The University of Reading

Information Content of Spatially Correlated Observation Errors

L.M. Stewart, S.L. Dance and N.K. Nichols

NUMERICAL ANALYSIS REPORT 4/06

Department of Mathematics
The University of Reading
Whiteknights, PO Box 220
Reading
Berkshire RG6 6AX

Department of Mathematics

Abstract

In the algorithmic process of data assimilation, errors are contributed both from the mis-calculation of observations and the mis-estimation of the background state. The specification of these errors is important as they respectively determine the extent to which the observations and background influence the analysis. For ease of computation, observation errors are often assumed to be independent of each other, i.e., uncorrelated; however, in a reality where instrument noise and model error are present this is not the case.

It is suspected that under the assumption of uncorrelated errors, information from observations is not utilised optimally, and so the benefits of an increased amount of satellite data are limited. In this report, we represent the information content of observations though the measures of Shannon Information Content and degrees of freedom of signal, and make comparisons between three different approaches to observation error correlation specification. Information change from using the traditional pre-processing technique of variance enlargement is also considered.

It is shown that, using empirically derived observation error correlations for Atmospheric Motion vectors (AMVs) data, incorrectly assuming uncorrelated observation errors leads to a significant loss of information. As the number of observations increases, the greater the difference in information is between analyses using a correlated and an uncorrelated observation error covariance matrix, and hence the more important it becomes to correctly specify these correlations. Also, enlarging the variance of observation errors is shown to have a detrimental effect on the amount of information obtained from the data, when used as an alternative to correctly modelled correlations.

From our findings, we can see that accurate observation error correlation specification is needed so that the information gathered from observational data increases in line with the improvements in the techniques of obtaining this data. A question for further study is how to implement these correlations in a computational inexpensive manner.

Contents

1 Introduction

In determining an accurate, high-resolution representation of the current state of the atmosphere for use as an initial condition in Numerical Weather Prediction, it is inadequate to solely use observations of atmospheric variables (e.g., temperature, wind speed, humidity and pressure) because of the insufficient quantity available. But combining these observations with knowledge of the behaviour (i.e, evolution in time) and structure (often embodied in a computer model) of the atmosphere provides us with a more consistent representation. For example, if we know the typical structure of an anticyclone, then an analysis can be drawn using only scattered observations. The sequential marrying of such observations and representation is known as data assimilation.

With the recent improvement of remote sensing techniques and methods of data assimilation, a significant amount of the data used in NWP models is now collected by satellites. For example, at the Met Office in May 2005, a global average of 1,627,480 observations from geostationary satellites were assimilated each day. This accounts for 88% of the total number of observations used [14]. The errors associated with these indirect measurements often stem from the same source, i.e, instrument type or a wrongly specified variable relationship.

The correlation properties of these errors are largely unknown and so for ease in many data assimilation algorithms, the correlated error component is set to zero. Several pre-processing techniques can be used on the data to compensate for this loss of accuracy, such as data thinning [13], superobbing [1], and increasing the observation error variance used in the assimilation [4]. However, such processes involve sub-optimisation of the data, and so useful information is often lost.

The aim of this paper is to investigate the quantitative amount of information gained through the inclusion of correlated errors. First some background theory on observation and background errors, and the problem they play a role in, will be given. We will then look at information theory and the various methods used to evaluate the information content of a set of observations. Discussion will be given on some pre-processing techniques, and we will conduct a practical experiment comparing observation information content for different observation error covariance matrices. Conclusions and discussion of future work will then be given.

2 Satellite Data Assimilation and the Inverse Problem

Typically a satellite instrument measures a radiance L and relates it to geophysical parameters through the radiative transfer equation, as described in [16]:

$$\begin{aligned} L(\nu) &= \int_0^\infty B(\nu, T(z)) \frac{d\tau(\nu)}{dz} dz \\ &+ \text{surface emission} \\ &+ \text{surface scattering}, \end{aligned} \tag{1}$$

where ν is the frequency, $B(\nu, T(z))$ is the Planck radiance for temperature T at altitude z, $\tau(\nu)$ is the altitude z to space transmittance, and $\frac{d\tau(\nu)}{dz}$ can be interpreted as weightings over the atmospheric temperature profile. By selecting radiation at different frequencies, a satellite instrument can provide information on a range of geophysical variables.

There are two main approaches to the use of satellite data in data assimilation. In the first approach retrievals are produced before the main assimilation procedure by some optimal estimation adjusting atmospheric profiles to their background counterparts and measured radiances. In the second there is no need to perform the retrieval step separately as it is incorporated into the main analysis by finding the model variables that minimise the cost function, penalising for distance from the analysed state to both the background and observations.

The observations in [2], from which we will analyse data later in the paper, are not of this type, but instead come from satellite winds and radiosonde measurements. However, satellite wind measurements are operationally derived by cloud tracking in the infrared, water vapour, or visible channel, and still use the process of radiance inversion to determine the height level of vectors.

Ideally we would measure any desired geophysical quantity directly by satellite observation. However, due to the constraints imposed by the atmosphere and technology, this is often not possible, and so an inverse problem must be solved to convert between the measured quantity and the desired one. Taking m measurements for n state variables, gives us an n-dimensional state vector x^t, an m-dimensional measurement vector y, and a set of simultaneous equations with variables from x^t and y. Following the model used in [11], this measurement process can be expressed as a forward model which maps the state space to the measurement space,

$$y = Hx^t + \epsilon^o, \tag{2}$$

where H is the forward model and ϵ^o is the measurement noise. Due to the complex nature of the relationship between atmospheric variables, H is often non-linear. However, for ease of analysis, we will assume H is linear and accept some non-linearity error.

In addition we have some prior knowledge of what we expect the state vector to be, described as the background state x^b,

$$x^b = x^t + \epsilon^b, \tag{3}$$

where ϵ^b is the background noise. Ideally this background state will not be deteriorated by poor quality observations.

We assume both the measurement and background noise are unbiased,

$$\mathbb{E}[\epsilon^o] = \mathbb{E}[\epsilon^b] = 0, \tag{4}$$

where $\mathbb{E}$ is the expectation operator, and define the background and observation error covariances matrices respectively:

$$B = \mathbb{E}[\epsilon^b {\epsilon^b}^T] \tag{5}$$

$$R = \mathbb{E}[\epsilon^o {\epsilon^o}^T]. \tag{6}$$

We seek to optimally combine these two sources of knowledge to give the best representation of the present state of the atmosphere.

Assuming Gaussian pdfs and using Bayesian theory, we minimise the cost function, defined in [11],

$$J(x) = \frac{1}{2}[(x - x^b)^T B^{-1}(x - x^b) + (y - Hx)^T R^{-1}(y - Hx)] \tag{7}$$

to get the Best Linear Unbiased Estimate (BLUE) equations:

$$x^a = x^b + A(y - Hx^b) \tag{8}$$

$$A = BH^T(HBH^T + R)^{-1} = (B^{-1} + H^T R^{-1} H)^{-1} H^T R^{-1} \tag{9}$$

where x^a is the analysis vector with covariance $S_a = (H^T R^{-1} H + B^{-1})^{-1}$.

