450

Influence matrix diagnostic of a data assimilation system

Carla Cardinali[1], Sergio Pezzulli[2] and Erik Andersson[1]

Research Department

October 2003

[1] European Centre for Medium-Range Weather Forecast, UK
[2] Department of Meteorology, University of Reading, Reading, UK

To be published in Q.J.Roy.Meteor.Soc.

ECMWF
Shinfield Park
Reading
RG2 9AX
United Kingdom

For additional copies please contact: library@ecmwf.int

Series: Technical Memoranda

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

© Copyright 2004

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

Literary and scientific copyrights belong to ECMWF and are reserved in all countries. This publication is not to be reprinted or translated in whole or in part without the written permission of the Director. Appropriate non-commercial use will normally be granted under the condition that reference is made to ECMWF.

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.

Abstract

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 have been 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 Boreal Spring 2003 operational system, 15% of the global influence is due to the assimilated observations in any one analysis, and the complementary 85% is the influence of the prior (background) information, a short-range forecast containing information from earlier assimilated observations. About 25% of the observational information is currently provided by surface-based observing systems, and 75% by satellite systems.

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). Incorrect specifications of background and observation error covariance matrices can be identified, interpreted and better understood by the use of influence matrix diagnostics for the variety of observation types and observed variables used in the data assimilation system.

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 10^7 degrees of freedom and more than 10^6 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. 1983). 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 $\boldsymbol{y}$ into $\hat{\boldsymbol{y}}$, 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 we derive the corresponding concepts for linear statistical DA schemes. 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, 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 ECMWF's operational 4D-Var system) is derived in section 3.3. In section 4 we present results and selected examples related to data influence diagnostics, 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:

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon} \qquad 2.1$$

where $\mathbf{y}$ is an $m\times1$ vector for the response variable (predictand); $\mathbf{X}$ is an $m\times q$ matrix of q predictors; $\boldsymbol{\beta}$ is a $q\times1$ vector of parameters to be estimated (the regression coefficients) and $\boldsymbol{\varepsilon}$ is an $m\times1$ vector of errors (or fluctuations) with expectation $E(\boldsymbol{\varepsilon})=0$ and covariance $\mathrm{var}(\boldsymbol{\varepsilon})=\sigma^2\mathbf{I}_m$ (that is, uncorrelated observation errors). In fitting the model (2.1) 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 $\mathbf{X}$ is assumed to have full rank q.

The LS method provides the solution of the regression equation as $\boldsymbol{\beta}=(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{y}$. The fitted (or estimated) response vector $\hat{\mathbf{y}}$ is thus:

$$\hat{\mathbf{y}} = \mathbf{S}\mathbf{y} \qquad 2.2$$

where

$$\mathbf{S} = \mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T \qquad 2.3$$

is the *m×m influence matrix* (or hat matrix). It is easily seen that

$$\mathbf{S} = \frac{\partial \hat{\mathbf{y}}}{\partial \mathbf{y}} \qquad 2.4$$

and that

$$S_{ij} = \frac{\partial \hat{y}_j}{\partial y_i} \qquad S_{ii} = \frac{\partial \hat{y}_i}{\partial y_i} \qquad 2.5$$

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

$$0 \le S_{ii} \le 1 \qquad i = 1, 2, ..., m \qquad 2.6$$

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

$$\mathrm{var}(\hat{\mathbf{y}}) = \sigma^2\mathbf{S} \qquad \mathrm{var}(\mathbf{r}) = \sigma^2(\mathrm{I_m} - \mathbf{S}) \qquad 2.7$$

The trace of the influence matrix is

$$tr(\mathbf{S}) = \sum_{i=1}^{m} S_{ii} = q = rank(\mathbf{S}) \qquad 2.8$$

(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, $tr(\mathbf{I} - \mathbf{S}) = m - tr(\mathbf{S})$, 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 ith observation is deleted is

$$\hat{y}_i - \hat{y}_i^{(-i)} = \frac{S_{ii}}{(1-S_{ii})} r_i \qquad 2.9$$

where $\hat{y}_i^{(-i)}$ is the LS estimate of y_i obtained by leaving-out the ith observation of the vector $\mathbf{y}$ and the ith row of the matrix $\mathbf{X}$. 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):

$$\sum_{i=1}^{m} (y_i - \hat{y}_i^{(-i)})^2 = \sum_{i=1}^{m} \frac{(y_i - \hat{y}_i)^2}{(1-S_{ii})^2} \qquad 2.10$$

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, (2.9) shows how to compute self-sensitivities by the leaving out one experiment.

The definitions of influence matrix (2.4) and self-sensitivity (2.5) 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 (2.10) 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 we have 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

3.1 Linear statistical estimation in Numerical Weather Prediction

Data assimilation systems for NWP provide estimates of the atmospheric state $\mathbf{x}$ by combining meteorological observations $\mathbf{y}$ with prior (or background) information $\mathbf{x}_b$. A simple Bayesian Normal model provides the solution as the posterior expectation for $\mathbf{x}$, given $\mathbf{y}$ 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 $\mathbf{x}$, given $\mathbf{y}$ and $\mathbf{x}_b$. The optimal GLS solution to the analysis problem (see Lorenc, 1986) can be written

$$\mathbf{x}_a = \mathbf{K}\mathbf{y} + (\mathbf{I}_n - \mathbf{K}\mathbf{H})\mathbf{x}_b \quad 3.1$$

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

$$\mathbf{K} = (\mathbf{B}^{-1} + \mathbf{H}^T\mathbf{R}^{-1}\mathbf{H})^{-1}\mathbf{H}^T\mathbf{R}^{-1} \quad 3.2$$

Here, $\mathbf{H}$ is a $p \times n$ 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, $\mathbf{H}$ is defined to include also the propagation in time of the atmospheric state vector to the observation times using a forecast model.

Substituting (3.2) into (3.1) and projecting the analysis estimate onto the observation space, the estimate becomes

$$\hat{\mathbf{y}} = \mathbf{H}\mathbf{x}_a = \mathbf{H}\mathbf{K}\mathbf{y} + (\mathbf{I}_p - \mathbf{H}\mathbf{K})\mathbf{H}\mathbf{x}_b \quad 3.3$$

It can be seen that the analysis state in observation space ($\mathbf{H}\mathbf{x}_\mathbf{a}$) is defined as a sum of the background (in observation space, $\mathbf{H}\mathbf{x}_\mathbf{b}$) and the observations $\mathbf{y}$, weighted by the $p \times p$ square matrices $\mathbf{I} - \mathbf{H}\mathbf{K}$ and $\mathbf{H}\mathbf{K}$, respectively.

