Data Assimilation

Observation influence diagnostic of a data assimilation system

C. Cardinali

Research Department

June 2013

Series: ECMWF Lecture Notes

A full list of ECMWF Publications can be found on our web site under:
http://www.ecmwf.int/publications/

Contact: library@ecmwf.int

European Centre for Medium-Range Weather Forecasts
Shinfield Park, Reading, RG2 9AX, England

The information within this publication is given in good faith and considered to be true, but ECMWF accepts no liability for error, omission and for loss or damage arising from its use.

Summary

The influence matrix is used in ordinary least-squares applications for monitoring statistical multiple-regression analyses. Concepts related to the influence matrix provide diagnostics on the influence of individual data on the analysis, the analysis change that would occur by leaving one observation out, and the effective information content (degrees of freedom for signal) in any sub-set of the analysed data. In this paper, the corresponding concepts are derived in the context of linear statistical data assimilation in Numerical Weather Prediction. An approximate method to compute the diagonal elements of the influence matrix (the self-sensitivities) has been developed for a large-dimension variational data assimilation system (the 4D-Var system of the European Centre for Medium-Range Weather Forecasts). Results show that, in the ECMWF operational system, 18% of the global influence is due to the assimilated observations, and the complementary 82% is the influence of the prior (background) information, a short-range forecast containing information from earlier assimilated observations. About 20% of the observational information is currently provided by surface-based observing systems, and 80% by satellite systems.

A toy-model is developed to illustrate how the observation influence depends on the data assimilation covariance matrices. In particular, the role of high-correlated observation error and high-correlated background error with respect to uncorrelated ones is presented. Low-influence data points usually occur in data-rich areas, while high-influence data points are in data-sparse areas or in dynamically active regions. Background error correlations also play an important role: high correlation diminishes the observation influence and amplifies the importance of the surrounding real and pseudo observations (prior information in observation space). To increase the observation influence in the presence of high correlated background error, it is necessary to also take the observation error correlation into consideration. However, if the observation error variance is too large with respect to the background error variance the observation influence will not increase. Incorrect specifications of the background and observation error covariance matrices can be identified by the use of the influence matrix.

KEYWORDS: Observations Influence Data Assimilation Regression Methods

1 Introduction

Over the years, data assimilation schemes have evolved into very complicated systems, such as the four-dimensional variational system (4D-Var) (Rabier *et al*. 2000) at the European Centre for Medium-Range Weather Forecasts (ECMWF). The scheme handles a large variety of both space and surface-based meteorological observations. It combines the observations with prior (or background) information of the atmospheric state and uses a comprehensive (linearized) forecast model to ensure that the observations are given a dynamically realistic, as well as statistically likely response in the analysis.

Effective monitoring of such a complex system, with the order of 10^9 degrees of freedom and more than 10^7 observations per 12-hour assimilation cycle, is a necessity. The monitoring cannot be restricted to just a few indicators, but a complex set of measures is needed to indicate how different variables and regions influence the data assimilation (DA) scheme. Measures of the observational influence are useful for understanding the DA scheme itself: How large is the influence of the latest data on the analysis and how much influence is due to the background? How much would the analysis change if one single influential observation were removed? How much information is extracted from the available data? It is the aim of this work to provide such analytical tools.

We turn to the diagnostic methods that have been developed for monitoring statistical multiple regression analyses. In fact, 4D-Var is a special case of the Generalized Least Square (GLS) problem (Talagrand, 1997) for weighted regression, thoroughly investigated in the statistical literature.

The structure of many regression data sets makes effective diagnosis and fitting a delicate matter. In robust (resistant) regression, one specific issue is to provide protection against distortion by anomalous data. In fact, a single unusual observation can heavily distort the results of ordinary (non-robust) LS regression (Hoaglin *et al*. 1982). Unusual or influential data points are not necessarily bad data points: they may contain some of the most useful sample information. For practical data analysis, it helps to judge such effects quantitatively. A convenient diagnostic measures the effect of a (small) change in the observation y_i on the corresponding predicted (estimated) value $\hat{y}_i$. In LS regression this involves a straightforward calculation: any change in y_i has a proportional impact on $\hat{y}_i$. The desired information is available in the diagonal of the *hat matrix* (Velleman and Welsh, 1981), which gives the estimated values $\hat{y}_i$ as a linear combination of the observed values y_i . The term *hat matrix* was introduced by J.W. Tukey (Tukey, 1972) because the matrix maps the observation vector **y** into $\hat{\mathbf{y}}$, but it is also referred to as the *influence matrix* since its elements indicate the data influence on the regression fit of the data. The matrix elements have also been referred to as the *leverage* of the data points: in case of high *leverage* a unit y-value will highly disturb the fit (Hoaglin and Welsh, 1978). Concepts related to the influence matrix also provide diagnostics on the change that would occur by leaving one data point out, and the effective information content (degrees of freedom for signal) in the data.

These influence matrix diagnostics are explained in Section 2 for ordinary least-squares regression. In Section 3 the corresponding concepts for linear statistical DA schemes is derived. It will be shown that observational influence and background influence complement each other. Thus, for any observation y_i either very large or very small influence could be the sign of inadequacy in the assimilation scheme, and may require further investigation. A practical approximate method that enables calculation of the diagonal elements of the influence matrix for large-dimension variational schemes (such as ECMWFs operational 4D-Var system) is described in Cardinali *et al* 2004 and therefore not shown here. In Section 4 results and selected examples related to data influence diagnostics are presented, including an investigation into the effective information content in several of the main types of observational data. Conclusions are drawn in Section 5.

2 Classical Statistical Definitions of Influence Matrix and Self-Sensitivity

The ordinary linear regression model can be written:

$$\mathbf{y} = \mathbf{X}\beta + \varepsilon \tag{1}$$

where **y** is an $m \times 1$ vector for the response variable (predictand); **X** is an $m \times q$ matrix of q predictors; β is a $q \times 1$ vector of parameters to be estimated (the regression coefficients) and ε is an $m \times 1$ vector of errors (or fluctuations) with expectation E(ε)=0 and covariance var(ε)=$\sigma^2\mathbf{I}_m$ (that is, uncorrelated observation errors). In fitting the model (1) by LS, the number of observations m has to be greater than the number of parameters q in order to have a well-posed problem, and **X** is assumed to have full rank q.
The LS method provides the solution of the regression equation as
$\beta = (\mathbf{X}^\mathbf{T}\mathbf{X})^{-1}\mathbf{X}^\mathbf{T}\mathbf{y}$. The fitted (or estimated) response vector is thus:

$$\hat{\mathbf{y}} = \mathbf{S}\mathbf{y} \tag{2}$$

where

$$\mathbf{S} = \mathbf{X}(\mathbf{X^T X})^{-1}\mathbf{X^T} \tag{3}$$

is the $m \times m$ *influence matrix* (or hat matrix). It is easily seen that

$$\mathbf{S} = \frac{\delta \hat{\mathbf{y}}}{\delta \mathbf{y}} \tag{4}$$

and that

$$\begin{aligned} S_{ij} &= \frac{\delta \hat{y}_i}{\delta y_j} \\ S_{ii} &= \frac{\delta \hat{y}_i}{\delta y_i} \end{aligned} \tag{5}$$

for the off-diagonal ($i \neq j$) and the diagonal (i=j) elements, respectively. Thus, S$_{ij}$ is the rate of change of $\hat{y}_i$ with respect to y_j variations. The diagonal element S$_{ii}$ instead, measures the rate of change of the regression estimate $\hat{y}_i$ with respect to variations in the corresponding observation y_i. For this reason the *self-sensitivity* (or self-influence, or leverage) of the i^{th} data point is the i^{th} diagonal element S$_{ii}$, while an off-diagonal element is a *cross-sensitivity* diagnostic between two data points. Hoaglin and Welsh (1978) discuss some properties of the influence matrix. The diagonal elements satisfy:

$$\begin{aligned} &0 \leq S_{ii} \leq 1 \\ &i = 1, 2, \ldots, m \end{aligned} \tag{6}$$

as **S** is a symmetric and idempotent projection matrix (**S=S**^{2}). The covariance of the error in the estimate $\hat{y}$, and the covariance of the residual **r=y-ŷ** are related to **S** by

$$\begin{aligned} var(\hat{\mathbf{y}}) &= \sigma^2 \mathbf{S} \\ var(\mathbf{r}) &= \sigma^2 (I_m - \mathbf{S}) \end{aligned} \tag{7}$$

The trace of the influence matrix is:

$$tr(\mathbf{S}) = \sum_{i=1}^{m} S_{ii} = q = rank(\mathbf{S}) \tag{8}$$

(in fact **S** has m eigenvalues equals to 1 and m-q zeros). Thus, the trace is equal to the number of parameters. The trace can be interpreted as the amount of information extracted from the observations or *degrees of freedom for signal* (Wahba *et al*. 1995). The complementary trace, tr(**I-S**)=m-tr(**S**) , on the other hand, is the degree of freedom for noise, or simply the degree of freedom (df) of the error variance, widely used for model checking (F test).

