Manuscript prepared for Atmos. Chem. Phys.
with version 3.2 of the LaTeX class copernicus.cls.
Date: 8 March 2012

A Practical Method to Estimate Information Content in the Context of 4D-Var Data Assimilation. I: Methodology

A. Sandu[1], K. Singh[1], M. Jardak[1], K. W. Bowman[2], and M. Lee[2]

[1]Department of Computer Science, Virginia Polytechnic Institute and State University, 2202 Kraft Drive, Blacksburg, VA 24060, USA
[2]Jet Propulsion Laboratory, 4800 Oak Grove Drive, Pasadena, CA 91109, USA

Correspondence to: Adrian Sandu (sandu@cs.vt.edu)

Abstract. Data assimilation obtains improved estimates of the state of a physical system by combining imperfect model results with sparse and noisy observations of reality. all observations used in data assimilation are equally valuable. The ability to characterize the usefulness of different data points is important for analyzing the effectiveness of the assimilation system, for data pruning, and for the design of future sensor systems.

This paper focuses on the four dimensional variational (4D-Var) data assimilation framework. Metrics from information theory are used to quantify the contribution of observations to decreasing the uncertainty with which the system state is known. We establish an interesting relationship between different information-theoretic metrics and the variational cost function/gradient under Gaussian linear assumptions. Based on this insight we derive an ensemble-based computational procedure to estimate the information content of various observations in the context of 4D-Var. The approach is illustrated on a nonlinear test problem. In the companion paper (Singh et al., 2012a) the methodology is applied to a global chemical data assimilation experiment.

1 Introduction

The ability to characterize the usefulness of different observation locations in data assimilation is important in analyzing the effectiveness of the assimilation system, for data pruning/data selection, for the design of future sensor systems, and for defining strategies for targeting observations. In order to quantify the contribution of observations in improving the state estimate obtained through data assimilation, we employ metrics from information theory. Broadly speaking, the information content of a data in information theory describes the amount of novelty brought in by that data. Information theory was devised in the field of electrical engineering and since then has been applied to diverse areas as complexity theory, network analysis, financial mathematics and mathematical statistics.

In the context of data assimilation, the information content of observations is loosely defined by their contribution to decreasing the uncertainty in the state estimate (Fisher, 1922). Several of the information theoretic metrics employed here measure the decrease in the (co-)variance of the error (the trace of the Fisher information matrix, the Shannon information, and the degrees of freedom for signal). Others measure the benefit of data assimilation in terms of adjusting the mean of the distribution (the signal information). Relative entropy offers a combination of both mean and variance effects.

Information theory has been used in atmospheric sciences for uncertainty studies, instrument development, and data selection. Abramov and Majda (2004); Majda and Wang (2006) propose the use the relative entropy to quantify the lack of information in climate systems; their approach is applicable to non-Gaussian distributions and non-linear models. They demonstrate the methodology with two "toy" models, Burgers-Hopf and Lorenz '96 (Lorenz, 1996); the approach becomes computationally intractable for real large scale models. Information theoretic metrics like the entropy reduction and the degrees of freedom for signal are being used in the development of remote-sounding instruments (Rodgers, 1996, 1998, 2000; Rabier et al., 2002; Worden et al., 2004). Data selection strategies were defined using information theory (Rabier et al., 2002).

The information theory has recently been used in data assimilation to characterize the information content of various observations (i.e., the usefulness of these observations). Fisher (2003) proposes methods to estimate the entropy reduction and degrees of freedom for signal with large variational analysis systems. Cardinali et al. (2004) study the influence-matrix diagnostic of data assimilation systems. Xu (2006) analyzes the relative entropy versus Shannon entropy difference to measure information content from observations for data assimilation. Zupanski et al. (2007) discusses the use of information measures in ensemble data assimilation.

In this paper we discuss a characterization of the information content of observations in the context of four dimensional variational (4D-Var) data assimilation framework. The analysis carried out in this paper assumes that errors are normally distributed and that the model dynamics is linear. It is shown

that, under these assumptions, the posterior statistics of the variational cost function and its gradient can be used to quantify the information content of observations. This result leads to the following computational procedure. After data assimilation is complete, an ensemble of simulations is carried out with initial conditions drawn from (an approximation of) the analysis probability distribution. Mean values of the cost function and of adjoint norms are used to estimate the information content of various observations in the context of 4D-Var. Note that all information metrics obtained here are with respect to the beginning of each assimilation window (as 4D-Var provides the analysis in form of the model initial conditions).

The paper is organized as follows. Section 2 reviews the variational approach to data assimilation from a Bayesian perspective. Various metrics for information content are discussed in Section 3. Section 4 develops computationally feasible estimation techniques for the information content of observations in the context of 4D-Var data assimilation; this is the main contribution of this work. The numerical results are presented and discussed in Section 5. Section 6 summarizes the findings of this work and points to future research directions.

2 Variational Data Assimilation

Variational methods solve the data assimilation problem in an optimal control framework (Courtier and Talagrand, 1987; Le Dimet and Talagrand, 1986; Lions, 1971). Specifically, one finds the control variable values (e.g., initial conditions) which minimize the discrepancy between model forecast and observations; the minimization is subject to the governing dynamic equations, which are imposed as strong constraints. In this discussion, for simplicity of presentation, we focus on discrete models where the initial conditions are the control variables.

Consider that the true state of the system $\mathbf{x}^{\text{true}} \in \mathbb{R}^n$ is unknown and needs to be estimated form the available information. In order to obtain an estimate of $\mathbf{x}^{\text{true}}$ *data assimilation combines three different sources of information*, as follows.

The background (prior) probability density $\mathcal{P}^{\text{B}}(\mathbf{x})$ encapsulates our current knowledge of the true state of the system. Specifically, it describes the uncertainty with which one knows $\mathbf{x}^{\text{true}}$ at a given moment, before any (new) measurements are taken. The mean taken with respect to this probability density is denoted by $\mathbb{E}^{\text{B}}[\cdot]$. The current best estimate of the true state is called the apriori, or the *background state* $\mathbf{x}^{\text{B}}$. The background estimation errors $\varepsilon^{\text{B}} = \mathbf{x}^{\text{B}} - \mathbf{x}^{\text{true}} \in \mathcal{N}(\mathbf{0}, \mathbf{B})$ are assumed Gaussian and are characterized by the *background error covariance matrix* $\mathbf{B} \in \mathbb{R}^{n \times n}$. With many nonlinear models this assumption is difficult to justify, but is nevertheless widely used because of its convenience.

The model encapsulates our knowledge about physical and chemical laws that govern the evolution of the system. The model evolves an initial state $\mathbf{x}_0 \in \mathbb{R}^n$ at the initial time t_0 to future state values $\mathbf{x}_i \in \mathbb{R}^n$

at future times t_i,

$$\mathbf{x}_i = \mathcal{M}_{t_0 \to t_i}(\mathbf{x}_0). \tag{1}$$

The size of the state space in realistic chemical transport models is very large. For example, a GEOS-Chem simulation at the $2^o \times 2.5^o$ horizontal resolution has $n \in \mathcal{O}\left(10^8\right)$ variables.