Equation (3.3) is the analogue of (2.1), except for the last term on the right hand side. In this case, for each unknown component of $\mathbf{H}\boldsymbol{x}$, there are two data values: a real and a 'pseudo' observation. The additional term in (3.3) includes these pseudo-observations, representing prior knowledge provided by the observation-space background $\mathbf{H}\mathbf{x}_\mathbf{b}$. From (3.3) and (2.4), the analysis sensitivity with respect to the observations is obtained

$$\mathbf{S} = \frac{\partial \hat{\mathbf{y}}}{\partial \mathbf{y}} = \mathbf{K}^T\mathbf{H}^T \quad 3.4$$

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

$$\frac{\partial \hat{\mathbf{y}}}{\partial (\mathbf{H}\mathbf{x}_\mathbf{b})} = \mathbf{I} - \mathbf{K}^T\mathbf{H}^T = \mathbf{I}_\mathrm{p} - \mathbf{S} \quad 3.5$$

We focus here on the expressions (3.4) and (3.5). 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 $1\text{-}S_{ii}$. It also follows that the complementary trace, $\mathrm{tr}(\mathbf{I}-\mathbf{S})=p-\mathrm{tr}(\mathbf{S})$, 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 $\mathrm{tr}(\mathbf{S})$. 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, (2.6) is assured, but for more general non-diagonal **R**-matrices it is easy to find counter-examples to that property.

Inserting (3.2) into (3.4), we obtain

$$\mathbf{S} = \mathbf{R}^{-1}\mathbf{H}(\mathbf{B}^{-1} + \mathbf{H}^T\mathbf{R}^{-1}\mathbf{H})^{-1}\mathbf{H}^T \qquad 3.6$$

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

3.2 An idealized case, for illustration

Assume there are two observations, each coincident with a point of the background - that is $\mathbf{H}=\mathbf{I}_2$. 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 $R=\sigma_o^2\begin{pmatrix}1 & 0\\ 0 & 1\end{pmatrix}$. For this simple case S is obtained with

$$S_{11} = S_{22} = \frac{r+1-\alpha^2}{r^2+2r+1-\alpha^2} \qquad 3.7$$

$$S_{12} = S_{21} = \frac{\alpha r}{r^2+2r+1-\alpha^2} \qquad 3.8$$

where $r = \sigma_o^2 / \sigma_b^2$. We can see that if the observations are very close (compared to the scale-length of the background error correlation), i.e $\alpha \sim 1$, then

$$S_{11} = S_{22} = S_{12} = S_{21} \simeq \frac{1}{r+2} \qquad 3.9$$

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 ($\sigma_b>>\sigma_o$), that is r~0, then both observations have influence $S_{11}=S_{22}=1/2$, and the background sensitivities are $1-S_{11}=1-S_{22}=1/2$.

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

$$S_{11} = S_{22} = \frac{2-\alpha^2}{4-\alpha^2} \qquad 3.10$$

$$S_{12} = S_{21} = \frac{\alpha}{4-\alpha^2} \qquad 3.11$$

If the locations are far apart, such that $\alpha \sim 0$, we obtain $S_{11}=S_{22}=1/2$, the background sensitivity is also ½ and $S_{12}=S_{21}=0$. We can conclude 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*

$$\hat{y}_1 = \frac{1}{4-\alpha^2}\left[(2-\alpha^2)y_1 + 2x_1 - \alpha(x_2 - y_2)\right] \qquad 3.12$$

The estimate $\hat{y}_1$ depends on y_1, x_1 and an additional term due to the second observation. We see 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 [$\alpha \rightarrow \pm 1$, $(2-\alpha^2) \rightarrow 1$]. From the expression above we also see that the contribution from the second observation is increasing with the correlation's absolute value, implying a larger contribution due to the background x_2 and observation y_2 nearby observation y_1.

3.3 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, $\mathbf{J''}$ 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 χ, $\boldsymbol{\chi}=\mathbf{L}^{-1}(\mathbf{x}-\mathbf{x}_b)$, with **L** chosen such that $\mathbf{B}^{-1} = \mathbf{L}^T\mathbf{L}$. 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).

$$\mathbf{J''}^{-1} \simeq \mathbf{B} - \sum_{i=1}^{M} \frac{1-\lambda_i}{\lambda_i}(\mathbf{L}v_i)(\mathbf{L}v_i)^T \qquad 3.13$$

The summation in (3.13) 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**. In ECMWF's operational data assimilation system, the variances of analysis error are computed according to this method. The variances are inflated to provide estimates of short-term forecast (background) error variances to be used as background errors in the next analysis cycle (Fisher, 1996). 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 coniugate-gradient algorithm. Because the minimum is found within an iterative method, the operational number of iterations is sufficient to find the solution (with 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 (3.13) 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

$$\mathbf{B} = \frac{1}{N}\sum_{i=1}^{N}(\mathbf{L}u_i)(\mathbf{L}u_i)^T \qquad 3.14$$

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 3.13 and 3.14 into 3.6 we arrive at an approximate method for calculating **S**, that is practical for a large dimension variational assimilation (both 3D and 4D-Var):

$$\mathbf{S} = \mathbf{R}^{-1}\mathbf{H}[\frac{1}{N}\sum_{i=1}^{N}(\mathbf{L}u_i)(\mathbf{L}u_i)^T + \sum_{i=1}^{M}\frac{1-\lambda_i}{\lambda_i}(\mathbf{L}v_i)(\mathbf{L}v_i)^T]\mathbf{H}^T \qquad 3.15$$

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 $i \neq j$, that represents the influence of the jth observation to the analysis at the ith 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.(3.15) gives $\mathbf{S}=\mathbf{R}^{-1}\mathbf{HBH}^T$ which in particular can have diagonal elements larger than one if elements of $\mathbf{HBH}^T$ are larger than the corresponding elements of **R**. We investigate the effects of the approximations on the calculated S_{ii} values in section 4.4.

Note that Eq.(2.9) can be applied to quantify how much the analysis at a given observation location would change by deleting the observation itself. The change depends only on the self-sensitivity and the residual value at that location. Note also that the tr(**S**) provides estimates of the information content of the data with respect to the background, and is equal to the degrees of freedom for signal as studied by e.g. Purser and Huang, 1993, Rabier *et al.* (2002) and Fourrié and Thépaut (2003) in the context to remote-sensing retrieval applications. In particular, Rabier *et al.* (2002) use a "data resolution matrix", which is basically the same as the influence matrix. Recently Fisher (2003) computed an estimate of the global tr(**S**) by using the Bay *et al.*(1996) method, without explicitly computing the individual elements S_{ii}. Comparison with Fisher's (2003) estimate has provided validation of our method, in terms of the global trace.

4 Results

The diagonal elements of the influence matrix have been computed for the operational 4D-Var assimilation system. The calculations in Eq(3.15) have been carried out on 60 model levels at T95 spectral truncation n=2,802,912, which is the resolution used for estimation of analysis and background error variances in the operational system. The observation departures ($\mathbf{y}$-$\mathbf{Hx}_b$) were calculated by comparing the observations with a 12-hour forecast integration at T319 resolution. The experiment date is the 13th February 2003 at 12 UTC using observations between 03 and 15 UTC. The number of Hessian vectors used here is M=753 and the number of random **B** vectors is N=500 (M=40 and N=50 in operations). The counts of assimilated observations for each main observation type are given in Table.1. The global total number of observations in this case was p=1,489,777. A large proportion of the used data is provided by satellite systems (Thépaut and

Andersson 2003): QuikSCAT near-surface winds, SATOB cloud-drift winds, AMSU-A microwave radiances, HIRS infrared radiances, SSMI microwave imager, GOES and METEOSAT water-vapour radiances, and ozone data. The remainder are surface-based observing systems (see WMO 1996).