A zero self-sensitivity S$_{ii}$=0 indicates that the i^{th} observation has had no influence at all in the fit, while S$_{ii}$=1 indicates that an entire degree of freedom (effectively one parameter) has been devoted to fitting just that data point. The average self-sensitivity value is q/m and an individual element S$_{ii}$ is considered ‘large’ if its value is greater than three times the average (Velleman and Welsh, 1981). By a symmetrical argument a self-sensitivity value that is less than one-third of the average is considered ‘small’.

Furthermore, the change in the estimate that occurs when the i^{th} observation is deleted is

$$\hat{y}_i - \hat{y}_i^{(-i)} = \frac{S_{ii}}{(1 - S_{ii})} r_i \tag{9}$$

where $\hat{y}_i{}^{(-i)}$ is the LS estimate of y_i obtained by leaving-out the i^{th} observation of the vector **y** and the i^{th} row of the matrix **X**. The method is useful to assess the quality of the analysis by using the discarded observation, but impractical for large systems. The formula shows that the impact of deleting (y_i,$\mathbf{x}_i$) on $\hat{y}_i$ can be computed by knowing only the residual r_i and the diagonal element S_{ii} - the nearer the self-sensitivity S_{ii} is to one, the more impact on the estimate $\hat{y}_i$. A related result concerns the so-called cross-validation (CV) score: that is, the LS objective function obtained when each data point is in turn deleted (Whaba, 1990, theorem 4.2.1):

$$\sum_{i=1}^{m}(y_i-\hat{y}_i{}^{(-i)})^2=\sum_{i=1}^{m}\frac{(y_i-\hat{y}_i)^2}{(1-S_{ii})^2} \tag{10}$$

This theorem shows that the CV score can be computed by relying on the all-data estimate $\hat{\mathbf{y}}$ and the self-sensitivities, without actually performing m separate LS regressions on the leaving-out-one samples. Moreover, (9) shows how to compute self-sensitivities by the leaving-out-one experiment.

The definitions of influence matrix (4) and self-sensitivity (5) are rather general and can be applied also to non-LS and nonparametric statistics. In spline regression, for example, the interpretation remains essentially the same as in ordinary linear regression and most of the results, like the CV-theorem above, still apply. In this context, Craven and Wahba (1979) proposed the generalized-CV score, replacing in (10) S_{ii} by the mean $tr(\mathbf{S})/q$. For further applications of influence diagnostics beyond usual LS regression (and further references) see Ye (1998) and Shen *et al.* (2002). The notions related to the influence matrix that it has introduced here will in the following section be derived in the context of a statistical analysis scheme used for data assimilation in numerical weather prediction (NWP).

3 Observational Influence and Self-Sensitivity for a DA Scheme

(a) Linear statistical estimation in Numerical Weather Prediction

Data assimilation systems for NWP provide estimates of the atmospheric state **x** by combining meteorological observations **y** with prior (or background) information $\mathbf{x}_b$. A simple Bayesian Normal model provides the solution as the posterior expectation for **x**, given **y** and $\mathbf{x}_b$. The same solution can be achieved from a classical *frequentist* approach, based on a statistical linear analysis scheme providing the Best Linear Unbiased Estimate (Talagrand, 1997) of **x**, given **y** and $\mathbf{x}_b$. The optimal GLS solution to the analysis problem (see Lorenc, 1986) can be written

$$\mathbf{x}_a=\mathbf{K}\mathbf{y}+(\mathbf{I}_n-\mathbf{K}\mathbf{H})\mathbf{x}_b \tag{11}$$

The vector $\mathbf{x}_a$ is the ‘analysis’. The gain matrix **K** ($n\times m$) takes into account the respective accuracies of the background vector $\mathbf{x}_b$ and the observation vector **y** as defined by the $n\times n$ covariance matrix **B** and the $m\times m$ covariance matrix **R**, with

$$\mathbf{K}=(\mathbf{B}^{-1}+\mathbf{H}^{\mathbf{T}}\mathbf{R}^{-1}\mathbf{H})^{-1}\mathbf{H}^{\mathbf{T}}\mathbf{R}^{-1} \tag{12}$$

Here, **H** is a $m\times n$ matrix interpolating the background fields to the observation locations, and transforming the model variables to observed quantities (e.g. radiative transfer calculations transforming the models temperature, humidity and ozone into brightness temperatures as observed by several satellite instruments). In the 4D-Var context introduced below, **H** is defined to include also the propagation in time of the atmospheric state vector to the observation times using a forecast model. Substituting (12) into (**??**) and projecting the analysis estimate onto the observation space, the estimate becomes

$$\hat{\mathbf{y}}=\mathbf{H}\mathbf{x_a}=\mathbf{H}\mathbf{K}\mathbf{y}+(\mathbf{I_m}-\mathbf{H}\mathbf{K})\mathbf{H}\mathbf{x_b} \tag{13}$$

It can be seen that the analysis state in observation space ($\mathbf{Hx}_a$) is defined as a sum of the background (in observation space, $\mathbf{Hx}_b$) and the observations $\mathbf{y}$, weighted by the $m \times m$ square matrices **I-HK** and **HK** , respectively.
Equation (13) is the analogue of (1), except for the last term on the right hand side. In this case, for each unknown component of **Hx**, there are two data values: a real and a 'pseudo' observation. The additional term in (13) includes these pseudo-observations, representing prior knowledge provided by the observation-space background $\mathbf{Hx}_b$. From (13) and (4), the analysis sensitivity with respect to the observations is obtained

$$\mathbf{S} = \frac{\delta \hat{\mathbf{y}}}{\delta \mathbf{y}} = \mathbf{K^T H^T} \tag{14}$$

Similarly, the analysis sensitivity with respect to the background (in observation space) is given by

$$\frac{\delta \hat{\mathbf{y}}}{\delta (\mathbf{Hx_b})} = \mathbf{I} - \mathbf{K^T H^T} = \mathbf{I_m} - \mathbf{S} \tag{15}$$

Lets focus here on the expressions (14) and (15). The influence matrix for the weighted regression DA scheme is actually more complex (see Appendix 1), but it obscures the dichotomy of the sensitivities between data and model in observation space.

The (projected) background influence is complementary to the observation influence. For example, if the self-sensitivity with respect to the ith observation is S$_{ii}$, the sensitivity with respect the background projected at the same variable, location and time will be simply 1-S$_{ii}$. It also follows that the complementary trace, tr(**I-S**)=m-tr(**S**), is not the df for noise but for background, instead. That is the weight given to prior information, to be compared to the observational weight tr(**S**). These are the main differences with respect to standard LS regression. Note that the different observations can have different units, so that the units of the cross-sensitivities are the corresponding unit ratios. Self-sensitivities, however, are pure numbers (no units) as in standard regression. Finally, as long as **R** is diagonal, (6) is assured (see Section 3(b)), but for more general non-diagonal **R**-matrices it is easy to find counter-examples to that property.
Inserting (12) into (14), we obtain

$$\mathbf{S} = \mathbf{R^{-1}H(B^{-1} + H^T R^{-1} H)^{-1} H^T} \tag{16}$$

As $\mathbf{(B^{-1} + H^T R^{-1} H)^{-1}}$ is equal to the analysis error covariance matrix **A**, we can also write $\mathbf{S} = \mathbf{R^{-1} H A H^T}$.

(b) R diagonal

In this section it is shown that as long as **R** is diagonal (6) is satisfied. Equation (16) can be written as

$$\begin{aligned} \mathbf{S} &= \mathbf{R^{-1}H[B - BH^T(HBH^T + R)^{-1}HB]H^T} \\ &= \mathbf{R^{-1}HBH^T - R^{-1}HBH^T(HBH^T + R)^{-1}HBH^T} \end{aligned} \tag{17}$$

Lets introduce the matrix $\mathbf{V{=}HBH}^T$, (17) becomes

$$
\begin{aligned}
\mathbf{S} &= \mathbf{R}^{-1}\mathbf{V} - \mathbf{R}^{-1}\mathbf{V}(\mathbf{V}+\mathbf{R})^{-1}\mathbf{V} \\
&= \mathbf{R}^{-1}\mathbf{V}[\mathbf{I} - (\mathbf{V}+\mathbf{R})^{-1}\mathbf{V}] \\
&= \mathbf{R}^{-1}\mathbf{V}[(\mathbf{V}+\mathbf{R})^{-1}(\mathbf{V}+\mathbf{R}) - (\mathbf{V}+\mathbf{R})^{-1}\mathbf{V}] \\
&= \mathbf{R}^{-1}\mathbf{V}(\mathbf{V}+\mathbf{R})^{-1}\mathbf{R} \\
&= \mathbf{R}^{-1}[(\mathbf{V}+\mathbf{R})(\mathbf{V}+\mathbf{R})^{-1} - \mathbf{R}(\mathbf{V}+\mathbf{R})^{-1}]\mathbf{R} \\
&= \mathbf{R}^{-1}[\mathbf{I} - \mathbf{R}(\mathbf{V}+\mathbf{R})^{-1}]\mathbf{R} \\
&= \mathbf{I} - (\mathbf{V}+\mathbf{R})^{-1}\mathbf{R} \\
&= (\mathbf{V}+\mathbf{R})^{-1}\mathbf{V}
\end{aligned} \tag{18}
$$