Combining prior information and observations to get a maximum likelihood estimate of x^a is expected to require a different weighting on each one of these sources of knowledge; the A matrix provides these optimal weights.

2.1 Observation Error Covariance Matrix

The extent to which the background and observations influence the analysis is determined by their respective error values. Observation errors are calculated by examining innovations (differences between observations and the background), and come from three uncorrelated sources: instrument noise, forward model error and non-linearity error. However, errors resulting from each individual source are expected to be correlated due to the similarity of their origin.

We can write R in the form:

$$R = D^{1/2} C D^{1/2}, \tag{10}$$

where C is the correlation matrix

$$C = \begin{pmatrix} 1 & \rho_{12} & \dots & \rho_{1n} \\ \rho_{12} & 1 & \dots & \rho_{2n} \\ \vdots & \ddots & \ddots & \vdots \\ \rho_{1n} & \rho_{2n} & \dots & 1 \end{pmatrix},$$

D is the error variance matrix

$$D = \begin{pmatrix} {\sigma^2}_1 & 0 & \dots & 0 \\ 0 & {\sigma^2}_2 & \dots & 0 \\ \vdots & \ddots & \ddots & \vdots \\ 0 & 0 & \dots & {\sigma^2}_n \end{pmatrix},$$

ρ_{ij} is the observation error correlation between observation i and observation j, and σ_i^2 is the observation error variance of observation i.

Error correlations are very important when we have a high resolution model and a low density of observations, or vice-versa, as they specify how the observations will be smoothed. Correct correlation specifications are vital to the accuracy of observation weightings. Positive error correlations reduce the weight given to the average of observations but give more relative importance to differences between observed values. Taking correlations into account could therefore give the analysis of atmospheric quantities which are calculated using the difference of some measurable quantity a larger dependence on the observations.

The major problem with observation error correlations is their complexity; for example, satellite data retrieval processes can create artificial correlations, and interpolation errors are correlated whenever observations are dense compared to the resolution of the model. Characterisation of observation errors is easier for raw radiances because of the fewer processing steps involved, but still in many models, error correlations are taken to be zero and the R matrix diagonal. This is perhaps a reasonable assumption when observations are taken by separate immovable instruments (i.e, surface observation networks) but not for radiosonde or satellite measurements, where the same instrument in used. We will examine the repercussions of making this assumption later in the paper.

2.2 Background Error Covariance Matrix

The background error covariance matrix is produced from estimates of the error variance in the forecast, and if it is badly specified, we will not have an accurate idea of the variance of the final analysis value x^a. Various methods for calculating this matrix are described in [7].

Although in this report we do not concentrate our attention on background error correlations, they have two very important roles in an assimilation system: information spreading/smoothing, and conveyance of the balance properties between model variables [3]. In regions of dense noisy observations, background error correlations are used to smooth the analysis, and in regions of sparse observations, they are used to spread information from observations to surrounding grid points. Smoothing of increments ensures the analysis contains scales that are statistically compatible with the smoothness properties of physical fields.

3 Information Theory

When we ignore observation error correlations and use processes such as data thinning and superobbing to assimilate satellite data, we are neglecting a portion of this data, and so information that could perhaps be utilised is lost. Ideally we would select the optimal subset of observations such that the important information is retained in a numerically cheap way.

The information content of a set of observations is the number of linearly independent pieces of information contained in the set. For an observation to contain useful information it is required that the natural variability of the observation vector is greater than the measurement error. Considering the system given by [15]

$$\tilde{y} = \tilde{H}\tilde{x} + \tilde{\epsilon}, \tag{11}$$

where $\tilde{x}$ is the transformed state vector,

$$\tilde{x} = B^{-1/2}(x - x^b), \tag{12}$$

and $\tilde{y}$ is the transformed measurement vector,

$$\tilde{y} = R^{-1/2}y, \tag{13}$$

this condition reduces to, the singular value of $\tilde{H} = R^{-1/2}HB^{1/2}$ related to the observation being approximately greater than unity.

This is purely a requirement for measurable information; there are several different ways to calculate the information content of an observation set.

3.1 Shannon Information Content

This is a measure of the reduction of entropy (number of distinct internal states). Using pdfs as a measure of knowledge of the system, and working in state space, suppose $p(x)$ is the knowledge before the observation, $p(x|y)$ is the knowledge after, and $S(p)$ is the entropy. Then the Shannon Information Content, SIC, as defined by [15] is

$$SIC = S[p(x)] - S[p(x|y)] \tag{14}$$

where

$$S[p(x)] = -\int p(x)\lg_2[p(x)]dx, \tag{15}$$

$$S[p(x|y)] = -\int p(x|y)\lg_2[p(x|y)]dx. \tag{16}$$

In the linear Gaussian case, which we will be considering, it is algebraically convenient to use natural logs as opposed to $\lg_2$. Such a change purely results in a slight rescaling of the entropy definition by $\ln 2 = 0.69$, but makes equation manipulation considerably easier. Using this approach, we get the equations:

$$S[p(x)] = n\ln(2\pi e)^{1/2} + \frac{1}{2}\ln|B|$$
$$S[p(x|y)] = n\ln(2\pi e)^{1/2} + \frac{1}{2}\ln|S_a|$$

where $|B|$ and $|S_a|$ are the determinants of matrices B and S_a respectively.

$$\begin{aligned} SIC &= \frac{1}{2}\ln\left|S_a^{-1}B\right| \\ &= \frac{1}{2}\ln\left|(H^TR^{-1}H + B^{-1})B\right| \\ &= \frac{1}{2}\ln\left|H^TR^{-1}HB + I\right| \\ &= \frac{1}{2}\ln\left|B^{1/2}H^TR^{-1}HB^{1/2} + I\right| \\ &= \frac{1}{2}\ln\left|\tilde{H}^T\tilde{H} + I\right| \\ &= \frac{1}{2}\sum_{i=1}^{m}\ln(1+\lambda_i^2) \end{aligned} \tag{17}$$

where λ_i are the singular values of $\tilde{H}$.

Physically entropy can be thought of as a ‘measure of the volume of the state space occupied by a pdf which describes the knowledge of the state’, and by taking an observation, the volume of uncertainty is reduced.