Type of Data	Description	Number
SYNOP	Surface Observations from land and ship stations: measuring p_s, T, RH , u and v	36 889
QuikSCAT	Satellite microwave scatterometer: derived measurement is u and v at the ocean surface	114 220
DRIBU	Drifting buoy measuring p_s, T, RH, u and v	3 419
SATOB	Satellite cloud drift winds	102 090
TEMP	Radiosondes from land and ship measuring p_s, T, RH , u and v	61 205
AMSUA	Satellite microwave sounder radiances	641 274
PILOT	Sondes and Wind profiler measuring u and v	45 060
HIRS	Satellite infrared radiances	163 134
AIREP	Aircraft measurements of T, u and v	119545
SSMI	Satellite microwave imager radiances	116 739
PAOB	Surface pressure Bogus observations	219
GOES	Geostationary satellite infrared sounder radiances	35 302
OZONE	Satellite ozone retrieval	11 058
METEOSAT	Geostationary satellite infrared sounder radiances	39 623
Total		1 489 777

Table 1 Data counts by observation type, 20030213-12 UTC. The total number of data is p=1,489,777.

4.1 Self-sensitivity examples

Self-sensitivities for SYNOP surface pressure observations are shown in Fig.1. Each box indicates the observation influence at the observation location. Data points with influence greater than one will be investigated in section 4.4. Low-influence data points have large background influence (see 3.4 and 3.5), which is the case in data-rich areas such as North America and Europe (observation influence ~ 0.2). In data-sparse areas individual observations have larger influence: in the Polar regions, where there are only few isolated observations, $S_{ii} \sim 1$ and the background has small influence on the analysis.

In dynamically active areas (Fig.1: North Atlantic), several fairly isolated observations have large influence on the analysis. This is due to the evolution of the background-error covariance matrix as propagated by the forecast model in 4D-Var (Thépaut *et al.* 1993, 1996). As a result, the data assimilation scheme can fit these observations more closely.

Similar features can be seen in Fig.2 showing the influence of u-component wind observations for aircraft (AIREP) data between 300 and 200 hPa. Isolated flight tracks over Atlantic and Pacific oceans show larger influences than measurements over data-dense areas over America and Europe.

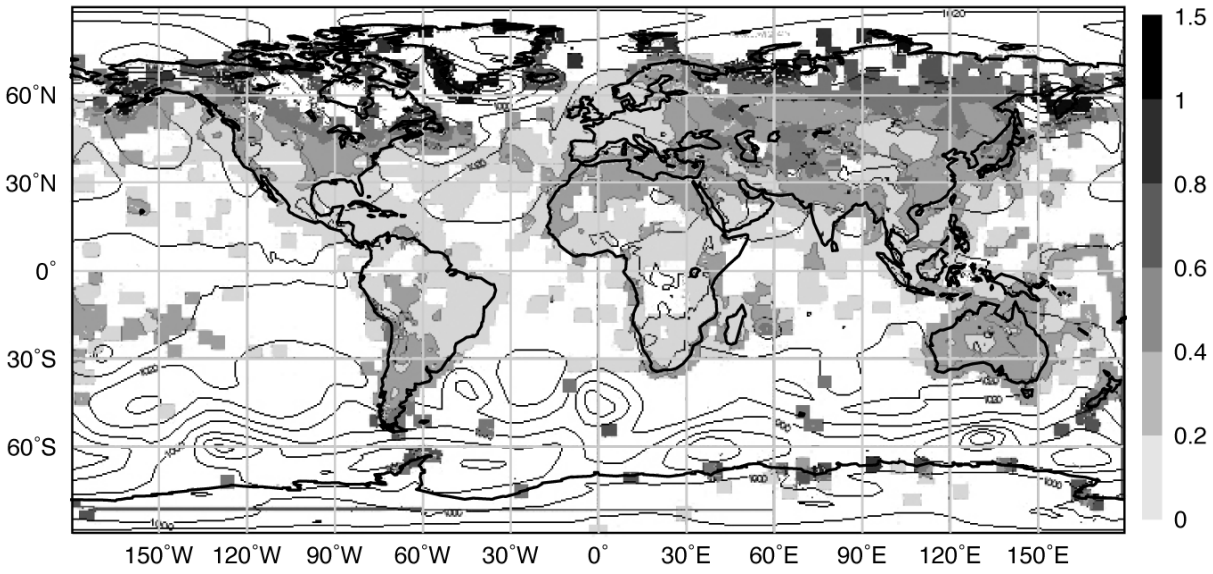

Figure 1 Observation Influence of surface pressure SYNOP data valid for 13 February 200312 UTC

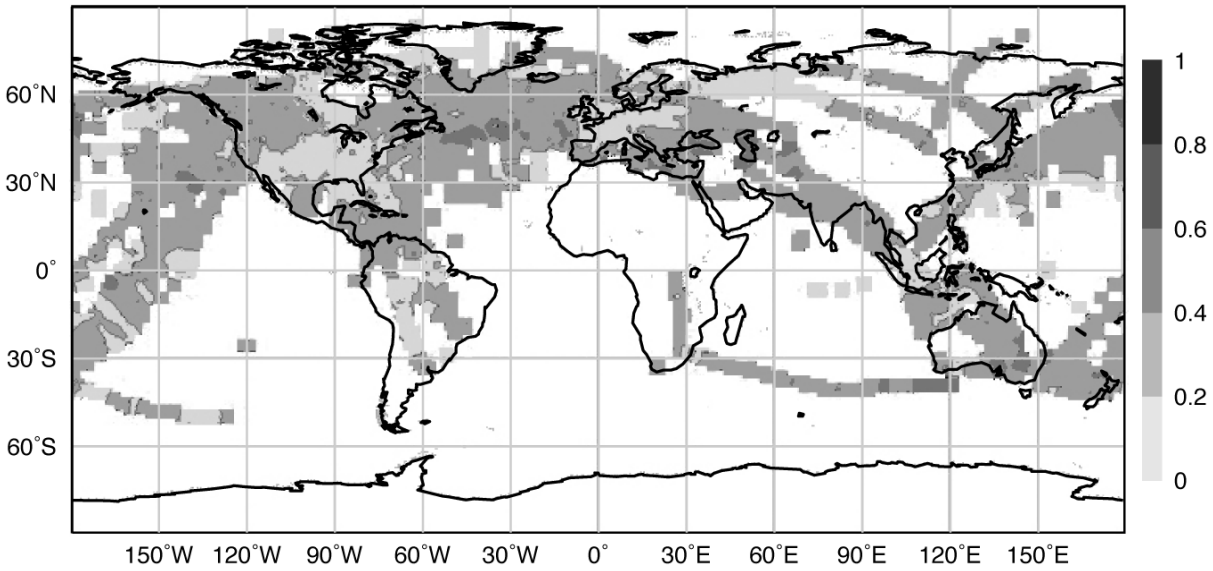

Figure 2 Observation Influence of wind u-component from AIREP data between 200 and 300 hPa. Analysis experiment valid on 13 February 2003 12 UTC.