Since **V** and **R** are positive definite covariance matrices, the matrix **(V+R)** is positive definite as well. In fact by definition for a non-zero vectors **z** with real entries the quantity $\mathbf{z}^{\mathbf{T}}(\mathbf{V}+\mathbf{R})\mathbf{z} = \mathbf{z}^{\mathbf{T}}\mathbf{V}\mathbf{z} + \mathbf{z}^{\mathbf{T}}\mathbf{R}\mathbf{z} > \mathbf{0}$. Lets consider the following theorem: If **D** is positive definite matrix then $\mathbf{D}^{-1}$ is positive definite and defining $\mathbf{D}^{-1}$=$\{\delta_{ij}\}$, D=$\{d_{ij}\}$ we have: $\delta_{ii} \geq 1/d_{ii}$ where the equality holds if and only if $d_{i1} = .. = d_{ii-1} = d_{ii+1} = .. = d_{in=0}$.
The diagonal elements of $\mathbf{D}^{-1}$=**(V+R)**$^{-1}$=$\{\delta_{ij}\}$ are then larger than the diagonal elements of **(V+R)**. Moreover, if **V**=$\{v_{ij}\}$ and **R**=diag(r_i) we obtain

$$\delta_{ii} \geq \frac{1}{v_{ii} + r_i} \tag{19}$$

And since the i-diagonal element of **(V+R)**$^{-1}$**R** is $(\delta_{i1}, \ldots, \delta_{in}) \begin{pmatrix} 0 \\ \vdots \\ r_i \\ \vdots \\ 0 \end{pmatrix} = \delta_{ii} r_i$

$$\delta_{ii} r_i \geq \frac{r_i}{v_{ii} + r_i} \tag{20}$$

From (18) considering that the product of two positive definite matrix is still a positive definite matrix

$$0 < S_{ii} = 1 - \delta_{ii} r_i \leq 1 - \frac{r_i}{v_{ii} + r_i} = \frac{v_{ii}}{v_{ii} + r_i} < 1 \tag{21}$$

(21) proves that the diagonal elements of the influence matrix for the weighted regression DA scheme are bound between (0,1).

(c) Toy model

Lets assume a simplified model with two observations, each coincident with a point of the background - that is **H=I**$_2$. Assume the error of the background at the two locations have correlation α, that is / **B**=$\sigma_b^2 \begin{pmatrix} 1 & \alpha \\ \alpha & 1 \end{pmatrix}$, with variance σ_b^2, and that similarly **R**=$\sigma_o^2 \begin{pmatrix} 1 & \beta \\ \beta & 1 \end{pmatrix}$ with variance σ_o^2 and correlation β. For this simple case **S** is obtained from (14)

$$S_{11} = S_{22} = \frac{\sigma_b^2\sigma_o^2(1-\alpha\beta)+\sigma_b^4(1-\alpha^2)}{\sigma_b^4(1-\alpha^2)+\sigma_o^4(1-\beta^2)+2\sigma_b^2\sigma_o^2(1-\alpha\beta)} \tag{22}$$

$$S_{12} = S_{21} = \frac{\sigma_b^2\sigma_o^2(\alpha-\beta)}{\sigma_b^4(1-\alpha^2)+\sigma_o^4(1-\beta^2)+2\sigma_b^2\sigma_o^2(1-\alpha\beta)} \tag{23}$$

For $\alpha \neq \pm 1$ and $\beta \neq \pm 1$ (**R** and **B** are full rank matrices). Lets define $r = \sigma_o^2/\sigma_b^2$, (22) and (23) reduce too

$$S_{11} = S_{22} = \frac{r(1-\alpha\beta)+1-\alpha^2}{r^2(1-\beta^2)+1-\alpha^2+2r(1-\alpha\beta)} \tag{24}$$

$$S_{12} = S_{21} = \frac{r(\alpha-\beta)}{r^2(1-\beta^2)+1-\alpha^2+2r(1-\alpha\beta)} \tag{25}$$

Figure 1 shows the diagonal elements of the influence matrix as a function of $r, S_{ii} = S_{ii}(r)$ (Eq. 24). From now on, S$_{ii}$ is also indicated as Observation Influence (OI). In general, the observation influence decreases with the increase of r. For highly correlated ($\alpha = 0.9$, $\beta = 0.9$) and diagonal $(\alpha = 0, \beta = 0)$ **R** and **B** matrices, the observation influence as a function of r is the same (solid grey line and dash thick line, respectively). Maximum observation influence is achieved when **B** is diagonal $(\alpha = 0)$ and **R** is highly correlated $(\beta = 0.9)$ (thin black line). The observation influence will constantly decrease from the 'maximum curve' with the decrease of the correlation degree in **R** (**B** still diagonal). And the minimum observation influence curve is achieved when **R** is diagonal $(\beta = 0)$ and **B** is highly correlated $(\alpha = 0.9)$ (thick solid line). It is worth to notice that if the observation error variance is larger than the background error variance $(\sigma_o^2 > \sigma_b^2)$ introducing the observation error correlation will slightly increase the observation influence and for $\sigma_o^2 \gg \sigma_b^2$ the observations will not be more influent in the analysis despite **R** is not diagonal.

(i) ***R*** *diagonal and* ***B*** *non-diagonal* $(\alpha \neq 0, \beta = 0)$. Equations (24) and (25) reduce respectively to

$$S_{11} = S_{22} = \frac{r+1-\alpha^2}{r^2+1-\alpha^2+2r} \tag{26}$$

$$S_{12} = S_{21} = \frac{r\alpha}{r^2+1-\alpha^2+2r} \tag{27}$$

It can be seen that if the observations are very close (compared to the scale-length of the background error correlation), i.e $\alpha \sim$1 (data dense area), then

$$S_{11} = S_{22} = S_{12} = S_{21} \cong \frac{1}{r+2} \tag{28}$$

Furthermore, if $\sigma_b = \sigma_o$, that is $r = 1$, we have three pieces of information with equal accuracy and $S_{11} = S_{22} = 1/3$. The background sensitivity at both locations is $1 - S_{11} = 1 - S_{22} = 2/3$. If the observation is much more accurate than the background ($\sigma_b \gg \sigma_o$), that is $r \sim 0$, then both observations have influence $S_{11} = S_{22} = 1/2$, and the background sensitivities are $1 - S_{11} = 1 - S_{22} = 1/2$.

Lets now turn to the dependence on the background-error correlation α, for the case $\sigma_b = \sigma_o$ $(r = 1)$. It is

$$S_{11} = S_{22} = \frac{2-\alpha^2}{4-\alpha^2} \tag{29}$$

$$S_{12} = S_{21} = \frac{\alpha}{4-\alpha^2} \tag{30}$$

If the locations are far apart, such that $\alpha \sim 0$, then $S_{11} = S_{22} = 1/2$, the background sensitivity is also $\frac{1}{2}$ and $S_{12} = S_{21} = 0$. It can be concluded that where observations are sparse, S_{ii} and the background-sensitivity are determined by their relative accuracies (r) and the off-diagonal terms are small (indicating that surrounding observations have small influence). Conversely, where observations are dense, S_{ii} tends to be small, the background-sensitivities tend to be large and the off-diagonal terms are also large.
It is also convenient to summarize the case $\sigma_b = \sigma_o$ $(r = 1)$ by showing the projected analysis at location *1*

$$\hat{y}_1 = \frac{1}{4-\alpha^2}[(2-\alpha^2)y_1 + 2x_1 - \alpha(x_2 - y_2)] \tag{31}$$

The estimate $\hat{y}_1$ depends on y_1, x_1 and an additional term due to the second observation. It is noticed that, with a diagonal **R**, the observational contribution is generally devalued with respect to the background because a group of correlated background values count more than the single observation $[\alpha \to \pm 1, (2-\alpha^2) \to 1]$. From the expression above we also see that the contribution from the second observation is increasing with the correlations absolute value, implying a larger contribution due to the background x_2 and observation y