Observations represent snapshots of reality available at several discrete time moments. Specifically, measurements $\mathbf{y}_i \in \mathbb{R}^m$ of the true state are taken at times t_i, $i = 1,\ldots,N$

$$\mathbf{y}_i = \mathcal{H}(\mathbf{x}_i) - \varepsilon_i^{\text{obs}} \ , \quad i = 1,\ldots,N \ . \tag{2}$$

The observation operator $\mathcal{H}$ maps the model state space onto the observation space. The *observation error* term $\varepsilon_i^{\text{obs}}$ accounts for both the measurement (instrument) errors, as well as representativeness errors (i.e., errors in the accuracy with which the model can reproduce reality). Typically observation errors are assumed unbiased and normally distributed

$$\varepsilon_i^{\text{obs}} \in \mathcal{N}(0, \mathbf{R}_i) \ , \quad i = 1,\ldots,N \ . \tag{3}$$

Moreover, observation errors at different times ($\varepsilon_i^{\text{obs}}$ and $\varepsilon_j^{\text{obs}}$ for $i \neq j$) are assumed to be independent.

Based on these three sources of information data assimilation computes the analysis (posterior) probability density $\mathcal{P}^{\text{A}}(\mathbf{x})$. Specifically, $\mathcal{P}^{\text{A}}(\mathbf{x})$ describes the uncertainty with which one knows $\mathbf{x}^{\text{true}}$ after all the information available from measurements has been accounted for. The mean taken with respect to this probability density is denoted by $\mathbb{E}^{\text{A}}[\cdot]$. The best estimate $\mathbf{x}^{\text{A}}$ is called the aposteriori, or the *analysis state*. The analysis estimation errors $\varepsilon^{\text{A}} = \mathbf{x}^{\text{A}} - \mathbf{x}^{\text{true}}$ are characterized by the *analysis error covariance matrix* $\mathbf{A} \in \mathbb{R}^{n \times n}$.

If both the the background and the observation errors are Gaussian, and the error propagation through the model (1) is linear, then he probability density of the analysis (estimation) errors ε^{A} is also Gaussian,

$$\varepsilon^{\text{A}} = \mathbf{x}^{\text{A}} - \mathbf{x}^{\text{true}} \in \mathcal{N}(0, \mathbf{A}) \quad \Leftrightarrow \quad \mathcal{P}^{\text{A}}(\mathbf{x}) = \mathcal{N}\left(\mathbf{x}^{\text{A}}, \mathbf{A}\right) \ . \tag{4}$$

2.1 The Bayesian point of view to data assimilation

The estimation problem is posed in a Bayesian framework. The analysis probability density is the probability density of the state *conditioned by all the available observations* $\mathbf{y} = [\mathbf{y}_1, \cdots, \mathbf{y}_N]$. Bayes theorem allows to express the analysis probability density as follows:

$$\mathcal{P}^{\text{A}}(\mathbf{x}) = \mathcal{P}(\mathbf{x}|\mathbf{y}) = \frac{\mathcal{P}(\mathbf{y}|\mathbf{x}) \cdot \mathcal{P}^{\text{B}}(\mathbf{x})}{\mathcal{P}(\mathbf{y})}, \tag{5}$$

The denominator $\mathcal{P}(\mathbf{y})$ is the marginal probability density of the observations and plays the role of a scaling factor. The probability of the observations conditioned by the states $\mathcal{P}(\mathbf{y}|\mathbf{x})$ is the probability

that the observation errors in (2) assume certain values. If the observation errors at different times are independent, and the observation errors are Gaussian (3), we have that

$$\mathcal{P}(\mathbf{y}|\mathbf{x}) = \frac{1}{(2\pi)^{mN/2}\sqrt{\prod_{i=1}^{N}\det\mathbf{R}_i}} \exp\left(-\frac{1}{2}\sum_{i=1}^{N}(\mathcal{H}(\mathbf{x}_i)-\mathbf{y}_i)^T\mathbf{R}_i^{-1}(\mathcal{H}(\mathbf{x}_i)-\mathbf{y}_i)\right) . \tag{6}$$

In the maximum likelihood approach one looks for the argument that maximizes the posterior distribution, or, equivalently, minimizes its negative logarithm:

$$\mathbf{x}^{\mathrm{A}} = \underset{\mathbf{x}}{\operatorname{argmax}}\,\mathcal{P}^{\mathrm{A}}(\mathbf{x}) = \underset{\mathbf{x}}{\operatorname{argmin}}\,\mathcal{J}(\mathbf{x}) , \quad \mathcal{J}(\mathbf{x}) = -\ln\mathcal{P}^{\mathrm{A}}(\mathbf{x}) . \tag{7}$$

In this context the data assimilation problem is formulated as an optimization problem. Using (5) the minimization cost function can be written as

$$\underbrace{-\ln\mathcal{P}^{\mathrm{A}}(\mathbf{x})}_{\mathcal{J}(\mathbf{x})} = \underbrace{-\ln\mathcal{P}^{\mathrm{B}}(\mathbf{x})}_{\mathcal{J}^{\mathrm{B}}(\mathbf{x})+\mathrm{const}} \underbrace{-\ln\mathcal{P}(\mathbf{y}|\mathbf{x})}_{\mathcal{J}^{\mathrm{obs}}(\mathbf{x})+\mathrm{const}} \underbrace{+\ln\mathcal{P}(\mathbf{y})}_{\mathrm{const}} . \tag{8}$$

The minimization function has two terms: the first one ($\mathcal{J}^{\mathrm{B}}$) comes from the negative logarithm of the background probability density, while the second one ($\mathcal{J}^{\mathrm{obs}}$) comes from the negative logarithm of the observation error probability density. Some scaling factors of the probability densities are usually left out as they give a constant component of the cost function and do not affect the minimization. The third term ($-\ln\mathcal{P}(\mathbf{y})$) does not depend on $\mathbf{x}$ and can also be left out of the minimization function. Under the assumption that the background errors are normally distributed, and after leaving out constant terms, we have that

$$\mathcal{J}^{\mathrm{B}}(\mathbf{x}) = \frac{1}{2}\left(\mathbf{x}-\mathbf{x}^{\mathrm{B}}\right)^T\mathbf{B}^{-1}\left(\mathbf{x}-\mathbf{x}^{\mathrm{B}}\right) . \tag{9}$$

Similarly, under the assumption that observation errors are normally distributed and independent (6), and after leaving out the constant terms,

$$\mathcal{J}^{\mathrm{obs}}(\mathbf{x}) = \sum_{i=1}^{N}\mathcal{J}_i^{\mathrm{obs}}(\mathbf{x}); \quad \mathcal{J}_i^{\mathrm{obs}}(\mathbf{x}) = \frac{1}{2}(\mathcal{H}(\mathbf{x}_i)-\mathbf{y}_i)^T\mathbf{R}_i(\mathcal{H}(\mathbf{x}_i)-\mathbf{y}_i) . \tag{10}$$

Because observation errors are independent each set of observations $\mathbf{y}_i$ at time t_i brings its own contribution $\mathcal{J}_i^{\mathrm{obs}}$ to the total cost function.

2.2 Four dimensional variational (4D-Var) data assimilation

In strongly-constrained 4D-Var data assimilation all observations (2) at all times $t_1,\cdots,t_N$ are simultaneously considered. The control parameters are the initial conditions $\mathbf{x}_0$; they uniquely determine the state of the system at all future times via the model equation (1). The background state is the prior value of the initial conditions $\mathbf{x}_0^{\mathrm{B}}$.