4.2 Trace diagnostic

We define the Global Average Influence (*GAI*) as the globally averaged observation influence. It is given by

$$GAI = \frac{tr(\mathbf{S})}{p} \qquad 4.1$$

where p is the total number of observations. In our experiment we found that *GI*=0.15. Consequently, the average background global influence to the analysis at observation points is equal to 0.85 (see 3.5).

Another index of interest is the Partial Influence (*PAI*) for any selected subset of data

$$PAI = \frac{\sum_{i \in I} S_{ii}}{p_I} \qquad 4.2$$

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, a particular meteorological variable, for example. In Fig.3 the *PAI* for the different observation types is plotted for three different geographical areas: the Northern Hemisphere extra tropics (*PAI*=15%), the Tropics (*PAI*=17.5%) and the Southern Hemisphere extra tropics (*PAI*=12%). It is clear that METEOSAT and GOES radiances have a relatively large average influence (*PAI*=35%) in the Tropics and in the Southern Hemisphere. In the Northern Hemisphere, DRIBU (drifting buoy) observations have *PAI*=40%, mainly from surface pressure measurements. On the other hand, used ozone observations have very small average influence on the analysis. Also AMSU-A radiance data have a low average influence *PAI*<10%. Similar diagnostics can be obtained for any other data subset of interest. The slightly higher influence in the Northern Hemisphere with respect to the Southern Hemisphere, is explained by the fact that the observations influence is higher for isolated observations and in very dynamically active areas mainly in data-sparse regions. The last one is due to the inflation of background error variances implied by 4D-Var.

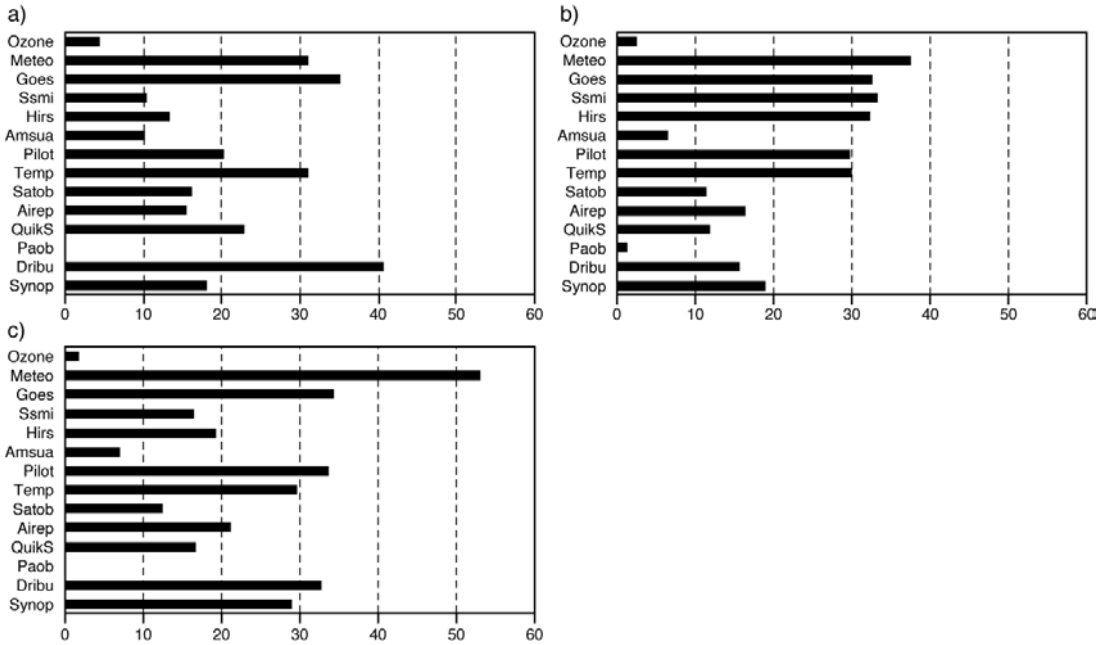

Figure 3 Average Self-sensitivity (AS, in %) for each of the main observation types, for the Northern Hemisphere extra-tropics (a) PAI=15%, the Tropics (b) PAI=17.5% and the Southern Hemisphere extra-tropics (c) PAI=12%.

As we found relatively high influence from satellite measurements that have strong dependence on atmospheric humidity (i.e. GOES, METEOSAT, SSMI and HIRS), we show in Fig.4 a scatter plot for all types of humidity-related observations (every type is a subset *I*). In the top panel, normalized sub-traces of tr(**S**) are plotted versus p_I/p for the tropical area. The sub-traces represent the influence of relative humidity (*rh*) from SYNOP and DRIBU, specific humidity (*q*) from TEMP and radiances from METEOSAT (2 satellites), GOES (1 satellite), SSMI (3 satellites, 7 channels) and HIRS (2 satellites, 8 channels). The sum of

all subsets I defines the subset F. On the x-axis, the relative number of observations per sub-trace is normalized with respect to the total observation number (frequency=p_I/p).

In the mid panel, the residuals ($\sum_{i\in I} S_{ii} / Trace(S) - p_I / p$) with respect to the bisect regression line are plotted versus p_I/p (as before). The mid plot highlights humidity data types with particularly high or low self-sensitivity. We find that the radiances from the HIRS (channels 11 and 12), GOES and METEOSAT satellite data types have a significantly higher (p-value=0) self-sensitivity than all the other assimilated data. Also, these observation types are significantly (p-value=8%) more influential than the other humidity-related observations. In fact, normalizing the residuals with respect to partial amounts ($\sum_{i\in HIRS} S_{ii} / \sum_{i\in F} S_{ii} + \sum_{i\in SSMI} S_{ii} / \sum_{i\in F} S_{ii} + ...$ and for the x-axis $p_{HIRS} / p_F + p_{SSMI} / p_F + ...$), Fig. 4 (bottom panel) shows the high influence of HIRS channel 11, in particular. Further investigations of this and similar diagnostics should be performed to find out if this is indication that aspects of the background error or observation error covariances need to be adjusted, or if what we see is the correct behaviour given the relative accuracies of the background and the observation error for the various data types.