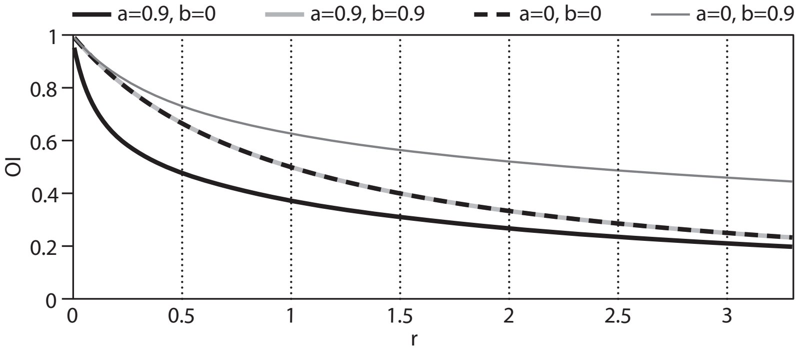

*Figure 1: Self-Sensitivities or Observation Influence (OI) as a function of the ratio between the observation error variance and the background error variance. Four different cases are shown: highly correlated **B** and uncorrelated **R** (thick black line). Highly correlated **R** and highly correlated **B** (thick grey line). Uncorrelated **B** and highly correlated **R** (thin grey line). Uncorrelated **R** and uncorrelated **B** (dashed black line).*

4 Results

The diagonal elements of the influence matrix have been computed for the operational 4D-Var assimilation system at T159 spectral truncation 91 model levels for October 2011. For the calculation details see

Cardinali *et al* 2004 and Appendix B. The observation departures $(\mathbf{y}\text{-}\mathbf{Hx}_b)$ were calculated by comparing the observations with a 12-hour forecast integration at T511 resolution. The assimilated observations for each main observation type are given in Table 1. A large proportion $\sim$ (98%) of the used data is provided by satellite systems.

(a) Trace diagnostic: Observation Influence and DFS

The global average Observation Influence (OI) is defined as

$$OI = \frac{tr(\mathbf{S})}{p} \tag{32}$$

where m is the total number of observations. For October 2011 OI=0.18. Consequently, the average background global influence to the analysis at observation points is equal to 0.82 (see 15). It is clear that in the ECMWF system the global observation influence is quite low.

In Fig. 2 the *OI* for the all different observation types is plotted. In general, *OI* of conventional observations (SYNOP, DRIBU, PROFILER, PILOT, DROP, TEMP, Aircraft) is larger than the satellite one. The largest *OI* is provided by DRIBU surface pressure observations because they are located over the oceans that are in general very poor observed (less than continental areas). Moreover, DRIBU and SYNOP observations are very high quality measurements and the observation error variances is quite small, likely smaller than the background error variance (see 'toy model' in section 3 (c)). Similarly, the $OI\sim$0.4-0.5 of the remaining conventional data is due to their quite small observation error variance. In Section 3 (c) it has been proved that if R is diagonal the *OI* is bounded between (0,1) but from Fig. 2, we can see that DRIBU *OI* is higher than 1. This is due to the approximation of the numerical solution and, in particular, the use in the influence matrix calculation of an estimate of the analysis covariance matrix A (see Cardinali *et al* 2004 for details). On the contrary, the *OI* influence of satellite data is quite small. The largest influence is provided by GPS-RO observations ($\sim$0.4) which again are accurate data and likely with uncorrelated observation error (Healy and Thpaut 2006), followed by AMSU-A measurements ($\sim$0.3). All the other observations have an influence of about 0.2. Recently, changes in the assimilation of 'All-Sky' observations (TMI and SSMIS) have increased their influence in the analysis (Cardinali and Prates 2011, Geer *et al* 2011)

In Section 2 it has been shown that tr($\mathbf{S}$) can be interpreted as a measure of the amount of information extracted from the observations. In fact, in non-parametric statistics, tr($\mathbf{S}$) measures the 'equivalent number of parameters' or *degrees of freedom for signal (DFS)*. Having obtained values of all the diagonal elements of $\mathbf{S}$ (using 16) we can now obtain reliable estimates of the information content in any subset of the observational data. However, it must be noted that this theoretical measure of information content does not necessarily translate on value of forecast impact. Figure 3 shows the information content for all main observation types. It can be seen that AMSU-A radiances are the most informative data type, providing 23% of the total observational information, IASI follows with 17% and AIRS with 16%. The information content of Aircraft (10%) is the largest among conventional observations, followed by TEMP and SYNOP ($\sim$4%). Noticeable is the 7% of GPS-RO (4^{th} in the satellite *DFS* ranking) that well combines with the 0.4 value for the average observation influence. In general, the importance of the observations as defined by e.g. the *DFS* well correlates with the recent data impact studies by Radnoti *et al*, (2010).

Data name	Data kind	Information
OZONE (O3)	Backscattered solar UV radiation, retrievals	Ozone, stratosphere
GOES-Radiance	US geostationary satellite infrared sounder radiances	Moisture, mid/upper troposphere
MTSAT-Rad	Japanese geostationary satellite infrared sounder radiances	Moisture, mid/upper troposphere
MET-rad	EUMETSAT geostationary satellite infrared sounder radiances	Moisture, mid/upper troposphere
AMSU-B	Microwave sounder radiances	Moisture, troposphere
MHS	Microwave sounder radiances	Moisture, troposphere
MERIS	Differential reflected solar radiation, retrievals	Total column water vapour
GPS-RO	GPS radio occultation bending angles	Temperature, surface pressure
IASI	Infrared sounder radiances	Temperature, moisture, ozone
AIRS	Infrared sounder radiances	Temperature, moisture, ozone
AMSU-A	Microwave sounder radiances	Temperature
HIRS	Infrared sounder radiances	Temperature, moisture, ozone
ASCAT	Microwave scatterometer backscatter coefficients	Surface wind
MODIS-AMV	US polar Atmospheric Motion Vectors, retrievals	Wind, troposphere
Meteosat-AMV	EUMETSAT geostationary Atmospheric Motion Vectors, retrievals	Wind, troposphere
MTSAT-AMV	Japanese geostationary Atmospheric Motion Vectors, retrievals	Wind, troposphere
GOES-AMV	US geostationary Atmospheric Motion Vectors, retrievals	Wind, troposphere
PROFILER	American, European and Japanese Wind profiles	Wind, troposphere
PILOT	Radiosondes at significant level from land stations	Wind, troposphere
DROP	Dropsondes from aircrafts	Wind, temperature, moisture, pressure, troposphere
TEMP	Radiosondes from land and ships	Wind, temperature, moisture, pressure, troposphere
Aircraft	Aircraft measurements	Wind, temperature, troposphere
DRIBU	Drifting buoys	Surface pressure, temperature, moisture, wind
SYNOP	Surface Observations at land stations and on ships	Surface pressure, temperature, moisture, wind

Table 1: Observation type assimilated on October 2011. The total number of data in one assimilation cycle is on average m$\sim$25,000,000.

Similar information content of different observation types may be due to different reasons. For example, DRIBU and OZONE information content is similarly small but whilst OZONE observations have a very small average influence (Fig.2) and dense data coverage, DRIBU observations have large mean influence but much lower data counts (Fig.2). Anyhow, the OZONE data are important for the ozone assimilation in spite of their low information content per analysis cycle. In fact, OZONE is generally a long-lived species, which allows observational information to be advected by the model over periods of several days.

The difference between *OI* and *DFS* comes from the number of observation assimilated. Therefore, despite the generally low observation influence of satellite measurements, they show quite large *DFS* because of the large number assimilated. A large discrepancy between *OI* and *DFS* points on those observation types where a revision of the assigned covariance matrices **R** and **B** will be beneficial: more information extracted from e.g. satellite measurements.

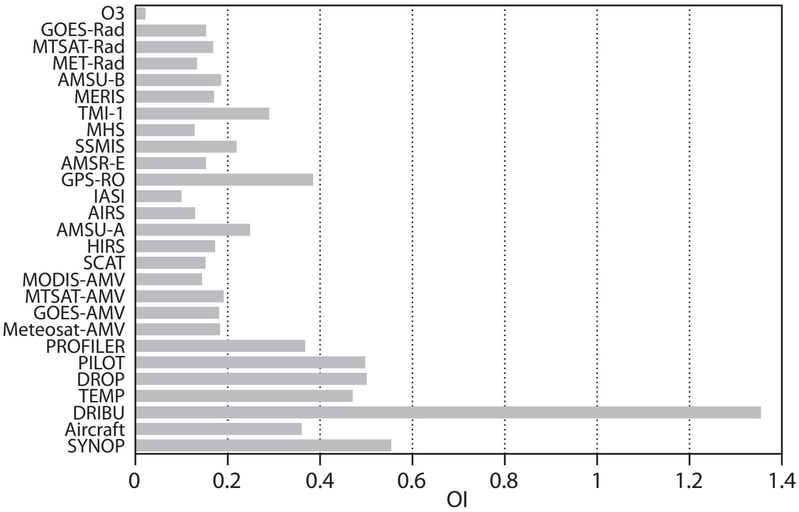

Figure 2: Observation Influence (OI) of all assimilated observations in the ECMWF 4DVar system in October 2011. Observation types are described in Table 1.