Given the background value of the initial state $\mathbf{x}_0^{\mathrm{B}}$, the covariance of the initial background errors $\mathbf{B}_0$, the observations $\mathbf{y}_i$ and the corresponding observation error covariances $\mathbf{R}_i$, $i = 1, \cdots, N$, the 4D-Var problem looks for the maximum likelihood estimate $\mathbf{x}_0^{\mathrm{A}}$ of the true initial conditions by solving the optimization problem (7). Combining (8), (9), and (10) leads to the 4D-Var cost function:

$$\mathcal{J}(\mathbf{x}_0) = \frac{1}{2}\left(\mathbf{x}_0 - \mathbf{x}_0^{\mathrm{B}}\right)^T \mathbf{B}_0^{-1}\left(\mathbf{x}_0 - \mathbf{x}_0^{\mathrm{B}}\right) + \frac{1}{2}\sum_{i=1}^{N}\left(\mathcal{H}(\mathbf{x}_i) - \mathbf{y}_i\right)^T \mathbf{R}_i^{-1}\left(\mathcal{H}(\mathbf{x}_i) - \mathbf{y}_i\right). \tag{11}$$

Note that the departure of the initial conditions from the background is weighted by the inverse background covariance matrix, while the differences between the model predictions $\mathcal{H}(\mathbf{x}_i)$ and observations $\mathbf{y}_i$ are weighted by the inverse observation error covariances.

The 4D-Var analysis is computed as the initial condition which minimizes (11) subject to the model equation constraints (1)

$$\mathbf{x}_0^{\mathrm{A}} = \operatorname{argmin} \mathcal{J}(\mathbf{x}_0) \quad \text{subject to} (1). \tag{12}$$

The model (1) propagates the optimal initial condition (11) forward in time to provide the analysis at future times, $\mathbf{x}_i^{\mathrm{A}} = \mathcal{M}_{t_0 \to t_i} \mathbf{x}_0^{\mathrm{A}}$.

The optimization problem (12) is solved numerically using a gradient-based technique. The gradient of (11) reads

$$\nabla \mathcal{J}(\mathbf{x}_0) = \mathbf{B}_0^{-1}\left(\mathbf{x}_0 - \mathbf{x}_0^{\mathrm{B}}\right) + \sum_{i=1}^{N}\left(\frac{\partial \mathbf{x}_i}{\partial \mathbf{x}_0}\right)^T \mathbf{H}_i^T \mathbf{R}_i^{-1}\left(\mathcal{H}(\mathbf{x}_i) - \mathbf{y}_i\right). \tag{13}$$

The 4D-Var gradient requires not only the linearized observation operator $\mathbf{H}_i = \mathcal{H}'(\mathbf{x}_i)$, but also the transposed derivative of future states with respect to the initial conditions. The 4D-Var gradient can be obtained effectively by forcing the adjoint model with observation increments, and running it backwards in time. The construction of an adjoint model requires considerable effort.

3 Information Metrics and Gaussian Probabilities

The 4D-Var data assimilation of the observations $\mathbf{y}$ changes the distribution of errors (uncertainty) in the initial conditions from the background probability density $\mathcal{P}^{\mathrm{B}}(\mathbf{x})$ to the analysis probability density $\mathcal{P}^{\mathrm{A}}(\mathbf{x})$. If the data assimilation is beneficial the uncertainty associated with the new distribution $\mathcal{P}^{\mathrm{A}}$ is smaller than the uncertainty associated with the original distribution $\mathcal{P}^{\mathrm{B}}$.

Roughly speaking, the *information content* of the observations $\mathbf{y}$ is measured by the decrease in uncertainty from before data assimilation ($\mathcal{P}^{\mathrm{B}}$) to after data assimilation ($\mathcal{P}^{\mathrm{A}}$). The information content depends not only on the data ($\mathbf{y}_i$) but also on the data accuracy ($\mathbf{R}_i^{-1}$), on the background uncertainty ($\mathbf{B}_0^{-1}$), and on the model dynamics $\mathcal{M}$.

We are interested to rigorously quantify the information content of observations in 4D-Var. For this we use several information theoretic metrics, which are reviewed below.

3.1 Fisher information matrix

The Fisher information matrix (FIM) (Fisher, 1922) associated with the probability density function $\mathcal{P}(\mathbf{x})$ is defined as

$$\mathcal{F}(\mathcal{P}) = \int_{\mathbb{R}^n} \left[\frac{\partial(-\ln\mathcal{P}(\mathbf{x}))}{\partial \mathbf{x}}\right] \left[\frac{\partial(-\ln\mathcal{P}(\mathbf{x}))}{\partial \mathbf{x}})\right]^T \mathcal{P}(\mathbf{x})\, d\mathbf{x} \quad \in \mathbb{R}^{n\times n}\,. \tag{14}$$

The trace of the FIM offers a measure of the total level of uncertainty associated with the distribution.

Under the assumption that the background errors are normally distributed the Fisher information matrix of the background error probability density $\mathcal{P}^{\mathrm{B}}(\mathbf{x}) = \mathcal{N}(\mathbf{x}_0^{\mathrm{B}}, \mathbf{B}_0)$ is just the inverse of the background error covariance:

$$\mathcal{F}\left(\mathcal{P}^{\mathrm{B}}\right) = \int_{\mathbb{R}^n} \left[\nabla \mathcal{J}^{\mathrm{B}}(\mathbf{x}_0)\right] \left[\nabla \mathcal{J}^{\mathrm{B}}(\mathbf{x}_0)\right]^T \mathcal{P}^{\mathrm{B}}(\mathbf{x}_0)\, d\mathbf{x}_0 = \mathbf{B}_0^{-1}. \tag{15}$$

Here we have used the relation (8) to link the background error probability densities with the background part of the 4D-Var cost function.

Similarly, assuming that the analysis error probability density is Gaussian (4) the analysis Fisher information matrix is

$$\mathcal{F}\left(\mathcal{P}^{\mathrm{A}}\right) = \int_{\mathbb{R}^n} [\nabla \mathcal{J}(\mathbf{x}_0)][\nabla \mathcal{J}(\mathbf{x}_0)]^T \mathcal{P}^{\mathrm{A}}(\mathbf{x}_0)\, d\mathbf{x}_0 = \mathbf{A}_0^{-1}\,. \tag{16}$$

The information content of the observations used in data assimilation can be measured as the trace of the background FIM (total uncertainty in the background) minus the trace of the analysis FIM (total uncertainty in the analysis) (Rodgers, 1998, 2000). In the Gaussian case this reduces to the trace of difference between the analysis and background error covariance matrices

$$\mathcal{I}^{\mathrm{FIM}} = \mathrm{trace}\left(\mathcal{F}\left(\mathcal{P}^{\mathrm{A}}\right)\right) - \mathrm{trace}\left(\mathcal{F}\left(\mathcal{P}^{\mathrm{B}}\right)\right) = \mathrm{trace}\left(\mathbf{A}_0^{-1} - \mathbf{B}_0^{-1}\right). \tag{17}$$

3.2 Shannon information

The entropy associated with a probability density is defined as (Shannon and Weaver, 1949; Bartlett, 1962)

$$\mathcal{H}(\mathcal{P}) = \int_{\mathbb{R}^n} \mathcal{P}(\mathbf{x}) \ln(\mathcal{P}(\mathbf{x}))\, d\mathbf{x}$$

and offers a measure of the *average uncertainty* with which one knows the state $\mathbf{x}$, if the estimation error has a probability density $\mathcal{P}$.