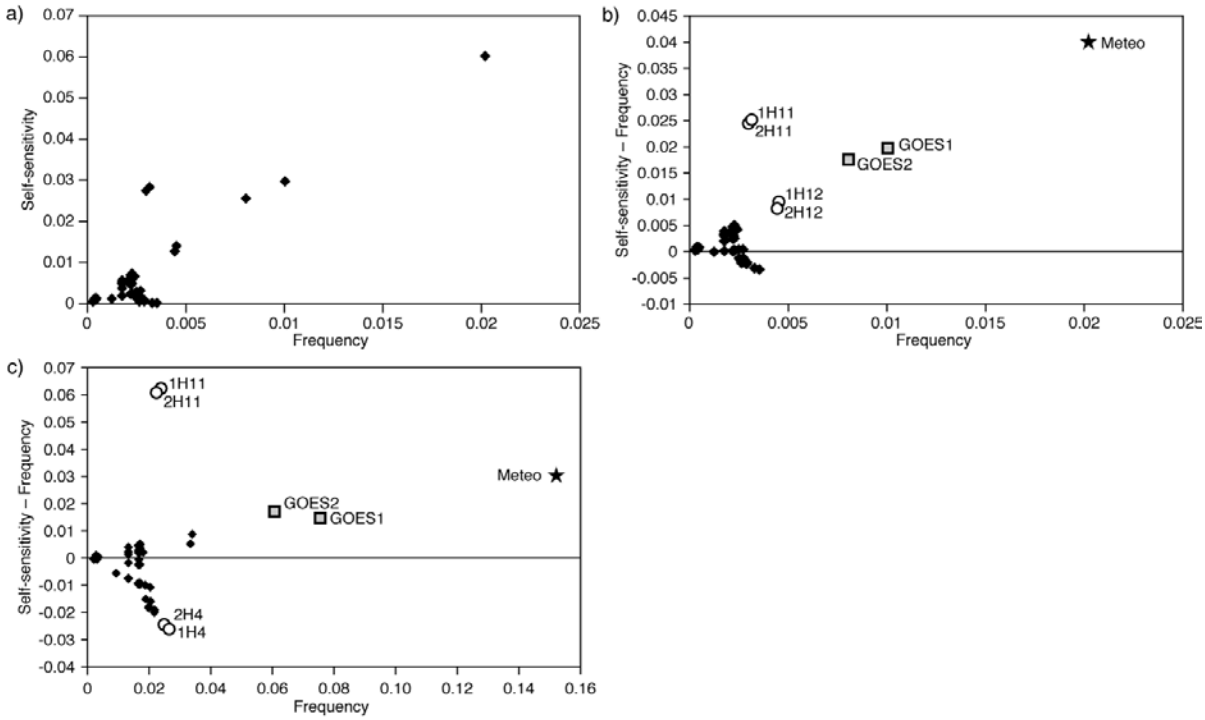

Figure 4 a) normalized sub-traces of Tr(S) are plotted versus pI/p for the tropical area and for relative humidity from SYNOP and DRIBU, specific humidity from TEMP and radiances from METEOSAT, GOES, SSMI and HIRS. Residuals (self-sensitivities minus frequency) with respect to the bisect line plotted versus pI/p are shown in b). c) is similar but values are normalized with respect to partial amount

4.3 Information content

In Section 2 we showed that tr(**S**) can be interpreted as a measure of the amount of information extracted from the observations. In fact, in non-parametric statistics, tr(**S**) measures the 'equivalent number of parameters' or *degrees of freedom for signal*. Having obtained values of all the diagonal elements of **S** (using 3.15) we can now obtain reliable estimates of the information content in any subset of the observational data. In Figure 5 we illustrate this in one example. It must be noted that this theoretical measure of information

content does not necessarily translate to value in terms of forecast impact. The figure shows the information content for all main observation types. We see that AMSU-A radiances are the most informative data type, providing 22% of the total observational information. HIRS follows with 17% and SSMI with 13%. The information content of AIREP, QuikSCAT, TEMP, GOES and Meteosat each is 6-8% while SYNOP, SATOB and PILOT are each less than 5%. A large part of the AMSU-A information is with respect to stratospheric temperature. Most of the SSMI and HIRS information is with respect to humidity. DRIBU and OZONE information content is small: ozone observations have a very small average influence (Fig.3) and dense data coverage while DRIBU observations have large mean influence per observation but much lower data counts (Fig.3). The ozone data are important for the ozone assimilation in spite of their low information content per analysis cycle. Ozone is generally a long-lived species, which allows observational information to be advected accurately by the model over periods of several days. The importance of AMSU-A correlates well with recent data impact studies by (Bouttier and Kelly, 2001; Kelly 2003 personal communication).

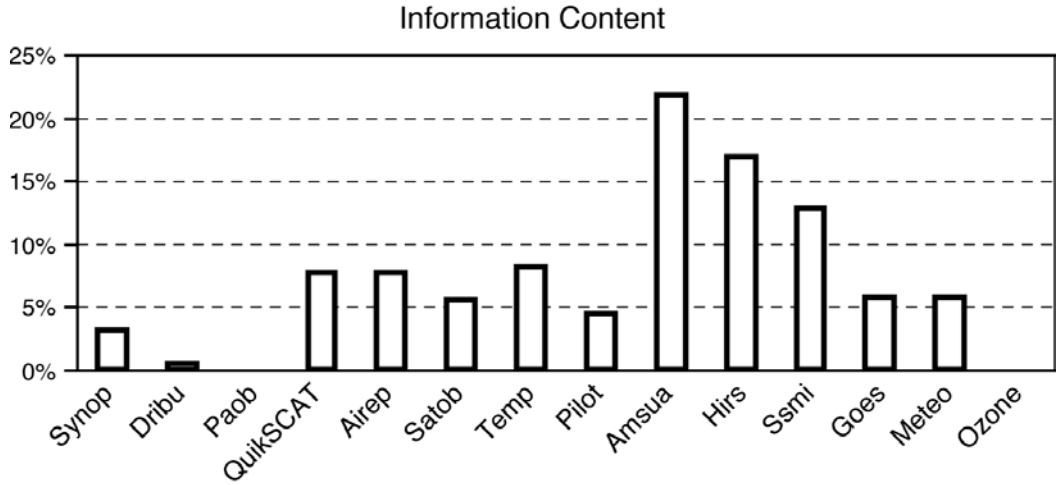

Figure 5 Information content i.e. degree of freedom for signal for the main data types in the assimilation

Our results are in good qualitative agreement with those obtained by Fisher (2003) using *Bay et al.'s* (1996) algorithm. The advantage of the influence-matrix approach is that it allows the computation of information content for any subset of data by straightforwardly summing up the elements corresponding each subset. The methods based on a global trace estimate, however, require that the analysis scheme is re-run once per data sub-division (Fisher 2003).

4.4 High-influence data

From (2.6) it is known that self-sensitivities cannot be larger than one but we have seen from Fig.1 that some S_{ii} are greater than one. In this paragraph the nature of the upper-bound values is explained.

4.4.1. Self-sensitivity calculation approximation

In Fig. 6a the self-sensitivities for HIRS channel-11 are shown. Values greater than one are shown in black. Here only M=40 Hessian eigen-vectors and N=50 random B vectors have been used to compute the self-sensitivities. In this worst case, 67% of the self-sensitivities are greater than one, in contrast with (2.6), which states that the diagonal elements of S are bounded to the interval between zero and one. The occurrence of $S_{ii}>1$ is partly due to the approximate nature of our method for calculation of self-sensitivity. Approximations in both of the two terms of (3.15) contribute to the problem. In the second term the number

of Hessian eigen-vectors is truncated to M. The term is therefore underestimated, and S_{ii} will tend to be over-estimated. 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 humidity analysis with its short co-variance length scales ~180 km, and large observational impacts) a large M is required to approximate B-A accurately. We repeated the calculations for several values of M>40, to the upper limit M=735 that could reasonably be afforded given the available computer resources. In Fig.6b we show the result corresponding to Fig.6a but for M=753 Hessian vectors. The number of S_{ii}>1 for this HIRS channel is now reduced by 30%. The approximate computation is mostly affecting the self-sensitivities close to the upper bound leaving the self-sensitivities <0.7 almost unaffected.