Another index of interest is the partial Observation Influence (OI_p) for any selected subset of data

$$OI_p = \frac{\sum_{i \subset I} S_{ii}}{p_I} \tag{33}$$

where p_I is the number of data in subset I. The subset I can represent a specific observation type, a specific vertical or horizontal domain or a particular meteorological variable. In Fig. 4 the *OI* of Aircraft data (I) is plotted as a function of pressure layers and for all observed parameters: temperature (t), zonal (u) and meridional (v) component of the wind. The largest *OI* is provided by temperature observations (∼0.4) similar distributed on the different pressure layers. Wind observations have larger influence (0.4) on the top of the atmosphere (above 400 hPa) than on the bottom one (0.2) due to the fact that there are very few wind observations on the troposphere and lower stratosphere mainly over the oceans. At those levels, temperature information is also provided by different satellite platforms (in terms of brightness temperature or radiance). In Fig. 5 the Aircraft *DFS* with respect to different pressure levels and observed parameters is shown. The largest *DFS* in the lower troposphere (below 700 hPa) for temperature measurements (∼10% with respect to the total Aircraft *DFS*) with respect to wind ones is due to the

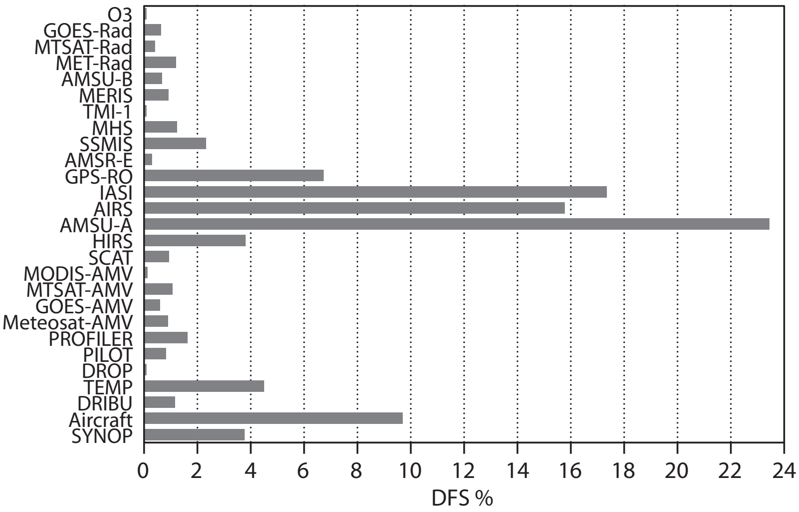

Figure 3: Degree of Freedom for Signal (DFS) of all observations assimilated in the ECMWF 4DVar system in October 2011. Observation types are described in Table 1.

largest temperature influence. For all the other levels, the *DFS* is quite similar to the *OI* distribution with the exception of the layer from 200 to 300 hPa where the increase to ∼50% is due to the increase of number of observations assimilated. Figure 6 and 7 shows the AMSU-A *OI* and *DFS*, respectively, for all the channels assimilated. A large part of the AMSU-A information is with respect to stratospheric temperature and the largest *OI* at that atmospheric layer is from channel 9 and 10 (∼0.4) (Fig. 6). Channel 5 (∼700 hPa) shows a very large ∼0.8 *OI*, the largest influence among all the channels. The reason of this large *OI* is unclear, and investigation is in due course to understand the cause. The channels observation influence distribution is similar to the *DFS* distribution (Fig. 7): channel 9 and 10 count for 18% of the AMSU-A DFS and channel 5 for 24%.

(b) Geographical map of OI

The geographical map of Observation Influence for SYNOP and DRIBU surface pressure observations is shown in Fig.8. Each box indicates the observation influence per observation location averaged among all the October 2011 measurements. Data points with influence greater than one are due to the approximation of the computed diagonal elements of influence matrix (see Cardinali *et al*, 2004 and Appendix B).

Low-influence data points have large background influence (see 14 and 15), which is the case in data-rich areas such as North America and Europe (observation influence ∼0.2) (see also Section 3 (c)). In data-sparse areas individual observations have larger influence: in the Polar regions, where there are only few isolated observations, the *OI* is very high (theoretically ∼1) and the background has very small influence on the analysis.
In dynamically active areas (Fig. 8: e.g. North Atlantic and North Pacific), several fairly isolated observations have large influence on the analysis. This is also due to the evolution of the background-error covariance matrix as propagated by the forecast model in 4D-Var (Thpaut *et al*. 1993, 1996). As a result, the data assimilation scheme can fit these observations more closely.

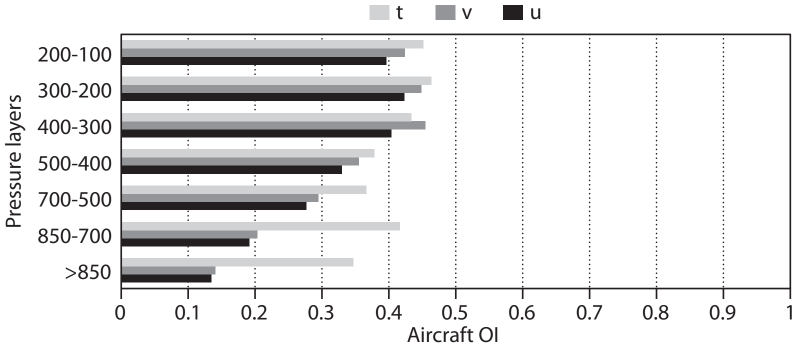

Figure 4: Observation Influence (OI) for Aircraft observations and for October 2011 grouped by pressure layer and observed parameter. Parameters are temperature (t) light grey bar; meridional wind (v) dark grey bar and zonal wind (u) black bar.

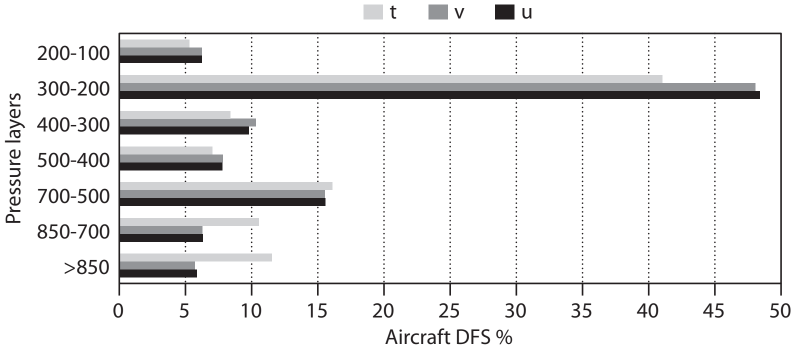

Figure 5: Degree of Freedom for Signal (DFS) in percentage for Aircraft observations and for October 2011 grouped by pressure layer and observed parameter. Parameters are temperature (t) light grey bar; meridional wind (v) dark grey bar and zonal wind (u) black bar. The percentage is relative to the total Aircraft DFS.

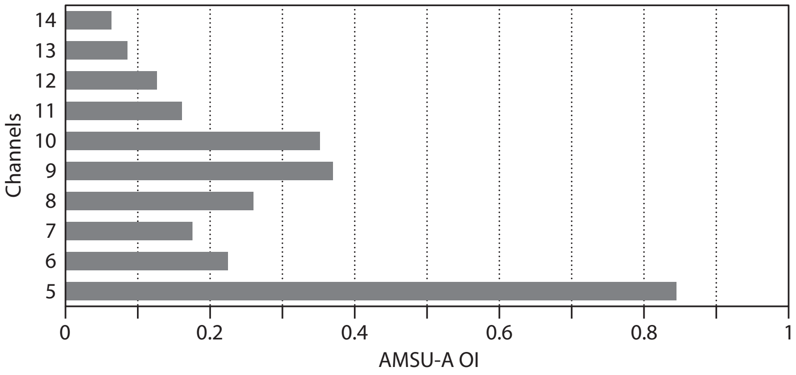

Figure 6: Observation Influence (OI) for AMSU-A observations and for October 2011 grouped by channels.

Similar features can be seen in Fig. 9, which shows the influence of u-component wind observations for Aircraft data above 400 hPa. Isolated flight tracks over Atlantic and Pacific oceans show larger influences than measurements over data-dense areas over America and Europe. The flight tracks over North Atlantic and North Pacific are also in dynamically active areas where the background error variances are implicitly inflated by the evolution of the background-error covariance matrix in the 4D-Var window. Figure 10 shows the geographical distribution of AMSU-A channel 8 observation influence. The largest

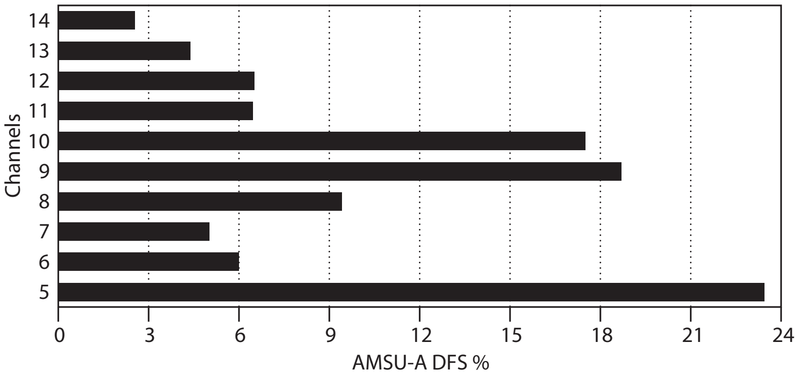

Figure 7: Degree of Freedom for Signal in percentage (DFS) for AMSU-A observations and for October 2011 grouped by channels. The percentage is relative to the total AMSU-A DFS.