3.2 Degrees of Freedom of Signal

Statistically degrees of freedom can be considered as the number of values in a statistic that are free to vary; the number of degrees of freedom in some observation data is a measure of the amount of information from the data that has been utilised. Obviously we seek observation groupings with a high number of degrees of freedom. To evaluate the number of degrees of freedom, we take the expected value of the minimum of the cost function given by (7):

$$\begin{aligned}\mathbb{E}[J(x)] &= \mathbb{E}[(x-x^b)^T B^{-1}(x-x^b)] + \mathbb{E}[(y-Hx)^T R^{-1}(y-Hx)] \\ &= \text{number of degrees of freedom of the data}\end{aligned}$$

But we know that some observations are worthless as their natural variability is less than the measurement error; these observations provide degrees of freedom related to noise, dof_n. To identify which observations provide us with useful information, we define a further transformed measurement vector $\overline{y} = U^T\tilde{y}$, where U is the matrix of left singular values of $\tilde{H}$. We find that the elements of $\overline{y}$ which vary more than the noise are those for which the singular value of $\tilde{H}$ is greater than unity. These transformed observations correspond to a measurable quantity, and provide degrees of freedom related to signal, dof_s. Large singular values correspond to well-observed directions and a significant reduction in error variance.

So, the total number of degrees of freedom of the data equals the degrees of freedom for signal (dof_s) plus the degrees of freedom for noise (dof_n); dof_s measure the part of the minimised $J(x)$ attributed to the state vector, and dof_n measure the part attributed to noise. Through linear transformations, under which degrees of freedom remain unchanged, we have a numeric representation for dof_s and dof_n in terms of singular values of $\tilde{H}$ (λ_i),

$$dof_s = \sum_{i=1}^{m} \frac{{\lambda_i}^2}{1+{\lambda_i}^2}, \tag{18}$$

$$dof_n = \sum_{i=1}^{m} \frac{1}{1+{\lambda_i}^2}, \tag{19}$$

which sum to the total number of measurements, m.

In terms of information theory, degrees of freedom of signal ‘indicate the amount of useful information contained in the observations‘; in practical terms, they indicate the number of quantities measured.

The above is a statistical approach to dof_s. Alternatively, we can use a linear algebra approach to get a numeric representation for dof_s.

We have an initial covariance matrix B, and performing an analysis to minimise the variance in observed directions gives us a posterior covariance matrix S_a. We evaluate this minimised variance by examining the eigenvalues of these two matrices: for large eigenvalues there is a large uncertainty in the direction of the associated eigenvector, and conversely, for small eigenvalues there is little uncertainty in the associated direction.

To perform a comparison of B and S_a, it would be simpler to have diagonal matrices. To this end consider a non-singular square matrix L, as in [8], such that

$$LBL^T = I \tag{20}$$

$$LS_aL^T = \hat{S}_a. \tag{21}$$

This transformation is not unique as we can replace L by X^TL where X is an orthogonal matrix, i.e, $X^TLBL^TX = X^TX = I$. Now, if we take X to be the matrix of eigenvectors of $\hat{S}_a$, then we simultaneously reduce B to the identity matrix of its eigenvalues and S_a to a diagonal matrix, Λ, i.e, $X^TLS_aL^TX = X^T\hat{S}_aX = \Lambda$.

So, after this transformation, the diagonal elements of the transformed B matrix are all unity, and each corresponds to an individual degree of freedom. The eigenvalues of $\hat{S}_a$ may therefore be interpreted as the relative reduction of variance in each of the N independent directions, where N is the dimension of the state vector. So, in the well-observed directions the corresponding eigenvalue of $\hat{S}_a$ will be small, and directions that have not been impacted by the observation will have large eigenvalues.

So, an alternative representation of dof_s is,

$$dof_s = N - trace(\Lambda), \tag{22}$$

and correspondingly for dof_n,

$$dof_n = trace(\Lambda). \tag{23}$$

This approach is equivalent to that of finding singular values of $\tilde{H}$, [8]; the benefit of using it over the statistical method will be seen later in the chapter.

3.3 Fisher Information Content

When estimating a parameter of a distribution, we want to obtain an estimate of maximum likelihood. In a Bayesian setting, this procedure is based on obtaining a set of measurements and maximising the probability of these measurements having occurred given a certain state vector, $p(y|x)$. In maximising the likelihood that we assign the correct value to a parameter, we are minimising the error in incorrectly estimating it; the Fisher Information Content, F, is a measure of this minimisation.

Rather than maximising the likelihood function $p(y|x)$, normally the log likelihood function is algebraically simpler to maximise and achieves the same goal. So, considering the function, as defined by [15],

$$\ln p(y|x) = -\frac{1}{2}(y - Hx)^T R^{-1}(y - Hx) + \text{a constant}, \tag{24}$$

the quantity

$$F = \mathbb{E}\left[\left(\frac{\delta \ln p(y|x)}{\delta x}\right)^2\right] \tag{25}$$

is known as the Fisher Information Matrix. This can also be written in the form,

$$\begin{aligned} F &= \mathbb{E}\left[\left(\frac{\delta \ln[p(x|y)/p(x)]}{\delta x}\right)^2\right] \\ &= \mathbb{E}\left[\left(\frac{\delta \ln p(x|y)}{\delta x} - \frac{\delta \ln p(x)}{\delta x}\right)^2\right] \end{aligned} \tag{26}$$

where $p(x)$ was the initial knowledge of the system and $p(x|y)$ was the knowledge after the observations.

Rewriting the SIC, we can see that a link exists between the two:

$$\begin{aligned} SIC &= \int [p(x|y)\ln p(x|y) - p(x)\ln p(x)]dx \\ &= \mathbb{E}[\ln p(x|y)] - \mathbb{E}[\ln p(x)]. \end{aligned} \tag{27}$$

If entropy is related to the volume of the state space, then the Fisher Information is related to the corresponding surface area.

In the linear Gaussian case, the Fisher Information Content is the inverse of the analysis covariance matrix S_a:

$$S_a^{-1} = F = H^T R^{-1} H + B^{-1}. \tag{28}$$

3.4 Alternative Formulae

However, suppose we were to accidentally, or even deliberately, compute the analysis covariance matrix, S_a, using the incorrect observation error covariance matrix, R; would it still be appropriate to use the above formulas to calculate information content? This question is important, as in our later analysis we will be calculating SIC and dof_s values for cases where correlations in the observation error covariance matrix are ignored.

If we were to knowingly use an incorrect R matrix in our calculations, then we would be accepting additional observation errors, which would need to be included for the analysis procedure to be correct. This can be achieved through the addition of an extra term to the analysis covariance matrix, as in [10], giving,

$$S_a^* = S_a + AR'A^T \tag{29}$$

$$A = BH^T(HBH^T + R_f)^{-1} \tag{30}$$

$$R' = R_t - R_f, \tag{31}$$

where R_t and R_f represent the true and false observation error covariance matrices respectively, and S_a and A are both evaluated at R_f.