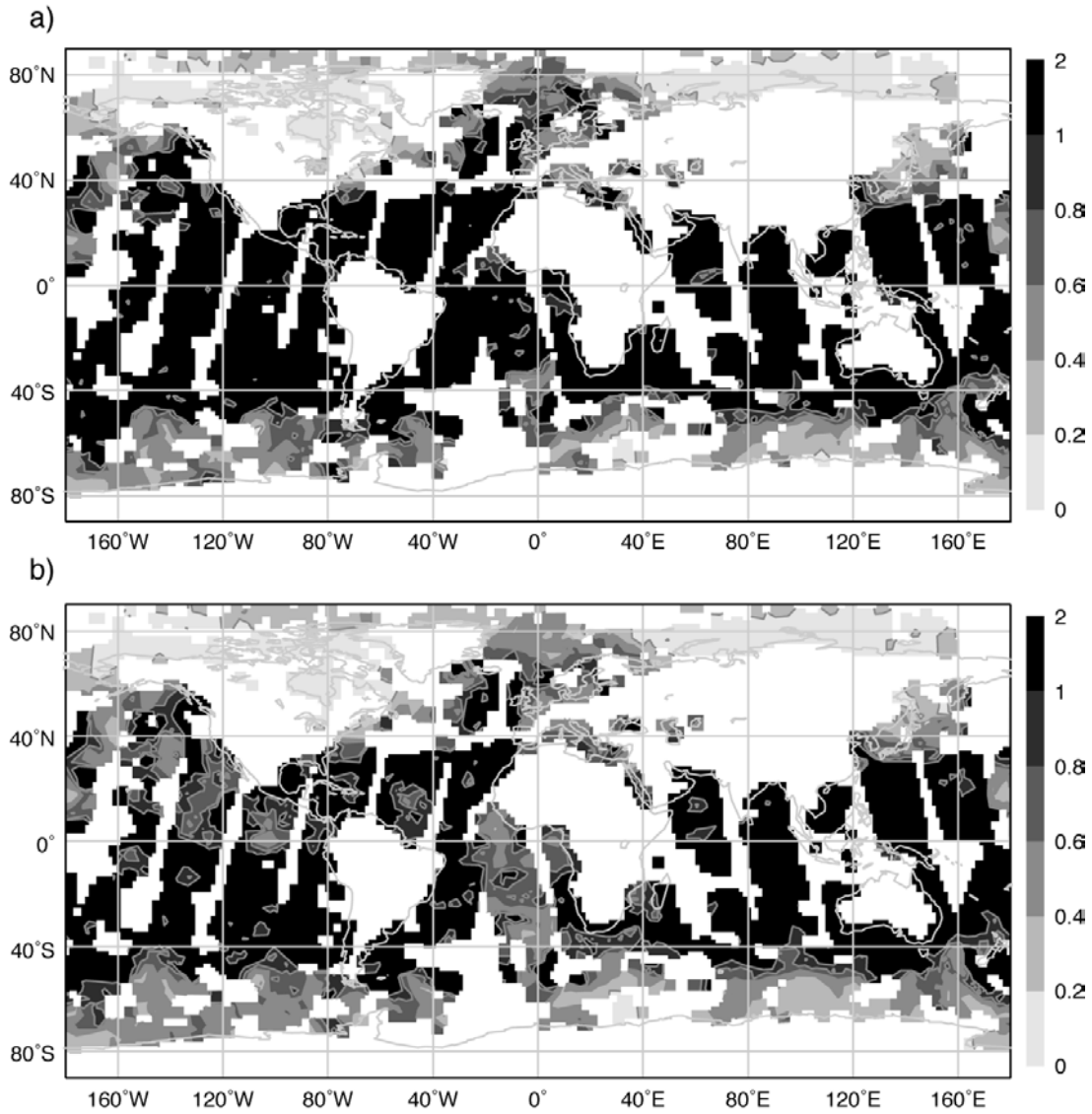

Figure 6 Radiances Influence from HIRS channel 11. a) M=40 Hessian eigen-vectors used. b) M=753 Hessian eigen-vectors used

HIRS channel 11 is the data set for which the problem is most acute. In fact, even for M=40, self-sensitivities greater than one appear only for a very small proportion (2%) of the global set of observations. In Fig.7, 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 (3.15). 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.

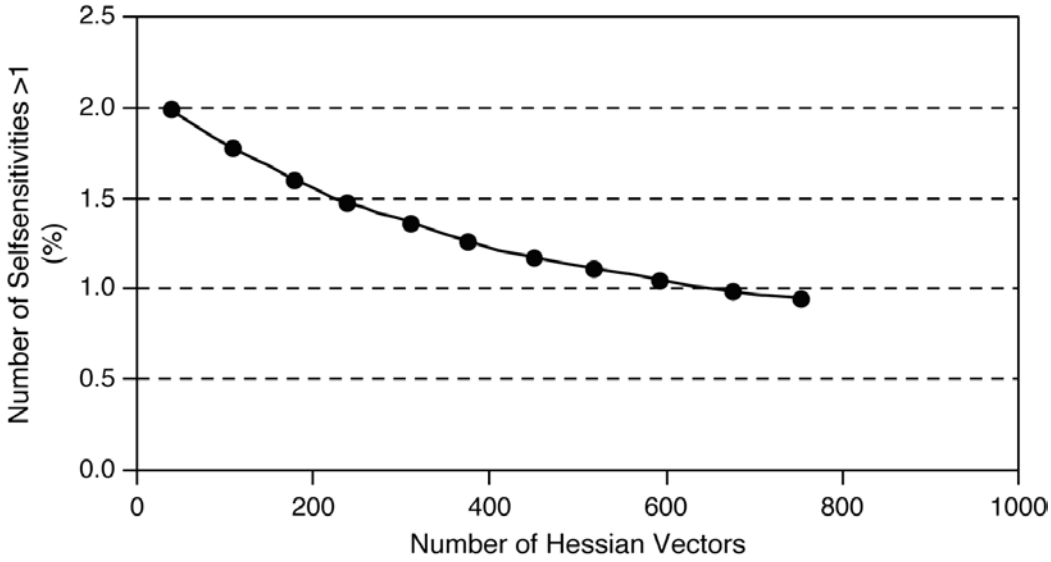

Figure 7 Proportion of self-sensitivity values greater than one (in %) versus the number of Hessian vectors used to compute the variances of B-A, using the approximate method described in Eq.(3.15)

4.4.2. Ill-conditioning problem

A set of linear equations is said to be *ill-conditioned* if small variations have a large effect on the exact solution $\hat{\mathbf{y}}$. For example, if $\mathbf{X}$ (or its analogue in 4D-Var) is nearly singular small changes in its element can cause large variations in data assimilation, serious effect on the stability and accuracy of the solution. A measure of ill-conditioning is the condition number $\mathcal{K}$ defined as the ratio between the largest and the smallest nonzero singular value of $\mathbf{X}$, that is, the square roots of the eigen-value of $\mathbf{XX}^{\mathbf{T}}$. Regression models are notoriously ill-conditioned when the degree is higher than five.