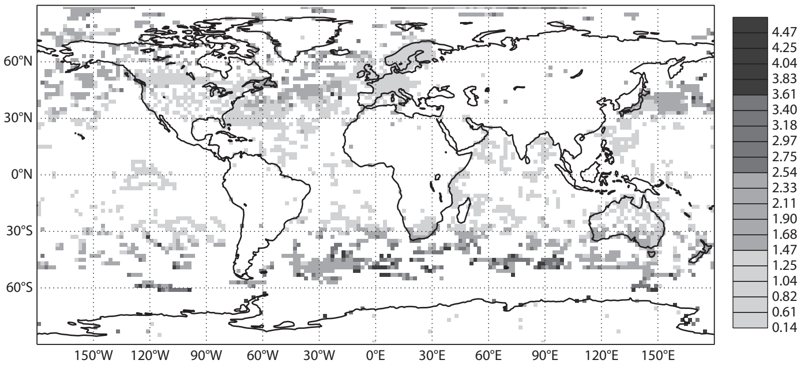

Figure 8: Observation Influence (OI) of SYNOP and DRIBU surface pressure observations for October 2011. High influential points are close to 1 and low influential points are close to 0.

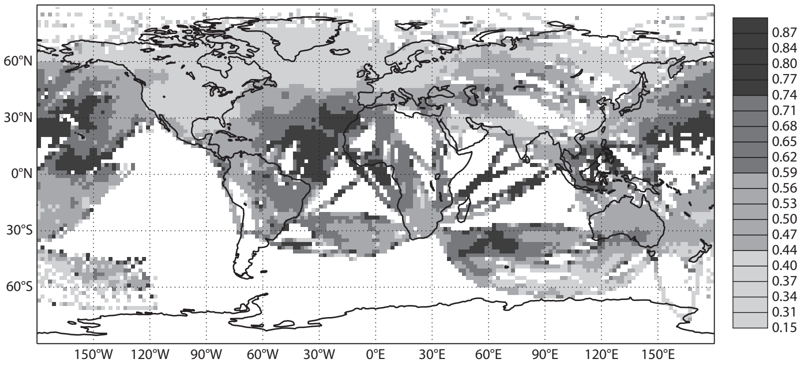

Figure 9: Observation Influence (OI) of Aircraft zonal wind component above 400 hPa for October 2011. High influential points are close to 1 and low influential points are close to 0.

influence is noticed in the extra-tropics and polar areas ($\sim$0.4) whilst in the tropics the maximum *OI* is $\sim$0.12. Since channel 8 observation error variances are geographically constant the main difference in the observed *OI* pattern is likely due to the **B** covariance matrix. It looks that either the background error correlation are higher or the background error variance are larger in the tropics than in the extra-tropics.

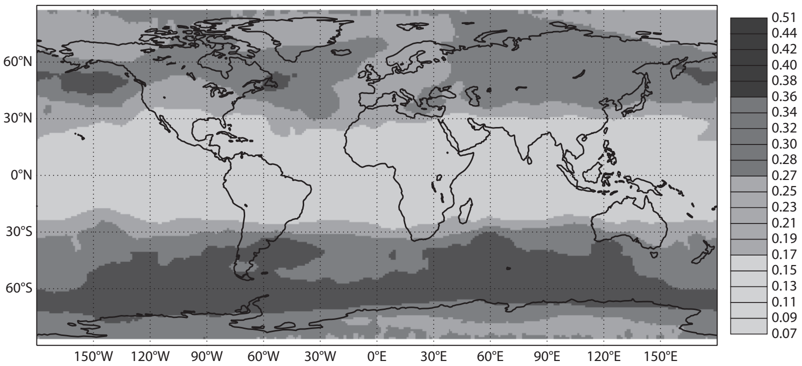

Figure 10: Observation Influence (OI) of AMSU-A channel 8 for October 2011. High influential points are close to 1 and low influential points are close to 0.

5 Conclusions

The influence matrix is a well-known concept in multi-variate linear regression, where it is used to identify influential data and to predict the impact on the estimates of removing individual data from the regression. In this paper the influence matrix in the context of linear statistical analysis schemes has been derived, as used for data assimilation of meteorological observations in numerical weather prediction (Lorenc 1986). In particular an approximate method to compute the diagonal elements of the influence matrix (the self-sensitivities or observation influence) in ECMWFs operational data assimilation system (4D-Var) has been derived and implemented. The approach necessarily approximates the solution due to the large dimension of the estimation problem at hand: the number of estimated parameters is of the order 10^9, and the number of observational data is around 25×10^6.

The self-sensitivity provides a quantitative measure of the observation influence in the analysis. In robust regression, it is expected that the data have similar self-sensitivity (sometimes called leverage) - that is, they exert similar influence in estimating the regression line. Disproportionate data influence on the regression estimate can have different reasons: First, there is the inevitable occurrence of incorrect data. Second, influential data points may be legitimately occurring extreme observations. However, even if such data often contain valuable information, it is constructive to determine to which extent the estimate depends on these data. Moreover, diagnostics may reveal other patterns e.g. that the estimates are based primarily on a specific sub-set of the data rather than on the majority of the data.

In the context of 4D-Var there are many components that together determine the influence given to any one particular observation. First there is the specified observation error covariance **R**, which is usually well known and obtained simply from tabulated values. Second, there is the background error covariance **B**, which is specified in terms of transformed variables that are most suitable to describe a large proportion of the actual background error covariance. The implied covariance in terms of the observable quantities is not immediately available for inspection, but it determines the analysis weight given to the data. Third, the dynamics and the physics of the forecast model propagate the covariance in time, and modify it according to local error growth in the prediction. The influence is further modulated by data density. Examples for surface pressure and aircraft wind observations have been shown indicating that low influence data points occur in data-rich areas while high influence data points are in data-sparse regions or in dynamically active areas. Background error correlations also play an important role. In fact, very high correlations drastically lessen the observation influence (it is halved in the idealized example presented in Section 3 (c)) in favour of background influence and amplify the influence of the surrounding observations. The observation influence pattern of AMSU-A channel 8 suggests some affectation of the correlation expresses by the **B** covariance matrix.

The global observation influence per assimilation cycle has been found to be 18%, and consequently the background influence is 82%. Thus, on average the observation influence is low compared to the influence of the background (the prior). However, it must be taken into account that the background contains observation information from the previous analysis cycles. The theoretical information content (the degrees of freedom for signal) for each of the main observation types was also calculated. It was found that AMSU-A radiance data provide the most information to the analysis, followed by IASI, AIRS, Aircraft, GPS-RO and TEMP. In total, about 20% of the observational information is currently provided by surface-based observing systems, and 80% by satellite systems. It must be stressed that this ranking is not an indication of relative importance of the observing systems for forecast accuracy. Nevertheless, recent studies on the 24-hour observation impact on the forecast with the adjoint methodology have shown similar data ranking (Langland and Baker, 2004; Zhu and Gelaro 2008; Cardinali 2009)

If the influence matrix were computed without approximation then all the self-sensitivities would have been bounded in the interval zero to one. With the approximate method used, out-of-bound self-sensitivities occur if the Hessian representation based on an eigen-vector expansion is truncated, especially when few eigen-vectors are used. However, it has been shown that this problem affects only a small percentage of the self-sensitivities computed, and in particular those that are closer to one.

Self-sensitivities provide an objective diagnostic on the performance of the assimilation system. They could be used in observation quality control to protect against distortion by anomalous data; this aspect has been explored by Junjie *et al*. (2009) in the context of Ensemble Kamlan Filter where the **B** is well known and the solution for the diagonal element of the Influence matrix is therefore very accurate. Junjie *et al*. (2009) have shown that the leaving-out-one observation, not practical for large system dimension, can be replaced by the Self-sensitivities (Eq. 9) that provide a similar diagnostic without performing separate least square regressions. Self-sensitivities also provide indication on model and observation error specification and tuning. Incorrect specifications can be identified, interpreted and better understood through observation influence diagnostics, partitioned e.g. by observation types, variable, levels, and regions.