For example, assume that the background error distribution is Gaussian. The entropy of the background probability density is given by the relation (Rodgers, 2000)

$$\mathcal{P}^{\mathrm{B}}(\mathbf{x}) = \mathcal{N}\left(\mathbf{x}_0^{\mathrm{B}}, \mathbf{B}_0\right) \quad \Rightarrow \quad \mathcal{H}\left(\mathcal{P}^{\mathrm{B}}\right) = n \ln\left(\sqrt{2\pi e}\right) + \frac{1}{2}\ln\det(\mathbf{B}_0)\,.$$

In this case, the entropy may be interpreted as a measure of the volume in phase space enclosed by a surface of constant probability.

Using the Bayes rule (5) the entropy of the analysis error probability distribution can be written as

$$\mathcal{H}\left(\mathcal{P}^{\mathrm{A}}\right) = \int \left[\ln \mathcal{P}^{\mathrm{B}}(\mathbf{x}) + \ln \mathcal{P}(\mathbf{y}|\mathbf{x}) - \ln \mathcal{P}(\mathbf{y})\right] \mathcal{P}^{\mathrm{A}}(\mathbf{x}) d\mathbf{x}.$$

The Shannon information content of observations $\mathbf{y}$ used in 4D-Var data assimilation is defined as the decrease in the average uncertainty with which the initial state is known. Specifically, the Shannon information content is given by the difference between the background entropy and the analysis entropy,

$$\mathcal{I}^{\mathrm{Shannon}} = \mathcal{H}\left(\mathcal{P}^{\mathrm{B}}\right) - \mathcal{H}\left(\mathcal{P}^{\mathrm{A}}\right) . \tag{18}$$

Under the assumption that both the background and the analysis error probability densities are Gaussian (4), the Shannon information content of the observations used in data assimilation is

$$\mathcal{I}^{\mathrm{Shannon}} = \frac{1}{2}\ln\det(\mathbf{B}_0) - \frac{1}{2}\ln\det(\mathbf{A}_0) = \frac{1}{2}\ln\det\left(\mathbf{B}_0\mathbf{A}_0^{-1}\right) = \frac{1}{2}\ln\det\left(\mathbf{A}_0^{-1/2}\mathbf{B}_0\mathbf{A}_0^{-1/2}\right). \tag{19}$$

3.3 Degrees of freedom for signal

The Degrees of freedom for signal (DFS) metric for the information content has been previously employed in meteorological data assimilation (Rodgers, 1996; Fisher, 2003; Cardinali et al., 2004; Stewart et al., 2008; Zupanski et al., 2007).

Consider the symmetric matrix square root $\mathbf{B}_0^{1/2}$ of the background covariance; we have that

$$\mathbf{B}_0 = \mathbf{B}_0^{1/2}\mathbf{B}_0^{1/2} , \quad \mathbf{B}_0^{-1} = \mathbf{B}_0^{-1/2}\mathbf{B}_0^{-1/2}.$$

Consider also the orthogonal matrix $\mathbf{Q}$ whose columns are the eigenvectors of the symmetric matrix $\mathbf{B}_0^{-1/2}\mathbf{A}_0\mathbf{B}_0^{-1/2}$

$$\mathbf{Q}^T\left(\mathbf{B}_0^{-1/2}\mathbf{A}_0\mathbf{B}_0^{-1/2}\right)\mathbf{Q} = \Sigma,$$

with Σ a diagonal matrix. The matrix $\mathbf{L} = \mathbf{B}_0^{-1/2}\mathbf{Q}$ has the property that it transforms simultaneously the background and the analysis covariances to diagonal forms (Fisher, 2003) when it is symmetrically applied:

$$\mathbf{L}^T\mathbf{B}_0\mathbf{L} = \mathbf{I}_{n\times n} , \quad \mathbf{L}^T\mathbf{A}_0\mathbf{L} = \Sigma .$$

The diagonal elements of the transformed background error covariance matrix are equal to unity and each corresponds to an individual degree of freedom. The eigenvalues of the transformed matrix Σ, on the other hand, can be interpreted as the relative reduction in variance in each of the n statistically independent directions corresponding to the n components of error in the state vector. The degrees of freedom for signal measures the total reduction in variance and is defined as

$$\mathcal{I}^{\mathrm{DFS}} = \mathrm{trace}\left(\mathbf{I}_{n\times n} - \Sigma\right) = n - \mathrm{trace}\left(\mathbf{B}_0^{-1/2}\mathbf{A}_0\mathbf{B}_0^{-1/2}\right) = n - \mathrm{trace}\left(\mathbf{B}_0^{-1}\mathbf{A}_0\right) \ . \tag{20}$$

The relative reduction in variance $\mathbf{B}_0^{-1}\mathbf{A}_0$ could also be interpreted as the gradient of the analysis in observation space with respect to the observations.

3.4 Relative entropy

The information content of the observations used in data assimilation can also be measured by the relative entropy (RE) of the analysis probability density with respect to the background probability density:

$$\mathcal{I}^{\mathrm{RE}} = \int_{\mathbb{R}^n} \mathcal{P}^{\mathrm{A}}(\mathbf{x}) \ln \frac{\mathcal{P}^{\mathrm{A}}(\mathbf{x})}{\mathcal{P}^{\mathrm{B}}(\mathbf{x})} d\mathbf{x} \ .$$

Under the assumption that both the background and the analysis error probability densities are Gaussian (4), the relative entropy of the analysis over the background is (Xu, 2006):

$$\mathcal{I}^{\mathrm{RE}} = \frac{1}{2}\left(\mathbf{x}_0^{\mathrm{A}} - \mathbf{x}_0^{\mathrm{B}}\right)^T \mathbf{B}_0^{-1}\left(\mathbf{x}_0^{\mathrm{A}} - \mathbf{x}_0^{\mathrm{B}}\right) \tag{21a}$$

$$+\frac{1}{2}\mathrm{trace}\left(\mathbf{B}_0^{-1/2}\mathbf{A}_0\mathbf{B}_0^{-1/2}\right) \tag{21b}$$

$$-\frac{n}{2} \tag{21c}$$

$$+\frac{1}{2}\ln\det\left(\mathbf{B}_0^{1/2}\mathbf{A}_0^{-1}\mathbf{B}_0^{1/2}\right) \ . \tag{21d}$$

The *signal part* of the relative entropy

$$\mathcal{I}^{\mathrm{Signal}} = \frac{1}{2}\left(\mathbf{x}_0^{\mathrm{A}} - \mathbf{x}_0^{\mathrm{B}}\right)^T \mathbf{B}_0^{-1}\left(\mathbf{x}_0^{\mathrm{A}} - \mathbf{x}_0^{\mathrm{B}}\right) \tag{22}$$

measures the reduction of uncertainty due to the change in the best estimate from the background state to the analysis state. The terms (21b), (21c), and (21d) together form the *dispersion part* of the relative entropy.