It is not uncommon that variates (regressors) can be in linear relationship with a set of variates (collinearity). This means that $\mathbf{X}^{\mathrm{T}}\mathbf{X}$ will be near singularity, the smallest eigen-value will be small and $\mathcal{K}$ will be large. In 4D-Var $\mathcal{K}$ is typically 10^4-10^5. Cases of poor convergence of the 4D-Var solution algorithm have been studied (Andersson *et al.* 2000). It was found that when the ratio σ_b/σ_o between specified background and observation errors is large, or the observation density is large, the 4D-var problem can be ill-conditioned and poor convergence in the minimization can result.

We have found that self-sensitivities exceeding the upper bound ($S_{ii}>1$) can provide indication of data that contribute to poor conditioning of the analysis, because of large σ_b/σ_o ratio. To illustrate this, self-sensitivities computed in a single analysis experiment have been compared to self-sensitivities computed

during a modified four-day assimilation cycle. In Fig. 8, self-sensitivities for DRIBU surface pressure observation over the North Atlantic (Fig. 8a) are shown together with the mean sea level pressure field. Higher self-sensitivities (>0.6) are close to a dynamically active area (a developing low). Figure 8b shows that the ratio σ_b/σ_o for the same observations is close to one. The patterns in both Fig. 8a and b are dependent on the background error variance specification, which in ECMWF practice is modified to account for error growth in the prediction over the time interval between analyses (12 hours). Operationally (and in Fig. 8a and b), the B matrix variances are inflated exponentially in time (Fisher 1996), and the new variances are used in the next analysis cycle. In Fig. 8c and d the error inflation has instead been performed in a flow dependent manner using the tangent linear forecast model (Andersson and Fisher 1999). The influence pattern for DRIBU surface pressure is now computed after having cycled the 4D-Var system for four days (the experiment started on 9 February 2003 12 UTC). The results shown in panels 8c and d are valid at the same time of Fig. 8a (13 February 2003 12 UTC). In the region of the low, self-sensitivities are now larger and some are exceeding one (Fig. 8c). Also the ratio σ_b/σ_o (Fig.8d) has increased, in some locations by a factor three. Variance inflation and propagation (applied to the background standard deviation) can thus generate collinearity and potentially ill-condition the estimation problem. We know that data points with self-sensitivities greater than one are due to the approximate nature of our computation, but when they occur they can nevertheless be valuable by drawing attention to those data points that may be contributing towards ill-conditioning of 4D-Var.

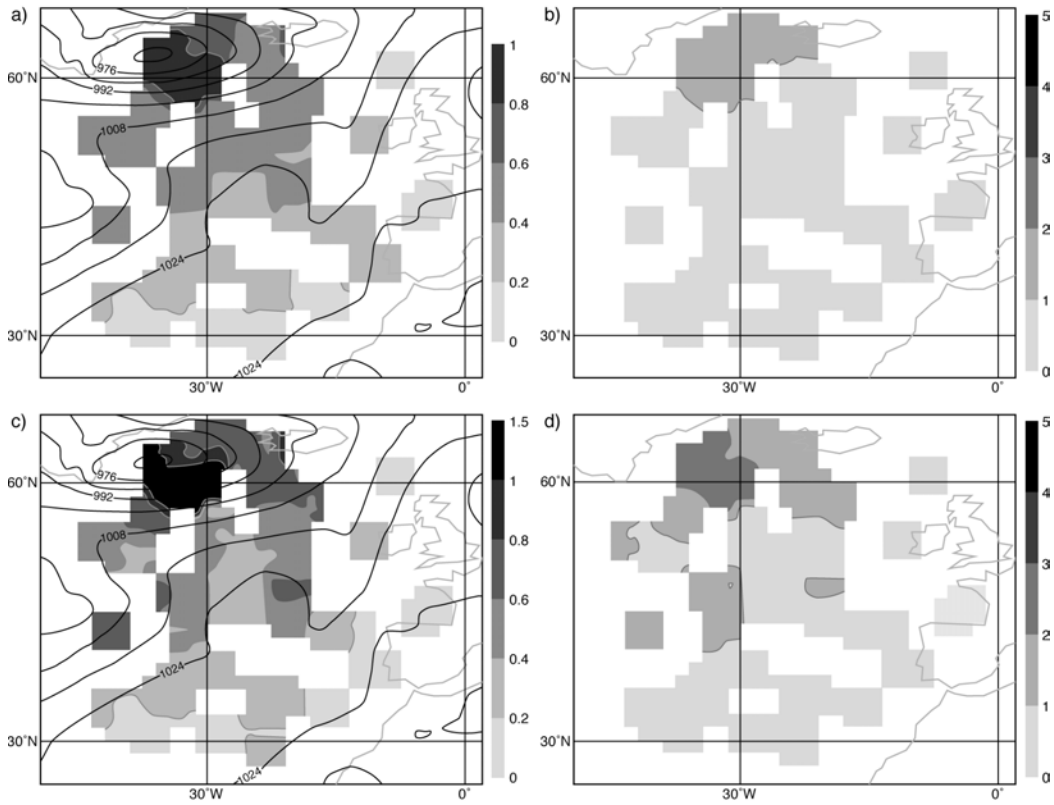

Figure 8 a) DRIBU surface pressure influence valid on 13 February 200312 UTC. b) Ratio between background and observation standard deviation for same experiment shown in a). c) DRIBU surface pressure influence from a flow-dependent background standard deviation experiment cycled for 4 days and valid on 13 February 200312 UTC. d) as b) bur for experiment shown in c).

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 we have derived the influence matrix in the context of linear statistical analysis schemes, as used for data assimilation of meteorological observations in numerical weather prediction (Lorenc 1986). In particular we derive and implement an approximate method to compute the diagonal elements of the influence matrix (the self-sensitivities) in ECMWF's operational data assimilation system (4D-Var). Our approach is necessarily approximate due to the large dimension of the estimation problem at hand: the number of estimated parameters is of the order 10^6 ($8*10^6$ in operational practice, and $3*10^6$ in the calculations presented here), and the number of observational data is around $1.5*10^6$.

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. We provided an illustration with respect to humidity-related observations in 4D-Var showing particularly large influence of HIRS channel 11.

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. We showed examples for surface pressure and aircraft wind observations 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 3.2) in favour of background influence and amplify the influence of the surrounding observations.

In this study the global observation influence per assimilation cycle has been found to be 15%, and consequently the background influence is 85%. 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 (22%), followed by HIRS (17%), SSMI (13%), AIREP, QuikSCAT, TEMP, Goes and Meteosat each providing 6-8 %. It must be stressed that this ranking is not an indication of relative importance of the observing systems for forecast accuracy. The results compare well with independent estimates provided by Fisher (2003).

If the influence matrix were computed without approximation then all the self-sensitivities will be bounded in the interval zero to one. With the approximate method used here 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 in this study, and in particular those that are closer to one. Remaining values greater than one can be due to large background to observation error ratio, which is one factor that is known to contribute towards ill-conditioning and poor convergence of the 4D-Var algorithm.

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 (however this aspect has not been explored within the current study). In fact, the leaving-out-one observation, that is not practical for large system dimension, uses the discarded observation to assess the quality of the analysis. It has been shown (Eq. 2.9) that Self-sensitivities 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. It would be interesting to see if a small sub group of very influential data (i.e. satellite observations) have the same impact in the forecast than the full amount of data. If this is the case, a dynamical thinning can be thought that selects, every assimilation cycle, the most influent partition of a particular remote sensing instrument measurements, from information based on the previous cycle. Clearly, it can be assumed that components of the observing network remains constant and background error variances remained almost unchanged for close assimilation cycles.