So to calculate the Shannon Information Content when an incorrect R matrix is used, we again consider equations (17) and (18), but with S_a^* as the analysis covariance matrix, to get

$$\begin{aligned} SIC &= \frac{1}{2}\ln|S_a^{*-1}B| \\ &= \frac{1}{2}\ln|[(H^TR_f^{-1}H+B^{-1})^{-1}+AR'A^T]^{-1}B|, \end{aligned} \tag{32}$$

which does not reduce down to a nice expression with singular values of $\tilde{H}$, but is still possible to evaluate.

To calculate the degrees of freedom of signal it is easiest to manipulate the formulas produced from the linear algebra analysis as they are given directly in terms of S_a. So, from equations (20), (21) and (22),

$$dof_s = N - tr(\Lambda^*) \tag{33}$$

where $X^TLS_a^*L^TX = X^TS_a^*X = \Lambda^*$.

However, we could argue that we always knowingly use an incorrect R matrix, as it is impossible to know observation errors exactly, and hence we should be consistent with our analysis. So, in cases where we use an incorrect observation error covariance matrix, the analysis covariance matrix S_a should just be evaluated at the incorrect R matrix, R_f.

So to summarise, we have three approaches to evaluating the information content:

Approach 1 : Assume that we are using the correct R matrix, R_t,
and evaluate S_a at this value
$S_a = (H^TR_t^{-1}H+B^{-1})^{-1}$

Approach 2 : We knowingly use an incorrect R matrix and include an
additional term in the error covariance matrix to
accurately model this
$S_a^* = (H^TR_f^{-1}H+B^{-1})^{-1}+AR'A^T$

Approach 3 : Accept that we are using an incorrect R matrix,

R_f, and evaluate S_a at this value

$S_a = (H^T R_f^{-1} H + B^{-1})^{-1}$

4 Observation Error Correlations - Traditional Approaches

As mentioned in Section 2, the complexity of observation error correlations usually means that the R matrix is taken to be diagonal despite knowing this not to be the case. In most cases, to compensate for the lack of correlation, the variances in the R matrix are inflated, so that the observations have a lower weighting in the analysis. The benefits of this approach are debatable.

In [4] the impact on the analysis of using uncorrelated R matrices with different levels of inflated variance, compared to that of the true correlated R matrix was examined. Results showed that error variances can be made at most 2-4 times larger than the standard deviation of the true R matrix before the model field becomes degraded through excessive error amplification. So, whatever benefit variance enlargement has, it is limited by the need for a physically accurate model representation. The paper also concluded overall that 'if the real observation matrix has significant correlations, the approximation of a diagonal error covariance will not realise the full potential of the observations', i.e, information will be lost.

Another approach to the problem of correlated errors is the process of superobbing [1], which uses a weighted average of the differences between observations and collocated backgrounds within a 3-d box to create one superob. This lowers the effect of correlated error by reducing the data density, and reduces uncorrelated error through averaging. The benefits of this approach are the lowered risk of smoothing atmospheric features, and better optimisation of data compared to conventional methods such as data thinning.

Suppose we have N observations in a 3-d box (y_i) with corresponding background values $({x_i}^b)$, then the superob value will be,

$$s = {x_0}^b + \sum_{i=1}^{N} w_i(y_i - {x_i}^b), \tag{34}$$

where ${x_0}^b$ is the background value at the superob location and w_i are the weightings, assumed in this case to be $1/N$.

Under the assumption that innovations are equally weighted, superobbing has been found to be most effective in boxes where uncorrelated error dominates. It is further suggested that this random error reduced by superobbing is not the primary source of error. So again, the process has a limited benefit on the case when correlated observation errors are present.

Suberobbing is a method of data thinning. Data thinning can be beneficial as it compromises between the risks of having too low a data density and having correlated observation errors. But under what conditions will the best balance between observation correlation and thinning be reached?

In a recent paper [13] it was found that when assimilating data with correlated errors as if they were uncorrelated and using optimal thinning, increasing observation density led to a significant improvement in analysis accuracy, and the extraction of most of the independent information. However, for spatially correlated errors, implemented correctly in the analysis, increasing the observation density beyond some threshold value yielded very little or no improvement in analysis accuracy. The conclusion was that wrongly treating observation errors as uncorrelated limits the use of high density observations.

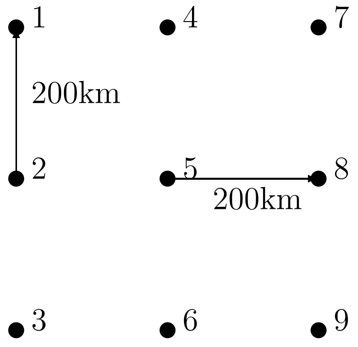
So, although implementation procedures exist for when correlated observation errors are ignored, research has shown that such methods result in a deficiency in data utilisation, and a suspected loss of information.

5 Calculation of Quantative Information Content for Atmospheric Motion Vectors

We know that to optimally extract essential information from a set of observations in an assimilation system, a good specification of observation error R is needed. It is suspected that the more accurately R is represented, the more information in the observations will be available, i.e, if observation error correlations are ignored then information will be lost. The aim of this section is to quantitatively evaluate the difference in information content between a diagonal covariance and a full covariance R matrix for a set of data on Atmospheric Motion Vectors (AMVs).

The data for this analysis comes from [2], which focuses on quantifying spatial correlations of random errors in AMVs. The paper analyses pairs of collocations between AMVs and radiosonde observations; for each pair, two different AMVs are collocated with radiosonde observations from two respective stations. Assuming the sonde errors are spatially uncorrelated, any AMV-sonde difference between stations is attributed to spatially correlated AMV errors. AMV/sonde collocation matching between stations occurs if the AMVs originate from the same imagery, and the difference between assigned pressures is less than 150hPa. This method is obviously very data intensive, and so can only take place in dense sonde networks.

Consider an idealised data set with observations on an $p \times p$ regular grid with 200km spacing, where the points are numbered column wise, i.e, for a 3×3 grid



We assume isotropic correlations given by the correlation function:

$$C_{ij} = \left(1 + \frac{r_{ij}}{L}\right) \exp\left(-\frac{r_{ij}}{L}\right) \tag{35}$$

where r_{ij} is the level spacing between point i and point j, and L is the length scale. Taking the correlated part of the AMV error as the square root of the variance, we assume that all error covariances are the same.

We will use the results from the satellite GOES-10 in the northern hemisphere (high latitude mid-level AMVs) to compute the information content for different sizes of grid: $L = 190$, $r = 200$, and $\sigma = 3.5$

Values of SIC and dof_s, for the three approaches we are considering are calculated via the various methods described in Section 3.

For **Approach 1**, the observation correlation matrix C is computed from the correlation function (35), and then combined with the p^2 by p^2 matrix of diagonal variances,

$$D = \begin{pmatrix} 12.25 & 0 & \dots & 0 \\ 0 & 12.25 & \dots & 0 \\ \vdots & \ddots & \ddots & \vdots \\ 0 & 0 & \dots & 12.25 \end{pmatrix},$$

via the formula $R = D^{1/2}CD^{1/2}$, to get the covariance matrix R. We assume this R matrix is correct, and denote it R_t.