Comparing (21a)–(21b)–(21c)–(21d) and (19), (20), (22) reveals that

$$\mathcal{I}^{\mathrm{RE}} = \underbrace{\mathcal{I}^{\mathrm{Signal}}}_{(21a)} + \underbrace{\mathcal{I}^{\mathrm{Shannon}}}_{(21d)} - \underbrace{(1/2)\,\mathcal{I}^{\mathrm{DFS}}}_{(21b)-(21c)} \ .$$

Let us have a closer look at the relative entropy between two Gaussian distributions:

$$
\begin{aligned}
\mathcal{I}^{\mathrm{RE}} &= \int_{\mathbb{R}^n} \mathcal{P}^{\mathrm{A}}(\mathbf{x}) \cdot \left(\ln \mathcal{P}^{\mathrm{A}}(\mathbf{x}) - \ln \mathcal{P}^{B}(\mathbf{x}) \right) d\mathbf{x} \\
&= \int_{\mathbb{R}^n} \mathcal{P}^{\mathrm{A}}(\mathbf{x}) \cdot \left(-\frac{1}{2} \ln \det \mathbf{A}_0 - \frac{1}{2} \left(\mathbf{x} - \mathbf{x}_0^{\mathrm{A}} \right)^T \mathbf{A}_0^{-1} \left(\mathbf{x} - \mathbf{x}_0^{\mathrm{A}} \right) \right. \\
&\qquad \left. + \frac{1}{2} \ln \det \mathbf{B}_0 + \frac{1}{2} \left(\mathbf{x} - \mathbf{x}_0^{\mathrm{B}} \right)^T \mathbf{B}_0^{-1} \left(\mathbf{x} - \mathbf{x}_0^{\mathrm{B}} \right) \right) d\mathbf{x} \\
&= \frac{1}{2} \ln \det \mathbf{A}_0^{-1} \mathbf{B}_0 - \frac{n}{2} + \frac{1}{2} \int_{\mathbb{R}^n} \mathcal{P}^{\mathrm{A}}(\mathbf{x}) \cdot \left(\mathbf{x} - \mathbf{x}_0^{\mathrm{B}} \right)^T \mathbf{B}_0^{-1} \left(\mathbf{x} - \mathbf{x}_0^{\mathrm{B}} \right) dx
\end{aligned}
$$

We see that the Shannon part (21d) of the relative entropy comes from the scaling factors of the Gaussian distributions (the difference between the logarithms of the $(2\pi)^{-n/2}$ $(\det \mathbf{B}_0)^{-1/2}$ and $(2\pi)^{-n/2}$ $(\det \mathbf{A}_0)^{-1/2}$ factors). Since 4D-Var cost functions do not account for this scaling we cannot hope to accurately recover the Shannon part of the dispersion just by analyzing the cost function.

The constant term (21c) comes from the integration (averaging) of the exponent of the analysis distribution; this is shown in Appendix A in relation (A2). The signal part (21a) and the DFS part (21b) come from the integration (averaging) of the exponent of the background distribution; this is shown in Appendix A in relation (A3).

The three terms (21a), (21c), and (21b) are represented in the 4D-Var cost function, and we could be estimated from statistics of different parts of the 4D-Var cost function.

4 Estimation of the Data Information Content in the Context of 4D-Var Data Assimilation

We seek to derive a computationally-easy way to estimate the information content of various observations in the context of 4D-Var. The proposed approach is based on an approximate sampling from the posterior error distribution in 4D-Var. Thus, our approach is a hybrid one: ensembles are used to infer the information content of observations used in variational data assimilation.

Sampling from the posterior probability density at t_0 is challenging since this probability density is not explicitly computed by 4D-Var. Approximate sampling can be performed using second order adjoints, and computing a few eigenvectors corresponding to the dominant eigenvalues of the inverse Hessian. An alternative approach is based on a subspace analysis of 4D-Var Cheng et al. (2010). A detailed discussion of sampling strategies is provided in the companion paper (Singh et al., 2012a).

Therefore, we assume that we have the ability to obtain the following sample of initial conditions from the posterior distribution:

$$
\mathbf{x}_0^r \in \mathcal{P}^{\mathrm{A}}(\mathbf{x}_0) \ , \quad r = 1, \cdots, \mathrm{N}_{\mathrm{ens}} \ . \tag{23}
$$

Based on it we can approximate expected values of any functional $f(\mathbf{x})$ with respect to the posterior

density by posterior ensemble averages as follows:

$$\mathbb{E}^{\mathrm{A}}[f(\mathbf{x}_0)] \approx \langle f(\mathbf{x}_0)\rangle^{\mathrm{A}} = \frac{1}{\mathrm{N_{ens}}}\sum_{r=1}^{\mathrm{N_{ens}}} f(\mathbf{x}_0^r) \ . \tag{24}$$

4.1 Estimation of the FIM information content

In the 4D-Var setting a gradient based optimization method is typically employed to minimize the cost function $\mathcal{J}(\mathbf{x})$. The gradients are evaluated by the adjoint model; specifically, the value of the adjoint variable at the initial time equals the gradient of the cost function with respect to the initial state

$$\lambda_0(\mathbf{x}_0) = \nabla_{\mathbf{x}_0}\mathcal{J}(\mathbf{x}_0) \ .$$

The adjoint variable depends on the forward model trajectory about which the linearization is performed. This is indicated explicitly by making the adjoint variable a function of the forward initial condition.

The trace of the analysis FIM (16) can be expressed as:

$$\mathrm{trace}\left(\mathcal{F}\left(\mathcal{P}^{\mathrm{A}}\right)\right) = \int_{\mathbb{R}^n} \mathrm{trace}\left(\lambda_0(\mathbf{x}_0)\lambda_0^T(\mathbf{x}_0)\right)\mathcal{P}^{\mathrm{A}}(\mathbf{x}_0)\,d\mathbf{x}_0 = \mathbb{E}^{\mathrm{A}}\left[\|\lambda_0(\mathbf{x}_0)\|^2\right].$$

The trace of the analysis FIM is the average value of the adjoint variable norm with respect to the analysis distribution. Using the sample of initial conditions (23) the statistical average can be approximated by the ensemble average.

Under the typical assumption that the background probability is Gaussian and using (15) and (17) we obtain the following estimate for the FIM information content of all observations:

$$\mathcal{I}^{\mathrm{FIM}} \approx \left\langle \|\lambda_0(\mathbf{x}_0)\|^2 \right\rangle^{\mathrm{A}} - \mathrm{trace}\left(\mathbf{B}_0^{-1}\right) \ . \tag{25}$$

4.1.1 Computational procedure for estimating the FIM information

After the data assimilation has been performed, one runs the forward *and the adjoint* models $\mathrm{N_{ens}}$ times starting with forward initial conditions sampled from the analysis probability density (23). Each run produces an adjoint gradient, whose norm is computed. The ensemble average of these gradient norms estimates the trace of the analysis FIM.

4.2 Estimation of the DFS information content

In this section we consider the idealized situation detailed in Appendix B. Specifically, we assume that the model is linear (B1), the observation operator is also linear (B2), and both the background errors

and the observation errors are normally distributed. The analysis relies on the properties of random quadratic functionals presented in Appendix A.

Consider running the model with random initial conditions taken from the distribution $\widehat{\mathbf{x}}_0 \in \mathcal{N}(\mu, \mathbf{C})$. Each run results in different values of the 4D-Var cost function; we are interested to understand the information provided by the statistics of the (ensemble of) cost function values.

