Abstract

I ag c c a a a a , a c b b -ca c a b a a - a bac - g a . T c fica a a c- c b a a bac g - fl c a a . F a c a , b a a a b ac , . ., c a ; , a a a a ca .

Contents

observation information content for di erent 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

between the measured quantity and the desired one. Taking measurements for

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

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]

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

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

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

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

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

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 $( )$ is the knowledge before the observation, $(\mid)$ is the knowledge after, and $S( )$ is the entropy. Then the Shannon Information Content, SIC, as defined by [15] is

$$SIC = S[\,\cdot\,)] - S[\,\cdot\,|\,)] \tag{14}$$

where

$$S[\,\cdot\,)] = -\int \,\cdot\,) \lg_2[\,\cdot\,)]\,\cdot\,, \tag{15}$$

$$S[\,\cdot\,|\,)] = -\int \,\cdot\,|\,) \lg_2[\,\cdot\,|\,)]\,\cdot\,. \tag{16}$$

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:

$$S[\,\cdot\,)] = \ln(2\pi\,)^{1/2} + \frac{1}{2}\ln|B|$$

$$S[\,\cdot\,|\,)] = \ln(2\pi\,)^{1/2} + \frac{1}{2}\ln|S_a|$$

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

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

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

So, an alternative representation of $_s$ is,

$$_s = N - \quad (\Lambda), \tag{22}$$

and correspondingly for $_n$,

$$_n = \quad (\Lambda). \tag{23}$$

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, ( ). 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 ( ), normally the log likelihood function is algebraically simpler to maximise and achieves the same goal. So, considering the function, as defined by [15],

$$\ln\ (\ ) = -\frac{1}{2}(\ - H\ )^T R^{-1}(\ - H\ ) + \text{a constant}, \tag{24}$$

the quantity

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

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

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

where ( ) was the initial knowledge of the system and (|) was the knowledge after the observations.

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

$$SIC \quad =$$

covariance matrix, to get

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

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),

$$_s = N - \quad (\Lambda^*) \tag{33}$$

where $^T L S_a^* L^T \quad = \quad ^T S_a^* \quad = \Lambda^*$.

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

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

Approach 2 : We knowingly use an incorrect R matrix and include an additional term in the error covariance matrix to accurately model this

$S_a^* = (H^T R_f^{-1} H + B^{-1})^{-1} + AR'A^T$

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 di erent 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 di erences 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 ($_i$) with corresponding background values ($_i^b$), then the superob value will be,

$$\mathcal{S} = \_0^b + \sum_{i\ 1}^{N} \ _i(\ _i - \_i^b), \tag{34}$$

where $\_0^b$ is the background value at the superob location and $\ _i$ 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 e ective 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

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

where $_{ij}$ 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 di erent sizes of grid: L = 190, = 200, and σ = 3.5

Values of SIC and $_s$, 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 ^{2} by ^{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 $_s$. Since the R , T c ( ) T

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 $_s$ 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

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 $_s$ will be the same.

This just one specific case; under what general conditions would $_s$ 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 in R_f, and κ_i as the di erence in variance between R_t and R_f, then the requirement for equal degrees of freedom of signal becomes

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

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:

$$\text{trace}[AR'A^T] = 0, \tag{37}$$

which expanded is,

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

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 $_s$ 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 di erence 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 $_s$ are the same for these two methods.

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:

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

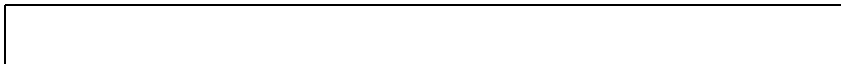
$$\text{Approach 2} \qquad F = [(R_f^{-1} + I)^{-1} + (I + R_f)^{-1}(R_t - R_f)(I + R_f)^{-1^T}]^{-1} \tag{40}$$

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

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 err0.6(f)- 4.44632.1(v)53=l.5(-1.1(o)-30e/F6 1 T0e/F6 1 .8(TDR61.

Size of Grid	Entropy Before	Approach 1 Entropy after	SIC	Approach 2 Entropy after	SIC	Approach 3 Entropy after	SIC