Assuming that we observe every desired atmospheric property directly, and we have uniform uncorrelated background errors, $H = I = B$, we can calculate the singular values λ_i of $\tilde{H} = R_t^{-1/2}HB^{1/2} = R_t^{-1/2}$. From these we can deduce the Shannon Information Content (17).

For ease of calculation, consider the linear algebra approach to dof_s. Since the background error covariance matrix is the identity, the requirement on the transforming matrix is reduced to $LL^T = I$, i.e. L is an orthogonal matrix. Choose $L = I$, then $S_a = \hat{S}_a$, and so we compute the eigenvalues of S_a to deduce the dof_s (22).

To evaluate the difference in information content when using a diagonal covariance and a full covariance R matrix, we must implement **Approach 2** and **Approach 3**.

In **Approach 2**, we know that we are using an incorrect R matrix in the form of the full R matrix, R_t, with the correlations ignored, i.e,

$$R_f = \begin{pmatrix} 12.25 & 0 & \dots & 0 \\ 0 & 12.25 & \dots & 0 \\ \vdots & \ddots & \ddots & \vdots \\ 0 & 0 & \dots & 12.25 \end{pmatrix}$$

The values for SIC and dof_s are given by evaluating equations (32) and (33) for R_f, under the assumption $H = I = B$.

Approach 3 is the evaluation of equations (17) and (22) as in **Approach 1**, only now with R_f as the observation error correlation matrix.

6 Results

6.1 Shannon Information Content and Degrees of Freedom of Signal

Tables 1 and 2 clearly show the benefits in terms of information content, of using a fully correlated R matrix as opposed to a diagonal approximation. For all sizes of grid examined, the Shannon Information Content and the number of degrees of freedom of signal are larger when the full R matrix is used (**Approach 1**) rather than either diagonal one (**Approach 2 or 3**). Comparing the two diagonal approximations (**Approach 2 and 3**), we see that when the Shannon Information Content is used as the measure of information, **Approach 2** gives a greater reduction in uncertainty. However, when the number of degrees of freedom of signal is used as the information measure, the value is no different for either approach.

To explain why the number of degrees of freedom of signal stay the same, we examine the trace of the matrices S_a and S_a^*, under the assumptions $H = I = B$. The trace of a matrix is equal to the sum of its eigenvalues, and so the equations for degrees of freedom of signal, (22) and (33), can be written:

Approach 2: $dof_s = N - \sum$ eigenvalues of S_a^*

Approach 3: $dof_s = N - \sum$ eigenvalues of S_a

where

$$\begin{aligned} S_a &= (I + R_f^{-1})^{-1} \\ S_a^* &= S_a + (I + R_f)^{-1}(R_t - R_f)(I + R_f)^{-1^T}. \end{aligned}$$

Since the matrix $R_t - R_f$ has zeros on the diagonal and $(I + R_f)^{-1}$ is a diagonal matrix, the second term of S_a^* will also have zeros on the diagonal and a zero trace. The sum of the eigenvalues of the second term of S_a^* is now zero, and hence the sum of the eigenvalues of S_a will be equal to the sum of the eigenvalues of S_a^*. Therefore, under **Approach 2** and **Approach 3**, the dof_s will be the same.

This just one specific case; under what general conditions would dof_s be the same for **Approach 2** and **Approach 3**? Firstly, suppose our assumption of $H = I = B$ still holds, and that the observation error variances in R_f are not necessarily the same as those in R_t, i.e, $R_t - R_f$ does not necessarily have zeros on the diagonal. Here, $(I + R_f)^{-1}$ is still diagonal, and if we define μ_i^2 as the approximated error variance of observation i in R_f, and k_i as the difference in variance between R_t and R_f, then the requirement for equal degrees of freedom of signal becomes

$$\sum_{i=1}^{n} \frac{k_i}{(1 + \mu_i^2)^2} = 0. \qquad (36)$$

So, if we use the correct variances in our approximation of the full observation error covariance matrix, then the number of degrees of freedom of signal will be the same for both **Approach 2** and **Approach 3**.

However, if $H \neq I \neq B$, then we are evaluating more complicated equations in (22) and (33). For **Approaches 2** and **3** to produce the same degrees of freedom, we require:

$$\mathrm{trace}[AR'A^T] = 0, \qquad (37)$$

which expanded is,

$$\mathrm{trace}[BH^T(HBH^T + R_f)^{-1}(R_t - R_f)(HBH^T + R_f)^{-1^T} HB^T] = 0. \qquad (38)$$

This equation cannot be simplified into a nice condition as in the case $H = I = B$, and is unlikely to hold in more realistic models.

From Figures 1 and 2 we can see that the Shannon Information Content and number of dof_s are directly proportional to the square of the number of columns (or rows) of the grid, ie. the number of observation points. The figures further demonstrate that, as the size of the grid increases, there is an increased difference between the information content using **Approach 1**, and the information content using **Approach 2** and **3**. Note that in Figure 2, the line for **Approach 2** is underneath the line for **Approach 3** since the number of dof_s are the same for these two methods.

In section 4, we discussed variance enlargement as an alternative to the use of a fully correlated matrix. Considering this approach, our results indicate that increasing the variance of R diagonal causes a reduction in the number of dof_s and SIC for all grid sizes (Table 2 and 3, Figures 3-6). However, when compared to the difference in information content between **Approach 1** and **Approach 2** or **3**, this reduction is not that significant. It is interesting that the number of dof_s is now different for **Approach 2** and **Approach 3**. This happens because equation (36) is no longer satisfied, i.e, the diagonal approximation of the full R matrix has incorrect variances on the diagonal.

Again comparing the two approaches of diagonal approximation, we see that, as in the standard variance case, if SIC is used as the measure of information content, **Approach 2** provides us with more information. Also, using **Approach 2**, the number of dof_s is larger. This makes intuitive sense, as we have a greater reduction in uncertainty through more observation information.

6.2 Fisher Information Content

Using the assumptions of a linear Gaussian data distribution and $H = B = I$, the Fisher Information Matrix, given by (28), takes the form:

$$\textbf{Approach 1} \qquad F = R_t^{-1} + I \tag{39}$$

$$\begin{aligned}\textbf{Approach 2} \qquad F = {} & [(R_f^{-1} + I)^{-1} \\ & +(I + R_f)^{-1}(R_t - R_f)(I + R_f)^{-1^T}]^{-1}\end{aligned} \tag{40}$$

$$\textbf{Approach 3} \qquad F = R_f^{-1} + I \tag{41}$$

The Fisher Information Matrix is a measure of the minimum error in estimating our variables. We have assumed that all our variables were observed directly, and so this error is purely measurement based when we assume that we are using the correct matrices.