Acknowledgements. The authors thank Mike Fisher for valuable discussion and suggestions. Olivier Talagrand provided helpful comments to the manuscript. Thank to Jan Haseler and Sami Saarinen for the technical support.

Appendix

Influence Matrix Calculation in Weighted Regression Data Assimilation Scheme

Under the *frequentist* approach, the regression equations for observation

$$\mathbf{y} = \mathbf{H}\boldsymbol{\theta} + \boldsymbol{\varepsilon}_o$$

and for background

$$\mathbf{x}_b = \boldsymbol{\theta} + \boldsymbol{\varepsilon}_b$$

are assumed to have uncorrelated error vectors $\boldsymbol{\varepsilon}_o$ and $\boldsymbol{\varepsilon}_b$, zero vector means and variance matrices **R** and **B**, respectively. The $\boldsymbol{\theta}$ parameter is the unknown system state (**x**) of dimension n. These regression equations are summarized as a weighted regression

$$\mathbf{z} = \mathbf{X}\boldsymbol{\theta} + \boldsymbol{\varepsilon}$$

where $\mathbf{z} = \left[\mathbf{y}^T\ \mathbf{x}_b{}^T\right]^T$ is *(p+n)×1* ; $\mathbf{X} = \left[\mathbf{H}^T\ \mathbf{I}_n\right]^T$ is *(p+n)×n* and $\boldsymbol{\varepsilon} = \left[\boldsymbol{\varepsilon}_o\ \boldsymbol{\varepsilon}_b\right]^T$ is *(m+n)×1* with zero mean and variances matrix

$$\boldsymbol{\Omega} = \begin{pmatrix} \mathbf{R} & 0 \\ 0 & \mathbf{B} \end{pmatrix}$$

The generalized LS solution for $\boldsymbol{\theta}$ is BLUE and is given by

$$\hat{\boldsymbol{\theta}} = (\mathbf{X}^T\boldsymbol{\Omega}^{-1}\mathbf{X})^{-1}\mathbf{X}^T\boldsymbol{\Omega}^{-1}\mathbf{z} \qquad \text{A1.1}$$

see Talagrand (1997). After some algebra this equation equals (3.1). Thus

$$\mathbf{z} = \mathbf{X}\hat{\boldsymbol{\theta}} = \left[\mathbf{H}^T\mathbf{x}_a{}^T\ \mathbf{x}_a{}^T\right]^T = \mathbf{X}(\mathbf{X}^T\boldsymbol{\Omega}^{-1}\mathbf{X})^{-1}\mathbf{X}^T\boldsymbol{\Omega}^{-1}\mathbf{z}$$

and by (2.5) the influence matrix becomes

$$\mathbf{S}_{zz} = \frac{\partial\hat{\mathbf{z}}}{\partial\mathbf{z}} = \frac{\partial\mathbf{X}\hat{\boldsymbol{\theta}}}{\partial\mathbf{z}} = \begin{pmatrix} \mathbf{S}_{yy} & \mathbf{S}_{yb} \\ \mathbf{S}_{by} & \mathbf{S}_{bb} \end{pmatrix} = \begin{pmatrix} \mathbf{R}^{-1}\mathbf{H}\mathbf{A}\mathbf{H}^T & \mathbf{R}^{-1}\mathbf{H}\mathbf{A} \\ \mathbf{B}^{-1}\mathbf{A}\mathbf{H}^T & \mathbf{B}^{-1}\mathbf{A} \end{pmatrix}$$

where $\mathbf{S}_{yy} = \frac{\partial\mathbf{H}\mathbf{x}_a}{\partial\mathbf{y}}; \mathbf{S}_{yb} = \frac{\partial\mathbf{x}_a}{\partial\mathbf{y}}; \mathbf{S}_{by} = \frac{\partial\mathbf{H}\mathbf{x}_a}{\partial\mathbf{x}_b}; \mathbf{S}_{bb} = \frac{\partial\mathbf{x}_a}{\partial\mathbf{x}_b}$. Note that $\mathbf{S}_{yy}$=**S** as defined in (3.4).

Generalized LS regression is different from ordinary LS because the influence matrix is not symmetric anymore. For idempotence, using (A1.1) it easy to show that $\mathbf{S}_{zz}\mathbf{S}_{zz}=\mathbf{S}_{zz}$. Finally,

$$\mathbf{S}_{bb} = \mathbf{B}^{-1}\mathbf{A} = \mathbf{I}_n - \mathbf{H}^T\mathbf{R}^{-1}\mathbf{H}\mathbf{A}$$

hence,

$$tr(\mathbf{S}_{bb}) = n - tr(\mathbf{H}^T\mathbf{R}^{-1}\mathbf{H}\mathbf{A}) = n - tr(\mathbf{S}_{yy})$$

it follows that

$$tr(\mathbf{S}_{zz}) = tr(\mathbf{S}_{yy}) + tr(\mathbf{S}_{bb}) = n$$

The trace of the influence matrix is still equal to the parameter's dimension.

References

Andersson, E. and Fisher, M. 1999: Background errors for observed quantities and their propagation in time. Proc. ECMWF Workshop on "Diagnosis of Data Assimilation Systems", Reading, U.K., 1-4 Nov. 1998, 81—90.

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.

Bay, Z., Fahey, M. and Golub, G. H.1996: Some large scale matrix computation problems. *J. Comput. Appl. Math.*, **74**, 21—89.

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

Bouttier, F, and Kelly, G., 2001: Observing system experiments in the ECMWF 4D-Var data assimilation system. *Q. J. R. Meteorol. Soc*, 127,1469-1488

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.

Fisher, M., 1996: The specification of background error variances in the ECMWF variational analysis system. Proc. ECMWF workshop on "Non-linear aspects of data assimilation", Reading, 9-11 September 1996, 645—652.

Fisher, M., 2003: Estimation of entropy reduction and degrees of freedom for signal for large variational analysis systems. *ECMWF Tech. Memo.*, **397,** pp 18.

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.

Fisher, M. and Andersson, E., 2001: Developments in 4D-Var and Kalman Filtering. *ECMWF Tech Memo.*, **347,** pp 36.

Fourrié, N. and Thépaut, J-N., 2003: Evaluation of the AIRS near-real-time channel selection for application to Numerical Weather Prediction. *Q. J. R. Meteorol. Soc*, **129**, 2425-2439.

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.

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

Purser, R. J. and Huang, H.-L.,1993: Estimating Effective Data Density in a Satellite Retrieval or an Objective Analysis. *J. Appl. Meteorol.*, **32,** 1092—1107.

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

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

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.

Thépaut, J.N. and Andersson, E. 2003: Assimilation of high-resolution satellite data. *ECMWF Newsletter*, **97,** 6—12.

Thépaut, 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.

Thépaut, J.N., Courtier, P., Belaud G. and Lemaître, 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.

WMO, 1996: Guide to meteorological instruments and methods of observation. Sixth Edition. WMO-No.8. Geneva, Switzerland.

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