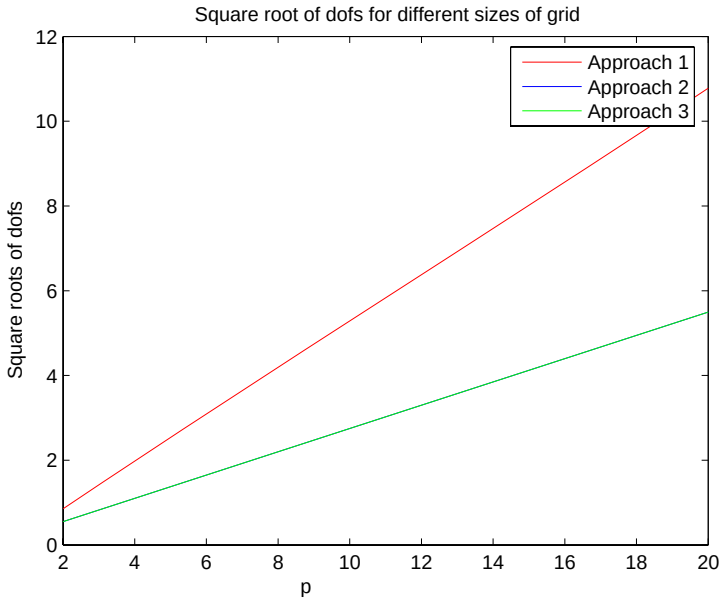


Figure 2: Relationship between grid size and $_{s}$

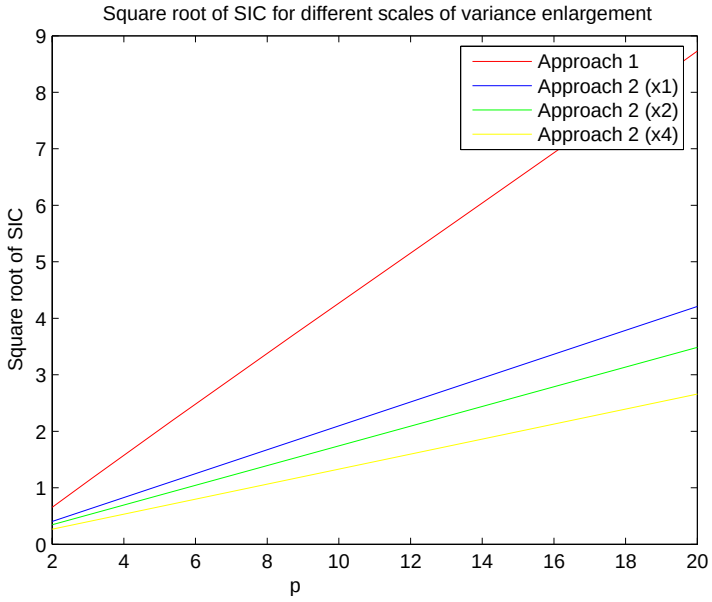


Figure 3: SIC for di erent scales of variance enlargement - Approach 2

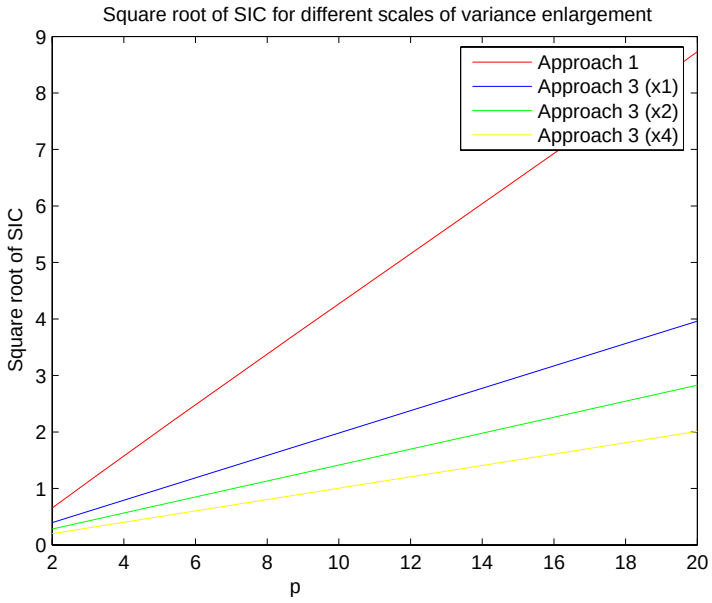
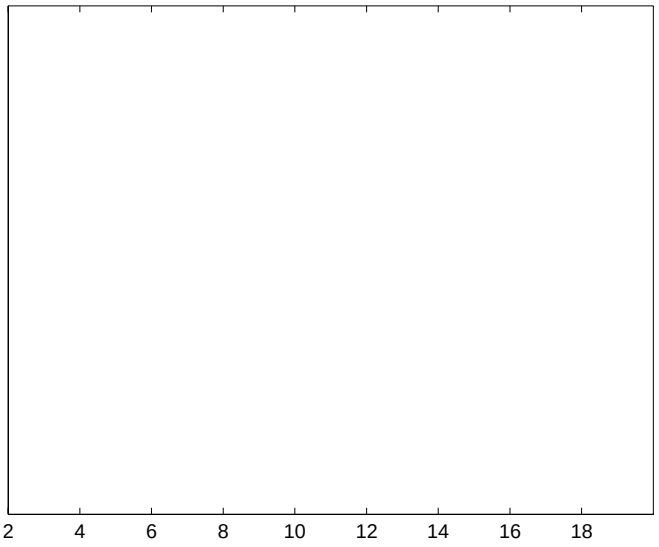
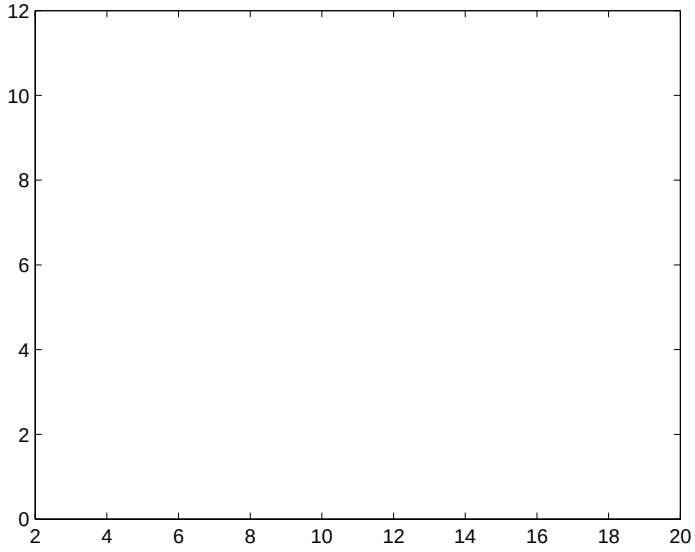


Figure 4: SIC for di erent scales of variance enlargement - Approach 3




12
10
8
6
4
2
0
2
4
6
8
10
12
14
16
18
20

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 ($\times$) 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 $_s$ 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 $_s$. 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

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.

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 O ce.

[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*