For a full R matrix, as in **Approach 1**, F will contain non-diagonal elements representing correlations between measurement errors. In **Approach 2**, we are using a diagonal R matrix which we are know is incorrect, so F will also have non-diagonal elements. However, these elements correspond to additionally accepted observation errors included to make the analysis correct, and not from individual correlations between measurements errors, since all variable measurements are assumed to be independent. In **Approach 3**, F is diagonal, which is expected as all variable measurement errors are assumed to be independent, as R is diagonal in this approach as well.

For example, in the 2×2 case for **Approach 1**:

$$F = \begin{pmatrix} 1.2583 & -0.1516 & -0.1516 & 0.0721 \\ -0.1516 & 1.2583 & 0.0721 & -0.1516 \\ -0.1516 & 0.0721 & 1.2583 & -0.1516 \\ 0.0721 & -0.1516 & -0.1516 & 1.2583 \end{pmatrix}$$

and for **Approach 2**:

$$F = \begin{pmatrix} 1.0892 & -0.0544 & -0.0544 & -0.0403 \\ -0.0544 & 1.0892 & -0.0403 & -0.0544 \\ -0.0544 & -0.0403 & 1.0892 & -0.0544 \\ -0.0403 & -0.0544 & -0.0544 & 1.0892 \end{pmatrix}$$

and for **Approach 3**:

$$F = \begin{pmatrix} 1.0816 & 0 & 0 & 0 \\ 0 & 1.0816 & 0 & 0 \\ 0 & 0 & 1.0816 & 0 \\ 0 & 0 & 0 & 1.0816 \end{pmatrix}$$

Here, when we have assumed correlations are present (**Approach 1**) or that we are using an incorrect R matrix (**Approach 2**), the minimum error in estimating an observation is greater than when we have assumed an idealised case of no correlations and the correct R matrix (**Approach 3**). This again makes intuitive sense. However, the Fisher Information Matrix would be more interesting to analyse if the variables were not observed directly; then the error in estimating these observations would arise from more than these two sources.

Size of	Entropy	**Approach 1**		**Approach 2**		**Approach 3**	
Grid	Before	Entropy after	*SIC*	Entropy after	*SIC*	Entropy after	*SIC*
2×2	5.676	5.2464	0.4296	5.5118	0.1642	5.5191	0.1569
3×3	12.770	11.5124	1.2576	12.3910	0.3790	12.4169	0.3531
4×4	22.703	20.2122	2.4908	22.0194	0.6836	22.0752	0.6278
5×5	35.473	31.3509	4.1221	34.3953	1.0777	34.4921	0.9809
6×6	51.082	44.9315	6.1505	49.5207	1.5613	49.6695	1.4125
7×7	69.528	60.9521	8.5759	67.3936	2.1344	67.6054	1.9226
8×8	90.812	79.4137	11.3983	88.0149	2.7971	88.3009	2.5111
9×9	114.934	100.3163	14.6177	111.3848	3.5492	111.7559	3.1781
10×10	141.894	123.6599	18.2341	137.5032	4.3908	137.9704	3.9236
20×20	567.575	491.3414	76.2336	549.8452	17.7298	551.8807	15.6943

Table 1: Shannon Information Content results

Size of Grid	Degrees of freedom of signal		
	Approach 1	**Approach 2**	**Approach 3**
2×2	0.7284	0.3019	0.3019
3×3	2.0286	0.6792	0.6792
4×4	3.9346	1.2075	1.2075
5×5	6.4418	1.8868	1.8868
6×6	9.5501	2.7170	2.7170
7×7	13.2595	3.6981	3.6981
8×8	17.5701	4.8302	4.8302
9×9	22.4818	6.1132	6.1132
10×10	27.9947	7.5472	7.5472
20×20	116.1857	30.1887	30.1887

Table 2: Degrees of freedom results

Size of Grid	**Approach 2**		**Approach 3**	
	SIC	dof_s	SIC	$dofs$
2×2	0.1201	0.2322	0.0800	0.1569
3×3	0.2711	0.5225	0.1800	0.3529
4×4	0.4829	0.9289	0.3200	0.6275
5×5	0.7555	1.4514	0.5001	0.9804
6×6	1.0889	2.0900	0.7201	1.4118
7×7	1.4830	2.8447	0.9801	1.9216
8×8	1.9379	3.7155	1.2802	2.5098
9×9	2.4536	4.7024	1.6202	3.1765
10×10	3.0301	5.8055	2.0003	3.9216
20×20	12.1374	23.2218	8.0011	15.6863

Table 3: Information content results (2× variance)

Size of Grid	**Approach 2**		**Approach 3**	
	SIC	dof_s	SIC	$dofs$
2×2	0.0707	0.1388	0.0404	0.0800
3×3	0.1591	0.3123	0.0909	0.1800
4×4	0.2828	0.5552	0.1616	0.3200
5×5	0.4420	0.8675	0.2525	0.5000
6×6	0.6365	1.2492	0.3636	0.7200
7×7	0.8665	1.7003	0.4950	0.9800
8×8	1.1318	2.2208	0.6465	1.2800
9×9	1.4325	2.8107	0.8182	1.6200
10×10	1.7685	3.4700	1.0101	2.0000
20×20	7.0755	13.8800	4.0405	8.0000

Table 4: Information content results (4× variance)