Note that $\widehat{\mathbf{x}}_0 - \mathbf{x}_0^{\mathrm{B}} \in \mathcal{N}\left(\mu - \mathbf{x}_0^{\mathrm{B}}, \mathbf{C}\right)$. A direct application of (A1a) reveals that the background component of the cost function has the following mean:

$$
\begin{aligned}
\mathcal{J}^{\mathrm{B}}(\widehat{\mathbf{x}}_0) &= \frac{1}{2}\left(\widehat{\mathbf{x}}_0 - \mathbf{x}_0^{\mathrm{B}}\right)^T \mathbf{B}_0^{-1}\left(\widehat{\mathbf{x}}_0 - \mathbf{x}_0^{\mathrm{B}}\right) \\
\mathbb{E}\left[\mathcal{J}^{\mathrm{B}}(\widehat{\mathbf{x}}_0)\right] &= \frac{1}{2}\left(\mu - \mathbf{x}_0^{\mathrm{B}}\right)^T \mathbf{B}_0^{-1}\left(\mu - \mathbf{x}_0^{\mathrm{B}}\right) + \frac{1}{2}\mathrm{trace}\left(\mathbf{B}_0^{-1}\mathbf{C}\right) \\
&= \mathcal{J}^{B}(\mu) + \frac{1}{2}\mathrm{trace}\left(\mathbf{C}^{1/2}\mathbf{B}_0^{-1}\mathbf{C}^{1/2}\right) .
\end{aligned}
$$

Since the dynamics is linear, for a given observation data vector $\mathbf{y}_i$ we have that

$$
\mathbf{H}_i\mathbf{M}_i\widehat{\mathbf{x}}_0 - \mathbf{y}_i \in \mathcal{N}\left(\mathbf{H}_i\mathbf{M}_i\mu - \mathbf{y}_i\,,\, \mathbf{H}_i\mathbf{M}_i\mathbf{C}\mathbf{M}_i^T\mathbf{H}_i^T\right) .
$$

Note that the above relation characterizes only the uncertainty in the initial conditions. The data is given; the same data values $\mathbf{y}_i$ are used for each initial condition $\widehat{\mathbf{x}}_0$.

The observation component of the cost function:

$$
\mathcal{J}^{\mathrm{obs}}(\widehat{\mathbf{x}}_0) = \frac{1}{2}\sum_{i=0}^{N}\left(\mathbf{H}_i\mathbf{M}_i\widehat{\mathbf{x}}_0 - \mathbf{y}_i\right)^T \mathbf{R}_i^{-1}\left(\mathbf{H}_i\mathbf{M}_i\widehat{\mathbf{x}}_0 - \mathbf{y}_i\right)
$$

has the following mean:

$$
\begin{aligned}
\mathbb{E}\left[\mathcal{J}^{\mathrm{obs}}(\widehat{\mathbf{x}}_0)\right] &= \frac{1}{2}\sum_{i=0}^{N}\mathbb{E}\left[\left(\mathbf{H}_i\mathbf{M}_i\widehat{\mathbf{x}}_0 - \mathbf{y}_i\right)^T \mathbf{R}_i^{-1}\left(\mathbf{H}_i\mathbf{M}_i\widehat{\mathbf{x}}_0 - \mathbf{y}_i\right)\right] \\
&= \frac{1}{2}\sum_{i=0}^{N}\left(\mathbf{H}_i\mathbf{M}_i\mu - \mathbf{y}_i\right)^T \mathbf{R}_i^{-1}\left(\mathbf{H}_i\mathbf{M}_i\mu - \mathbf{y}_i\right) \\
&\quad + \frac{1}{2}\sum_{i=0}^{N}\mathrm{trace}\left(\mathbf{R}_i^{-1}\mathbf{H}_i\mathbf{M}_i\mathbf{C}\mathbf{M}_i^T\mathbf{H}_i^T\right) \\
&= \mathcal{J}^{\mathrm{obs}}(\mu) + \frac{1}{2}\sum_{i=0}^{N}\mathrm{trace}\left(\mathbf{C}^{1/2}\mathbf{M}_i^T\mathbf{H}_i^T\mathbf{R}_i^{-1}\mathbf{H}_i\mathbf{M}_i\mathbf{C}^{1/2}\right) \\
&= \mathcal{J}^{\mathrm{obs}}(\mu) + \frac{1}{2}\mathrm{trace}\left(\mathbf{C}^{1/2}\left(\sum_{i=0}^{N}\mathbf{M}_i^T\mathbf{H}_i^T\mathbf{R}_i^{-1}\mathbf{H}_i\mathbf{M}_i\right)\mathbf{C}^{1/2}\right) \\
&= \mathcal{J}^{\mathrm{obs}}(\mu) + \frac{1}{2}\mathrm{trace}\left(\mathbf{C}^{1/2}\left(\mathbf{A}_0^{-1} - \mathbf{B}_0^{-1}\right)\mathbf{C}^{1/2}\right)
\end{aligned}
$$

Putting the two formulas together results in

$$\mathbb{E}[\mathcal{J}(\widehat{\mathbf{x}}_0)] - \mathcal{J}(\mu) = \frac{1}{2}\,\mathrm{trace}\left(\mathbf{C}^{1/2}\mathbf{A}_0^{-1}\mathbf{C}^{1/2}\right) . \tag{26}$$

4.2.1 Sampling independent variables

Recall that in the Gaussian case the Fisher information matrix (FIM) is just the inverse of the covariance. Let $\mathbf{C} = \sigma^2\mathbf{I}$. Then the total reduction in uncertainty is given by the trace of the difference between the analysis and the background FIMs:

$$\mathbb{E}\left[\mathcal{J}^{\mathrm{obs}}(\widehat{\mathbf{x}}_0)\right] - \mathcal{J}^{\mathrm{obs}}(\mu) = \frac{\sigma^2}{2}\,\mathrm{trace}\left(\mathbf{A}_0^{-1} - \mathbf{B}_0^{-1}\right) .$$

Consequently the FIM information content of all observations $\mathbf{y}_1 \cdots \mathbf{y}_N$ is

$$\mathcal{I}^{\mathrm{FIM}}_{\mathbf{y}_1\cdots\mathbf{y}_N} = \frac{2}{\sigma^2}\left(\mathbb{E}\left[\mathcal{J}^{\mathrm{obs}}(\widehat{\mathbf{x}}_0)\right] - \mathcal{J}^{\mathrm{obs}}(\mu)\right) .$$

The contribution of the observations $\mathbf{y}_i$ taken at time t_i to the decrease of the trace of FIM, i.e., the FIM information content of $\mathbf{y}_i$ is:

$$\begin{aligned}
\mathcal{I}^{\mathrm{FIM}}_{\mathbf{y}_i} &= \frac{2}{\sigma^2}\left(\mathbb{E}\left[\mathcal{J}_i^{\mathrm{obs}}(\widehat{\mathbf{x}}_0)\right] - \mathcal{J}_i^{\mathrm{obs}}(\mu)\right) \\
&= \frac{1}{\sigma^2}\,\mathbb{E}\left[\left(\mathbf{H}_i\mathbf{M}_i\widehat{\mathbf{x}}_0 - \mathbf{y}_i\right)^T\mathbf{R}_i^{-1}\left(\mathbf{H}_i\mathbf{M}_i\widehat{\mathbf{x}}_0 - \mathbf{y}_i\right)\right] \\
&\quad -\frac{1}{\sigma^2}\left(\mathbf{H}_i\mathbf{M}_i\mu - \mathbf{y}_i\right)^T\mathbf{R}_i^{-1}\left(\mathbf{H}_i\mathbf{M}_i\mu - \mathbf{y}_i\right) .
\end{aligned}$$

While in the linear case this expression does not depend on μ, in the nonlinear case we can take $\mu = \mathbf{x}_0^{\mathrm{A}}$ (after the analysis to assess the impact the observation *had* on the FIM) and $\mu = \mathbf{x}_0^{\mathrm{B}}$ (before the analysis to assess the impact the observation *will have* on the FIM).

