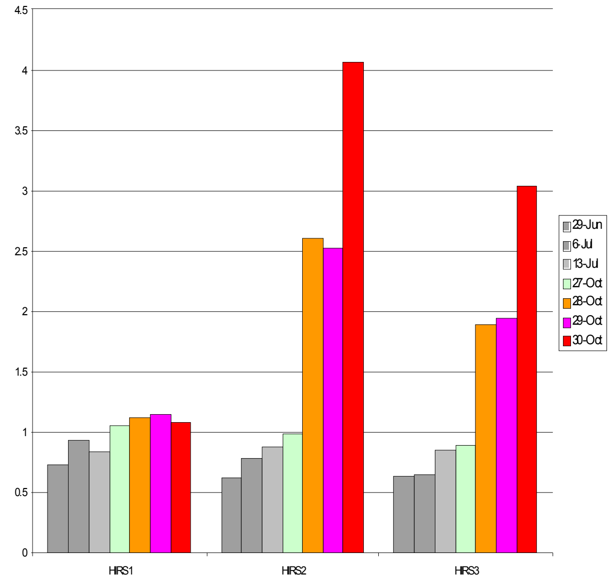

Figure 3 : Detection of an incident. Tuning coefficients computed for dates before the incident (dashed grey bars),the day before the incident (green bars), and during the incident (orange, purple and red bars).

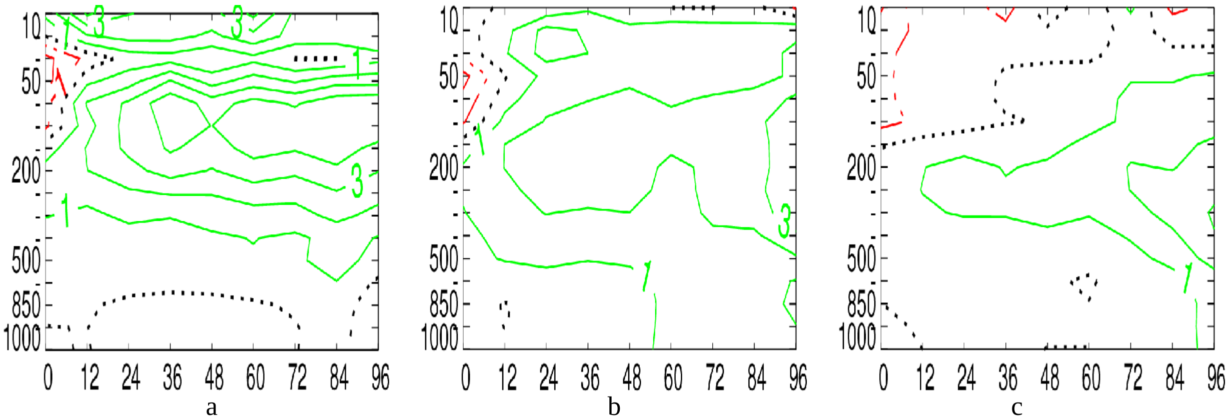

Figure 4 : Difference between the rms (geopotential TEMP observations minus forecast) for operational forecasts and "tuned" forecasts. The unit is the meter. Green lines show an improvement, red lines a deterioration. The x axis is the range of the forecast, the y axis is the pressure level. Panel 4a is for inter tropical areas, panel 4b for the southern hemisphere and panel 4c for the northern hemisphere.

CONTENTS