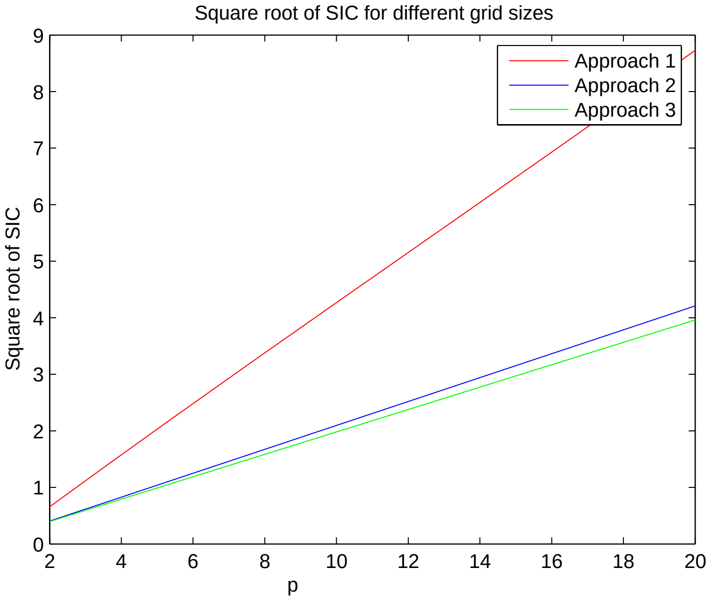


Figure 1: Relationship between grid size and SIC

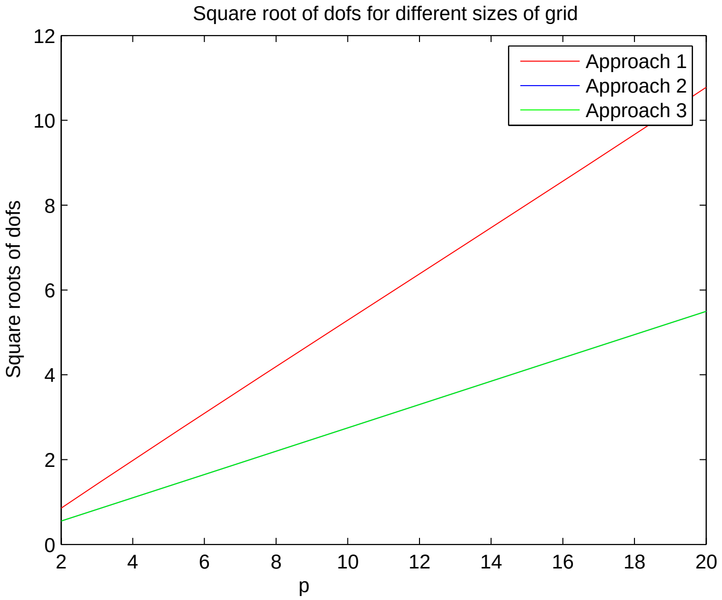


Figure 2: Relationship between grid size and dof_s

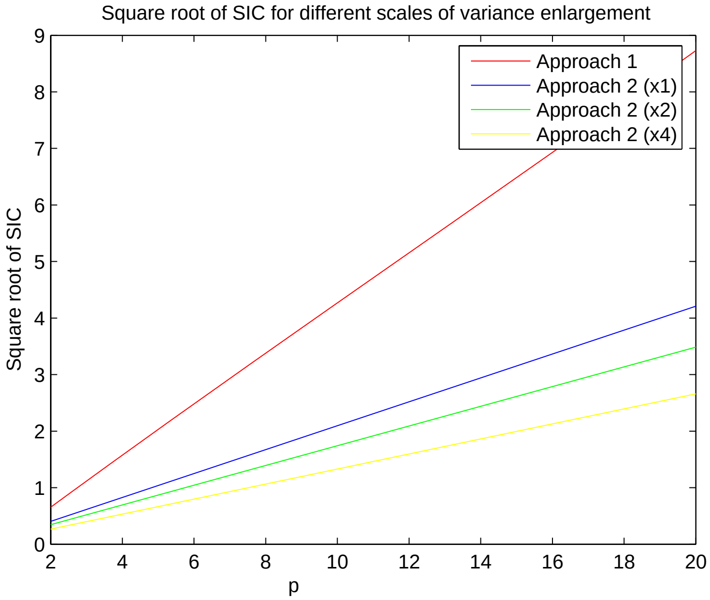


Figure 3: SIC for different scales of variance enlargement - **Approach 2**

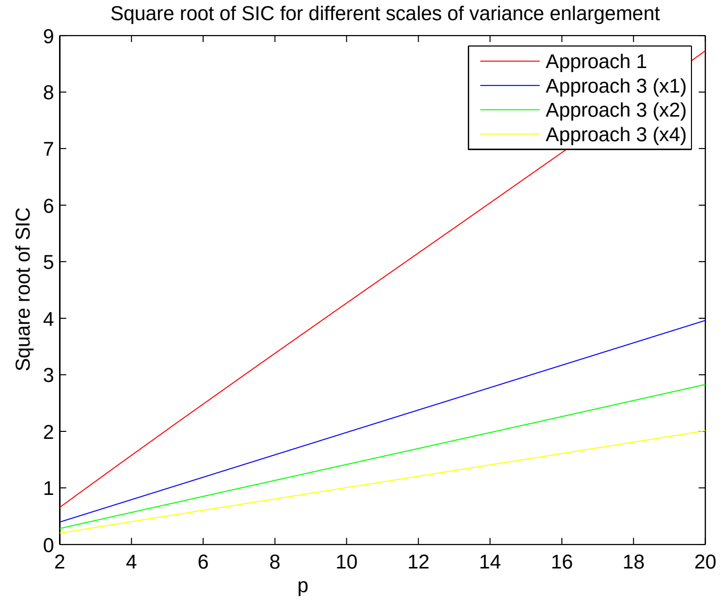


Figure 4: SIC for different scales of variance enlargement - **Approach 3**

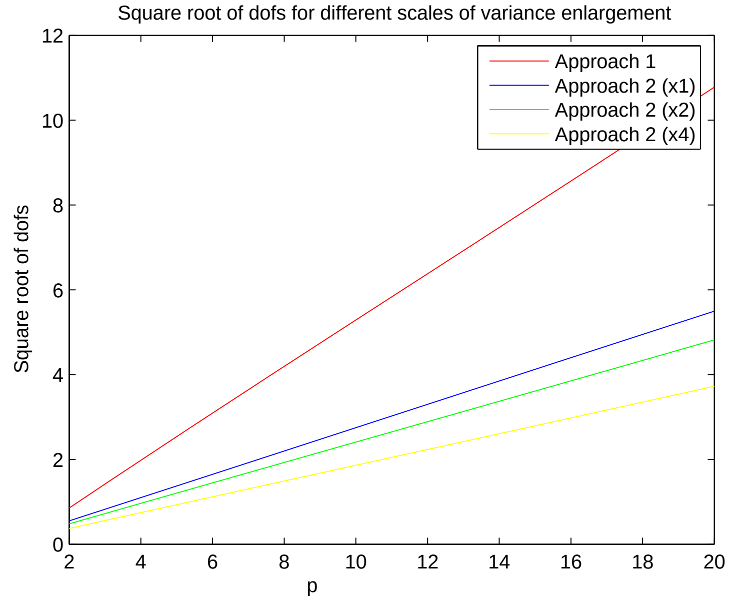


Figure 5: dof_s for different scales of variance enlargement - **Approach 2**

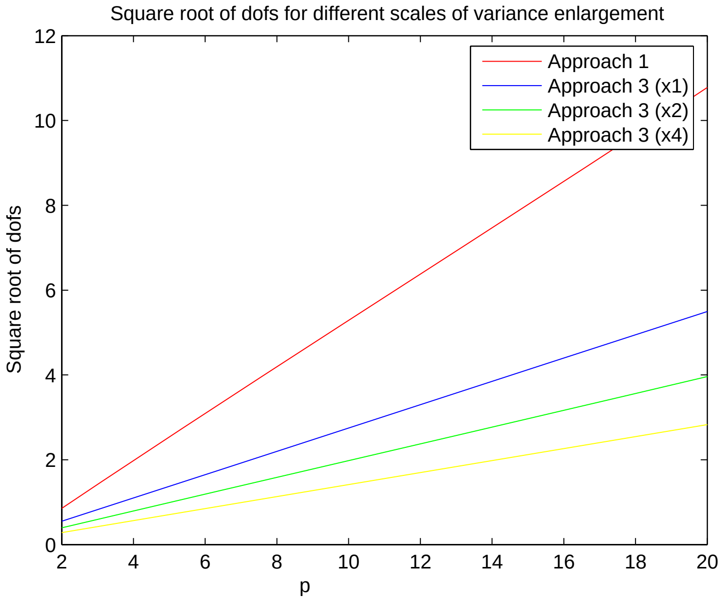


Figure 6: dof_s for different scales of variance enlargement - **Approach 3**

7 Conclusions and Future Work

Implications that using a diagonal error covariance matrix as opposed to a fully correlated one results in a significant loss in information, could possibly lead to a re-evaluation of the assumptions made when obtaining initial conditions through the solution of the inverse problem. So, it is important that any trends in information loss are identified and explained if possible.