4.2.2 Sampling from the analysis distribution

A sample $\widehat{\mathbf{x}}_0 \in \mathcal{N}\left(\mathbf{x}_0^{\mathrm{A}}, \mathbf{A}_0\right)$ from the posterior distribution leads to

$$\begin{aligned}
\mathbb{E}^{\mathrm{A}}\left[\mathcal{J}^{B}(\widehat{\mathbf{x}}_0)\right] &= \mathcal{J}^{B}(\mathbf{x}_0^{\mathrm{A}}) + \frac{1}{2}\mathrm{trace}\left(\mathbf{A}_0^{1/2}\mathbf{B}_0^{-1}\mathbf{A}_0^{1/2}\right) \\
\mathbb{E}^{\mathrm{A}}\left[\mathcal{J}^{\mathrm{obs}}(\widehat{\mathbf{x}}_0)\right] &= \mathcal{J}^{\mathrm{obs}}\left(\mathbf{x}_0^{\mathrm{A}}\right) + \frac{1}{2}\mathrm{trace}\left(\mathbf{A}_0^{1/2}\left(\mathbf{A}_0^{-1} - \mathbf{B}_0^{-1}\right)\mathbf{A}_0^{1/2}\right) \\
&= \mathcal{J}^{\mathrm{obs}}\left(\mathbf{x}_0^{\mathrm{A}}\right) + \frac{n}{2} - \frac{1}{2}\mathrm{trace}\left(\mathbf{A}_0^{1/2}\mathbf{B}_0^{-1}\mathbf{A}_0^{1/2}\right) \\
\mathbb{E}^{\mathrm{A}}\left[\mathcal{J}(\widehat{\mathbf{x}}_0)\right] &= \mathcal{J}(\mathbf{x}_0^{\mathrm{A}}) + \frac{n}{2} .
\end{aligned}$$

The signal part of the relative entropy (21a) is given by $\mathcal{J}^B(x_0^A)$. Attributing the contribution of each observation to the signal part of the entropy is more involved.

We have the following estimate of the DFS information content (21b) of all observations $\mathbf{y}_1 \cdots \mathbf{y}_N$:

$$\mathcal{I}^{\mathrm{DFS}}_{\mathbf{y}_1 \cdots \mathbf{y}_N} = n - \mathrm{trace}\left(\mathbf{A}_0^{1/2}\mathbf{B}^{-1}\mathbf{A}_0^{1/2}\right) = 2\mathbb{E}^{\mathrm{A}}\left[\mathcal{J}^{\mathrm{obs}}(\widehat{\mathbf{x}}_0)\right] - 2\mathcal{J}^{\mathrm{obs}}\left(\mathbf{x}_0^{\mathrm{A}}\right) . \tag{27}$$

This method allows to account for the contribution of each observation $\mathbf{y}_i$ to the DFS information as follows:

$$\begin{aligned}
\mathcal{I}^{\mathrm{DFS}}_{\mathbf{y}_i} &= 2\mathbb{E}^{\mathrm{A}}\left[\mathcal{J}_i^{\mathrm{obs}}(\widehat{\mathbf{x}}_0)\right] - 2\mathcal{J}_i^{\mathrm{obs}}\left(\mathbf{x}_0^{\mathrm{A}}\right) \\
&= \mathbb{E}^{\mathrm{A}}\left[(\mathcal{H}_i(\widehat{\mathbf{x}}_i) - \mathbf{y}_i)^T \mathbf{R}_i^{-1}(\mathcal{H}_i(\widehat{\mathbf{x}}_i) - \mathbf{y}_i)\right] \\
&\quad - \left(\mathcal{H}_i\left(\mathbf{x}_0^{\mathrm{A}}\right) - \mathbf{y}_i\right)^T \mathbf{R}_i^{-1}\left(\mathcal{H}_i\left(\mathbf{x}_0^{\mathrm{A}}\right) - \mathbf{y}_i\right)
\end{aligned}$$

For nonlinear models this relation holds within some approximation margin.

In practice the posterior expected value is replaced by the ensemble expected value

$$\mathcal{I}^{\mathrm{DFS}}_{\mathbf{y}_i} \approx 2\left\langle \mathcal{J}_i^{\mathrm{obs}}(\widehat{\mathbf{x}}_0)\right\rangle^{\mathrm{A}} - 2\mathcal{J}_i^{\mathrm{obs}}\left(\mathbf{x}_0^{\mathrm{A}}\right) . \tag{28}$$

4.2.3 Sampling from the background distribution

A sample $\widehat{\mathbf{x}}_0 \in \mathcal{N}\left(\mathbf{x}_0^{\mathrm{B}}, \mathbf{B}_0\right)$ from the background distribution leads to

$$\begin{aligned}
\mathbb{E}^{\mathrm{B}}\left[\mathcal{J}^B(\widehat{\mathbf{x}}_0)\right] &= \mathcal{J}^B(\mathbf{x}_0^{\mathrm{B}}) + \frac{n}{2} \\
\mathbb{E}^{\mathrm{B}}\left[\mathcal{J}^{\mathrm{obs}}(\widehat{\mathbf{x}}_0)\right] &= \mathcal{J}^{\mathrm{obs}}\left(\mathbf{x}_0^{\mathrm{B}}\right) + \frac{1}{2}\,\mathrm{trace}\left(\mathbf{B}_0^{1/2}\left(\mathbf{A}_0^{-1} - \mathbf{B}_0^{-1}\right)\mathbf{B}_0^{1/2}\right) \\
\mathbb{E}^{\mathrm{B}}\left[\mathcal{J}(\widehat{\mathbf{x}}_0)\right] &= \mathcal{J}\left(\mathbf{x}_0^{B}\right) + \frac{1}{2}\,\mathrm{trace}\left(\mathbf{B}_0^{1/2}\mathbf{A}_0^{-1}\mathbf{B}_0^{1/2}\right) .
\end{aligned}$$

4.2.4 Computational procedure for estimating the DFS information

After the data assimilation has been performed, one runs the forward model $\mathrm{N_{ens}}$ times. The initial conditions are sampled from the analysis distribution (23) (or from another distribution, e.g., diagonal, to obtain different statistics). An additional run is performed starting from the analysis initial conditions. During each run one records all individual contributions $\mathcal{J}_i^{\mathrm{obs}}$ of all observations $\mathbf{y}_i$ to the cost function. This data is post-processed according to (28). The ensemble average of the contributions $\mathcal{J}_i^{\mathrm{obs}}$, minus the contribution obtained from the analysis run, estimates (half of) the DFS information content of the data $\mathbf{y}_i$.