In the near future more satellite data will be used and likely be thinned. Thinning has to be performed either to reduce the observation error spatial correlation (Bormann *et al*. 2003) or to reduce the computational cost of the assimilation. The observation influence provides an objective way of selecting observations dependent on their local influence on the analysis estimate to be used in conjunction with forecast impact assessments. Recently, Bauer *et al* (2011) have shown that satellite measurements in sensitive areas as defined by singular vectors methodology have larger impact in the forecast than measurements in different regions and also larger or similar impact than the full amount of data. In this case, a dynamical thinning can be thought that selects, at every assimilation cycle, the most influent measurements partition of a particular remote sensing instrument, from information based on the previous cycle (see also Rabier *et al.*, 2002). Clearly, it can be assumed that components of the observing network remain constant and the background error variances remain almost unchanged for close assimilation cycles.

Acknowledgements

The author thanks Olivier Talagrand and Sergio Pezzulli for the fruitful discussions on the subject. Many thanks are given to Mohamed Dahoui and Anne Fouilloux for their precious technical support.

A Appendix: Influence Matrix Calculation in Weighted Regression Data Assimilation Scheme

Under the frequentist approach, the regression equations for observation

$$\mathbf{y} = \mathbf{H}\theta + \varepsilon_{\mathbf{o}} \tag{A.1}$$

and for background

$$\mathbf{x_b} = \theta + \varepsilon_{\mathbf{b}} \tag{A.2}$$

are assumed to have uncorrelated error vectors ε_o and ε_b, zero vector means and variance matrices **R** and **B**, respectively. The parameter is the unknown system state (**x**) of dimension n. These regression equations are summarized as a weighted regression

$$\mathbf{z} = \mathbf{X}\theta + \varepsilon \tag{A.3}$$

where $\mathbf{z} = [\mathbf{y^T x_b^T}]^{\mathbf{T}}$ is (m+n)×1 ; $\mathbf{X} = [\mathbf{H^T I_n}]^{\mathbf{T}}$ is (m+n)×n and $\varepsilon = [\varepsilon_o \varepsilon_b]^T$ is (m+n)×1 with zero mean and variances matrix

$$\Omega = \begin{pmatrix} \mathbf{R} & 0 \\ 0 & \mathbf{B} \end{pmatrix} \tag{A.4}$$

The generalized LS solution for θ is BLUE and is given by

$$\hat{\theta} = (\mathbf{X^T}\Omega^{-1}\mathbf{X})\mathbf{X^T}\Omega^{-1}\mathbf{z} \tag{A.5}$$

see Talagrand (1997). After some algebra this equation equals (11). Thus

$$\mathbf{z} = \mathbf{X}\hat{\theta} = [\mathbf{H^T x_a^T x_a^T}]^{\mathbf{T}} = \mathbf{X}(\mathbf{X^T}\Omega^{-1}\mathbf{X})^{-1}\mathbf{X^T}\Omega^{-1}\mathbf{z} \tag{A.6}$$

and by (5) the influence matrix becomes

$$\mathbf{S_{zz}} = \frac{\delta\hat{\mathbf{z}}}{\delta\mathbf{z}} = \frac{\delta\mathbf{X}\hat{\theta}}{\delta\mathbf{z}} = \begin{pmatrix} \mathbf{S_{yy}} & \mathbf{S_{yb}} \\ \mathbf{S_{by}} & \mathbf{S_{bb}} \end{pmatrix} = \begin{pmatrix} \mathbf{R^{-1}HAH^T} & \mathbf{R^{-1}HA} \\ \mathbf{B^{-1}AH^T} & \mathbf{B^{-1}A} \end{pmatrix} \tag{A.7}$$

where $\mathbf{S_{yy}} = \frac{\delta\mathbf{Hx_a}}{\delta\mathbf{y}}$; $\mathbf{S_{yb}} = \frac{\delta\mathbf{x_a}}{\delta\mathbf{y}}$; $\mathbf{S_{by}} = \frac{\delta\mathbf{Hx_a}}{\delta\mathbf{x_b}}$; $\mathbf{S_{bb}} = \frac{\delta\mathbf{x_a}}{\delta\mathbf{x_b}}$.Note that $\mathbf{S}_{yy}$=**S** as defined in (14).

Generalized LS regression is different from ordinary LS because the influence matrix is not symmetric anymore. For idempotence, using A.1 it easy to show that $\mathbf{S}_{zz}\mathbf{S}_{zz}$=$\mathbf{S}_{zz}$. Finally,

$$\mathbf{S_{bb}} = \mathbf{B}^{-\mathbf{1}}\mathbf{A} = \mathbf{I_n} - \mathbf{H^T}\mathbf{R}^{-\mathbf{1}}\mathbf{HA} \tag{A.8}$$

hence,

$$tr(\mathbf{S_{bb}}) = n - tr(\mathbf{H^T}\mathbf{R}^{-\mathbf{1}}\mathbf{HA}) = n - tr(\mathbf{S_{yy}}) \tag{A.9}$$

it follows that

$$tr(\mathbf{S_{zz}}) = tr(\mathbf{S_{yy}}) + tr(\mathbf{S_{bb}}) = n \tag{A.10}$$

The trace of the influence matrix is still equal to the parameters dimension.

B Appendix: Approximate calculation of self-sensitivity in a large variational analysis system

In a optimal variational analysis scheme, the analysis error covariance matrix **A** is approximately the inverse of the matrix of second derivatives (the Hessian) of the cost function J, i.e. $\mathbf{A}$=$(\mathbf{J}'')^{-1}$ (Rabier and Courtier, 1992). Given the large dimension of the matrices involved, **J** and its inverse cannot be computed explicitly. Following Fisher and Courtier (1995) we use an approximate representation of the Hessian based on a truncated eigen-vector expansion with vectors obtained through the Lanczos algorithm. The calculations are performed in terms of a transformed variable $\chi, \chi = \mathbf{L}^{-\mathbf{1}}(\mathbf{x} - \mathbf{x_b})$, with **L** chosen such that $\mathbf{B}^{-1}$=$\mathbf{L^T L}$. The transformation **L** thus reduces the covariance of the prior to the identity matrix. In variational assimilation **L** is referred to as the change-of-variable operator (Courtier *et al*. 1998).

$$\mathbf{J}''^{-\mathbf{1}} = \mathbf{B} - \sum_{\mathbf{i=1}}^{\mathbf{M}} \frac{\mathbf{1} - \lambda_\mathbf{i}}{\lambda_\mathbf{i}} (\mathbf{L}v_\mathbf{i})(\mathbf{L}v_\mathbf{i})^\mathbf{T} \tag{B.1}$$

The summation in B.1 approximates the variance reduction **B-A** due to the use of observations in the analysis. (λ_i, v_i) are the eigen-pairs of **A**. The Hessian eigen-vectors are also used to precondition the minimization (Fisher and Andersson, 2001). The computed eigen-values are not used to minimize the cost function but only to estimate the analysis covariance matrix. It is well known, otherwise, that the minimization algorithm is analogous to the conjugate-gradient algorithm. Because the minimum is found within an iterative method, the operational number of iterations is sufficient to find the solution (with the required accuracy) but does not provide a sufficient number of eigen-pairs to estimate the analysis error variances.

The diagonal of the background error covariance matrix **B** in B.1 is also computed approximately, using the randomisation method proposed by Fisher and Courtier (1995). From a sample of N random vectors $\mathbf{u}_i$ (in the space of the control-vector χ), drawn from a population with zero mean and unit Gaussian variance, a low-rank representation of **B** (in terms of the atmospheric state variables x) is obtained by

$$\mathbf{B} = \frac{1}{N} \sum_{i=1}^{N} (\mathbf{L}u_i)(\mathbf{L}u_i)^\mathbf{T} \tag{B.2}$$

This approximate representation of **B** has previously been used by Andersson *et al*. (2000) to diagnose background errors in terms of observable quantities, i.e. $\mathbf{HBH}^T$.

Inserting B.1 and B.2 into (16) an approximate method for calculating **S** is achieved, that is practical for a large dimension variational assimilation (both 3D and 4D-Var):

$$\mathbf{S} = \mathbf{R}^{-1}\mathbf{H}[\frac{\mathbf{1}}{\mathbf{N}}\sum_{\mathbf{i=1}}^{\mathbf{N}}(\mathbf{L}u_i)(\mathbf{L}u_i)^{\mathbf{T}} - \sum_{\mathbf{i=1}}^{\mathbf{M}}\frac{\mathbf{1}-\lambda_{\mathbf{i}}}{\lambda_{\mathbf{i}}}(\mathbf{L}v_{\mathbf{i}})(\mathbf{L}v_{\mathbf{i}})^{\mathbf{T}}]\mathbf{H}^{\mathbf{T}} \tag{B.3}$$

Only the diagonal elements of **S** are computed and stored - that is, the analysis sensitivities with respect to the observations, or self-sensitivities S_{ii}. The cross-sensitivity S_{ij} for i≠j, that represents the influence of the j^{th} observation to the analysis at the i^{th} location, is not computed. Note that the approximation of the first term is unbiased, whereas the second term is truncated such that variances are underestimated. For small M the approximate S_{ii} will tend to be over-estimates. For the extreme case M=0 Eq.(B.3) gives **S=R**$^{-1}$**HBH**T which in particular can have diagonal elements larger than one if elements of **HBH**^{T} are larger than the corresponding elements of **R**. The effect of the approximation on the calculated S_{ii} values is investigated.
Approximations in both of the two terms of (B.3) contribute to the problem. For the second term, the degree of over-estimation depends on the structure of the covariance reduction matrix **B-A**.