For all sizes of grid ($p \times p$) analysed, using the full R matrix (**Approach 1**), as opposed a diagonal one (**Approach 2** or **3**), gives a greater Shannon Information Content and a greater number of degrees of freedom of signal. So, as suspected, we lose information by using a diagonal R matrix.

For all three approaches, the SIC and number of dof_s are directly proportional to the number of observation points. However the gradient of proportionality varies; from Figure 1 and 2 we see that the gradient of the line for **Approach 1** is steeper than that for **Approach 2** or **3**. So, as we take more observations, the more important it becomes, in a sense of information optimisation, that we have a fully specified R matrix.

Comparing **Approach 2** and **Approach 3**, we find that **Approach 2** gives a greater reduction in uncertainty, i.e, a greater SIC, and more useful information, i.e, a greater number of dof_s. This implies that by using Bayesian philosophy, we have actually reduced the uncertainty more than we might think we have if we had simply used R_f as our 'correct' R matrix. In practice, we must always use this Bayesian philosophy of **Approach 2**, where we make calculations based on our best knowledge, which is never the truth.

Variance enlargement in a diagonal R matrix has a detrimental effect on the values of both SIC and dof_s for all grid sizes. As the scale of variance enlargement increases, the gradient of the line for the modified **Approach 2** and **Approach 3** decreases (Figures 3-6). So, as grid size increases, the greater the negative influence variance enlargement will have on information content. But as previously mentioned, when compared to the differences between **Approach 1** and **Approach 2** or **3**, this impact has limited significance.

The structure of this experiment is obviously very basic, and the assumptions of regular observation spacing and identical error variances are in reality not the case; but simplification is required, and modifying other factors in the experiment is likely to result in a more significant and telling change. The data we used from [2] was assumed to be directly observed, and any background error, uncorrelated and uniform. It would be more interesting to examine raw radiance data, or data directly converted from radiances for variables related to those that we are interested in (i.e, $H \neq I$). This could be done under the further assumption of correlated background errors; the subject and definition of which is examined in many papers.

Although our results suggest that it would be significantly beneficial, in terms of information utilised, to use a fully correlated R matrix when solving the inverse problem, we have not addressed the problems in implementing this. Obviously the inverse of R is considerably more computationally expensive to calculate when R is non-diagonal, especially as our grid size becomes larger, and hence more realistic. We need to have some idea of whether this extra information is worth the added computational time and expense, or even if it is possible to use a fully correlated R matrix in a numerically cheap way.

Also, as mentioned briefly in Section 2.1, using a correlated error covariance matrix reduces the weighting given to observations in the analysis, but gives more relative importance to the difference between observations. This could be examined further by considering the relationship between pressure and velocity in geostrophic motion:

$$\nabla p = \rho f \underline{v} \wedge \underline{k}, \tag{42}$$

where p is the pressure, ρ is the density, f is the Coriolis force, $\underline{v}$ is the wind velocity and $\underline{k}$ is a unit normal.

We would suspect that if error correlations were present in pressure observations then these observations would provide us with less knowledge about the pressure at each point, but give a better estimate for the pressure gradient, and hence more knowledge about the wind velocity. Whether, in addition, the conveyance of this balance property enables us to extract more information from the observations would also be interesting to investigate.

References

References

[1] Berger, H. and M. Forsythe, 2004. *Satellite Wind Superobbing.* Forecasting Research Technical Report No. 451, Met Office.

[2] Bormann, N., S. Saarinen, G. Kelly and J-N. Thepaut, 2002. *The spatial structure of observation errors in tmospheric Motion Vectors from geostationary satellite data.* EUMETSAT/ECMWF Fellowship Programme, Research Report No.12

[3] Bouttier, F. and P. Courtier, 1999. *Data assimilation concepts and methods.* Meteorological Training Course Lecture Series, ECMWF.

[4] Collard, A., 2004. *On the choice of observation errors for the Assimilation of AIRS brightness temperatures: A theoretical study.* ECMWF report.

[5] Cover, T.M. and J.A. Thomas, 1991. *Elements of Information Theory, Wiles Series in Telecommunications.* John Wiley and Sons.

[6] Daley, R., 1991. *Atmosheric Data Analysis.* Cambridge University Press.

[7] Fisher, M., 2003. *Background error covariance modelling, Recent developments in data assimilation for atsmosphere and ocean.* ECMWF Seminar Proceedings, 45-64.

[8] Fisher, M., 2003. *Estimation of Entropy Reduction and Degrees of Freedom for Signal for Large Variational Analysis Systems.* ECMWF Technical Memoranda

[9] Golub, G.H. and C.F. Van Loan, 1996. *Matrix computations.* The John Hopkins University Press, third edition.

[10] Healy, S.B. and A.A. White, 2003. *Use of discrete Fourier transforms in the 1D-Var retrieval problem. Q.J.R.Met.Soc.*,**131**:63-72.

[11] Johnson, C., 2003. *Information Content of Observations in Variational Data Assimilation.* PhD thesis, Department of Mathematics, University of Reading.

[12] Kalnay, E., 2003. *Atmospheric Data Modelling, Data Assimilation and Predictability.* Cambridge University Press.

[13] Liu, Z-Q. and F. Rabier, 2002. *The interaction between model resolution, observation resolution and observation density in data assimilation: A one-dimensional study. Q.J.R.Met.Soc.*,**128**:1367-1386.

[14] MetOffice, 2005. *Global Data Monitoring Statistics.*
URL http://www.metoffice.com/research/nwp/observations/
monitoring/month_rep/docs/05-gdms.html.

[15] Rodgers, C.D., 2000. *Inverse Methods of Atmospheric Sounding, Theory and Practice, volume 2 of Atmospheric, Oceanic and Planetary Physics.* World Scientific.

[16] Thepaut, J-N., 2004. *Satellite Data Assimilation in Numerical Weather Prediction: an Overview.* ECMWF report.