For an analysis in which observations lead to strongly localised covariance reduction (such as the surface pressure analysis with its short co-variance length scales ∼180 km, and large observational impacts) a large M is required to approximate **B-A** accurately.

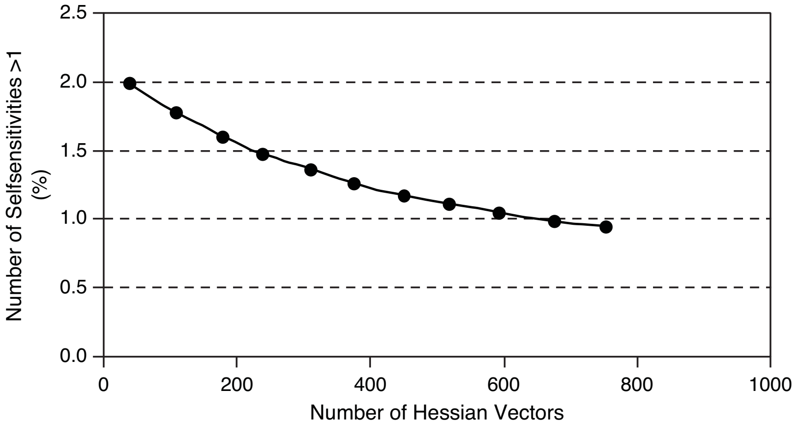

Figure 11: Proportion of self-sensitivity values greater than one (%) versus the number of Hessian vectors used to compute the variances of ***B-A****, using the approximate method described in B.3.*

In Fig. 11, the proportion of observations for which $S_{ii}>1$ is plotted versus M, the number of Hessian vectors used. The plot shows a gradual decrease of $S_{ii} > 1$ as M increases, as expected. The curve seems to approach 10,000 $S_{ii}>1$ (0.7% in the plot) for M somewhere between 1,000 and 2,000. However, increasing the number of Hessian vectors slightly increases the number of self-sensitivities less than zero (by 0.5%). This problem can be understood by looking at the approximations introduced through the first term of (B.3). The truncation N of the first term determines the randomisation sample size: larger N leads to smaller noise. The noise is unbiased - that is, the term is neither over nor under-estimated on average. The randomisation noise in the diagonal elements is in the order 10% with N=50 (Andersson *et al*. 2000). With N=500, values $S_{ii}<0$ have all disappeared.

References

Andersson, E., Fisher, M. Munro, R. and McNally, A, 2000: Diagnosis of background errors for radiances and other observable quantities in a variational data assimilation scheme, and the explanation of a case of poor convergence. *Q. J. R. Meteorol. Soc.* **126,** 1455-1472.

Bauer, P., R. Buizza, C. Cardinali and J- l Thpaut, 2011: Impact of singular vector based satellite data thinning on NWP. *Q. J. R. Meteorol. Soc*, **137**, 286-302.

Bormann, N., Saarinen, S., Kelly G. and Thpaut, J-N. 2003: The spatial structure of observation errors in atmospheric motion vectors from geostationary satellite data. *Mon. Wea. Rev.*, **131**, 706-718.

Cardinali, C., S. Pezzulli and E. Andersson, 2004: Influence matrix diagnostics of a data assimilation system. *Q. J. R. Meteorol. Soc.*, **130**, 2767-2786

Cardinali, C, 2009: Monitoring the forecast impact on the short-range forecast. *Q. J. R. Meteorol. Soc.*, **135**, 239-250.

Cardinali, C., and F. Prates, 2011: Performance measurement with advanced diagnostic tools of all-sky microwave imager radiances in 4D-Var *Q. J. R. Meteorol. Soc*, **137**, Issue 661, Part B, 20382046.

Courtier, P., Andersson, E., Heckley, W., Pailleux,, J., Vasiljevic, Hamrud, D. M., Hollingsworth, A., Rabier, F. and Fisher, M., 1998: The ECMWF implementation of three-dimensional variational assimilation (3D-Var). Part I: Formulation. *Q. J. R. Meteorol. Soc.* **124**, 1783-1807.

Craven, P., and Wahba, G., 1979: Smoothing noisy data with spline functions: estimating the correct degree of smoothing by the method of generalized cross-validation. *Numer. Math.*, **31**, 377-403.

Geer A.J. and P. Bauer 2011: Observation errors in all-sky data assimilation. *Q. J. R. Meteorol. Soc.* **137**, Issue 661, Part B, 20242037.

Fisher, M. and Courtier, P., 1995: Estimating the covariance matrices of analysis and forecast error in variational data assimilation. ECMWF Tech Memo., 220, pp 26.

Junjie, L., E., Kalnay, T., Miyoshi and C., Cardinali, 2009: Analysis sensitivity calculation within an Ensemble Kalman filter. *Q. J. R. Meteorol. Soc*. **135**, 1842-1851.

Healy, S. B. and J.-N., Thpaut, 2006: Assimilation experiments with CHAMP GPS radio occultation measurements. *Q. J. R. Meteorol. Soc.* **132**, 605-623.

Hoaglin, D. C., Mosteller, F. and Tukey J.W., 1982. Understanding Robust and Exploratory Data Analysis. *Wiley Series in Probability and Statistics*

Hoaglin, D. C., and Welsch, R. E. 1978: The hat matrix in regression and ANOVA. *The American Statisticians*, **32**, 17-22 and *Corrigenda* **32**, 146.

Langland R. and N.L Baker., 2004: Estimation of observation impact using the NRL atmospheric variational data assimilation adjoint system. *Tellus*, **56A**, 189-201.

Lorenc, A., 1986: Analysis methods for numerical weather prediction. *Q. J. R. Meteorol. Soc.*, **112**, 1177-1194.

Rabier, F., Jrvinen, H., Klinker, E., Mahfouf J.F., and Simmons, A., 2000: The ECMWF operational implementation of four-dimensional variational assimilation. Part I: experimental results with simplified physics. *Q. J. R. Meteorol. Soc*. **126**, 1143-1170.

Rabier, F., Fourri, N., Chafa D.and Prunet, P., 2002: Channel selection methods for infrared atmospheric sounding interferometer radiances. *Q. J. R. Meteorol. Soc.*, **128**, 1011-1027.

Rabier, F. and Courtier, P., 1992: Four-dimensional assimilation in the presence of baroclinic instability. *Q. J. R. Meteorol. Soc.*, **118**, 649-672.

Radnoti, G., P. Bauer, A. McNally, C. Cardinali, S. Healy and P. de Rosnay, 2010: ECMWF study on the impact of future developments of the space-based observing system on Numerical Weather Prediction. ECMWF Tec. Memo 638

Shen, X., Huang H., and Cressie N., 2002: Nonparametric hypothesis testing for a spatial signal. J.Am.Stat.Ass., **97**, 1122-1140

Talagrand, O., 1997: Assimilation of observations, an Introduction. *J. Meteorol. Soc. Japan*, **Vol 75**, N.1B,191-209.

Thpaut, J.N., Hoffman, R.N. and Courtier, P., 1993: Interactions of dynamics and observations in a four-dimensional variational assimilation. *Mon. Wea. Rev.*, **121**, 3393-3414.

Thpaut, J.N., Courtier, P., Belaud G. and Lematre, G., 1996: Dynamical structure functions in four-dimensional variational assimilation: A case study. *Q. J. R. Meteorol. Soc.*, **122**, 535-561.

Tukey, J. W. 1972: Data analysis, Computational and Mathematics. *Q. of Applied mathematics*, **30**, 51-65

Velleman, P. F., and Welsch, R. E., 1981: Efficient computing of regression diagnostics. *The American Statisticians*, **35**, 234-242.

Wahba, G., 1990: Spline models for observational data. SIAM, CBMS-NSF, *Regional Conference Series in Applied Mathematic*, **59**, pp 165

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.

Ye J., 1998: On measuring and correcting the effect of data mining and model selection. *J. Am. Stat. Ass.*, **93**, 120-131

Zhu, Y. and Gelaro, R., 2008: Observation Sensitivity Calculations Using the Adjoint of the Gridpoint Statistical Interpolation (GSI) Analysis System. *Mon. Wea. Rev.*, **136**, 335-3