4.3 Estimation of the RE information content

The relative entropy (RE) information content of all observations $\mathbf{y}_1 \cdots \mathbf{y}_N$ is measured by the relative entropy of the posterior probability density $\mathcal{P}^{\mathrm{A}}$ over the background probability density $\mathcal{P}^{\mathrm{B}}$

$$\begin{aligned}
\mathcal{I}^{\mathrm{RE}}_{\mathbf{y}_1 \cdots \mathbf{y}_N} &= \int_{\mathbb{R}^n} \mathcal{P}^{\mathrm{A}}(\mathbf{x}) \cdot \ln \frac{\mathcal{P}^{\mathrm{A}}(\mathbf{x})}{\mathcal{P}^{\mathrm{B}}(\mathbf{x})} d\mathbf{x} \\
&= \int_{\mathbb{R}^n} \mathcal{P}^{\mathrm{A}}(\mathbf{x}) \cdot \ln \frac{\mathcal{P}(\mathbf{y}|\mathbf{x})}{\mathcal{P}(\mathbf{y})} d\mathbf{x} \\
&= \int_{\mathbb{R}^n} \mathcal{P}^{\mathrm{A}}(\mathbf{x}) \cdot \Big(\ln \mathcal{P}(\mathbf{y}|\mathbf{x}) - \ln \mathcal{P}(\mathbf{y}) \Big) d\mathbf{x} \\
&= \mathbb{E}^{\mathrm{A}} [\ln \mathcal{P}(\mathbf{y}|\mathbf{x})] - \ln \mathcal{P}(\mathbf{y}) \\
&= \mathrm{const} - \mathbb{E}^{\mathrm{A}} \Big[\mathcal{J}^{\mathrm{obs}}(\mathbf{x}) \Big]
\end{aligned}$$

where we have made use of Bayes rule (5) to derive the second relation, and of (8) to derive the last equation. The marginal distribution of observations $\mathbf{y}$ does not depend on $\mathbf{x}$ and its expected value is a constant.

Assuming we can sample the posterior distribution this expected value can be approximated by the ensemble mean. The RE information content of all observations is estimated as

$$\mathcal{I}^{\mathrm{RE}}_{\mathbf{y}_1 \cdots \mathbf{y}_N} \approx \mathrm{const} - \Big\langle \mathcal{J}^{\mathrm{obs}}(\mathbf{x}) \Big\rangle^{\mathrm{A}} . \tag{29}$$

The relative entropy information content is larger when the 4D-Var process decreases more the observation part of the cost function. In other words, the lower the mismatch between model predictions and observations after assimilation the higher the relative entropy information content of observations is.

The RE information content of the particular observation $\mathbf{y}_i$ can be quantified as follows. Data assimilation using all observations $\mathbf{y}_1 \cdots \mathbf{y}_N$ results in a posterior probability density $\mathcal{P}^{\mathrm{A}}(\mathbf{x})$. Data assimilation using all observations except $\mathbf{y}_i$ results in another posterior probability density $\mathcal{P}^{\mathrm{A}}_{-i}(\mathbf{x})$. The RE information contribution of data $\mathbf{y}_i$ is measured by the relative entropy of the full-data posterior probability density $\mathcal{P}^{\mathrm{A}}$ over the partial-data posterior density $\mathcal{P}^{\mathrm{A}}_{-i}$. If the observation errors at different times are independent it can be shown that

$$\mathcal{I}^{\mathrm{RE}}_{\mathbf{y}_i} = \int_{\mathbb{R}^n} \mathcal{P}^{\mathrm{A}}(\mathbf{x}) \cdot \ln \frac{\mathcal{P}^{\mathrm{A}}(\mathbf{x})}{\mathcal{P}^{\mathrm{A}}_{-i}(\mathbf{x})} d\mathbf{x} = \mathrm{const}_i - \mathbb{E}^{\mathrm{A}} \Big[\mathcal{J}^{\mathrm{obs}}_i(\mathbf{x}) \Big] \approx \mathrm{const}_i - \Big\langle \mathcal{J}^{\mathrm{obs}}_i(\mathbf{x}) \Big\rangle^{\mathrm{A}} . \tag{30}$$

The constant comes from the marginal probability of the observation $\mathbf{y}_i$ and is different for each data point. Therefore it is difficult to apportion the information gain to individual observations using this metric.

An alternative, more computationally involved approach would be to repeat the data assimilation without the data point $\mathbf{y}_i$, and to build another ensemble drawn from $\mathcal{P}^{\mathrm{A}}_{-i}(\mathbf{x})$. For each data assimilation

experiment one computes the total RE information content (29). The information gain due to the data $\mathbf{y}_i$ is the measured by

$$\mathcal{I}^{\mathrm{RE}}_{\mathbf{y}_i} = \mathcal{I}^{\mathrm{RE}}_{\mathbf{y}_1 \cdots \mathbf{y}_N} - \mathcal{I}^{\mathrm{RE}}_{\mathbf{y}_1 \cdots y_{i-1}, y_{i+1} \cdots \mathbf{y}_N} \ . \tag{31}$$

4.3.1 Computational procedure for estimating the RE information

The computational procedure is similar to the one for the DFS information presented in Section 4.2.4. An ensemble of models is run with the initial conditions sampled from the analysis distribution (23). The ensemble average of the observation part $\mathcal{J}^{\mathrm{obs}}$ of the cost function estimates the RE information content of all observations (29), modulo a constant. This procedure can be repeated for different data assimilation scenarios, where individual data points are being withheld; the difference between the resulting metrics estimates the RE information content of the withheld data.

4.4 Estimation of the Shannon information content

We have seen that the Shannon information is related to the scaling of the Gaussian probability densities. This information is ignored by the 4D-Var cost function. Therefore, we cannot expect to obtain accurate estimates of the Shannon information content by mining the cost function information.

A (very) rough approximation can be obtained using the eigenvalues of the ensemble covariance matrices, as follows. Consider a set of perturbations drawn from the background ensemble, and a set of perturbations drawn from the analysis ensemble; in matrix notation

$$\Delta \mathbf{x}^{\mathrm{B}}_0 \in \mathbf{R}^{n \times \mathrm{N_{ens}}} \ ; \quad \Delta \mathbf{x}^{\mathrm{A}}_0 \in \mathbf{R}^{n \times \mathrm{N_{ens}}} \ ; \quad \mathrm{N_{ens}} \ll n \ .$$

The error covariance matrices are approximated by the ensemble covariance

$$\mathbf{B}_0 \approx \frac{1}{(\mathrm{N_{ens}} - 1)} \cdot \left(\Delta \mathbf{x}^{\mathrm{B}}_0 \right)^T \cdot \Delta \mathbf{x}^{\mathrm{B}}_0 \ ; \quad \mathbf{A}_0 \approx \frac{1}{(\mathrm{N_{ens}} - 1)} \cdot \left(\Delta \mathbf{x}^{\mathrm{A}}_0 \right)^T \cdot \Delta \mathbf{x}^{\mathrm{A}}_0 \ . \tag{32}$$

Denote the nonzero eigenvalues of the two ensemble covariance matrices by λ^B_i and λ^A_i respectively, $i = 1,2,\ldots,\mathrm{N_{ens}}$. The nonzero eigenvalues can be efficiently computed by solving small $\mathrm{N_{ens}} \times \mathrm{N_{ens}}$ eigenvalue problems since

$$\Lambda = \mathrm{eig} \underbrace{\left(\Delta \mathbf{x} \cdot \Delta \mathbf{x}^T \right)}_{n \times n} \in \mathbf{R}^n \ , \quad \lambda = \mathrm{eig} \underbrace{\left(\Delta \mathbf{x}^T \cdot \Delta \mathbf{x} \right)}_{\mathrm{N_{ens}} \times \mathrm{N_{ens}}} \in \mathbf{R}^{\mathrm{N_{ens}}} \quad \Rightarrow \quad \Lambda_i = \lambda_i \ , \ i = 1, \cdots, \mathrm{N_{ens}} \ . \tag{33}$$

An estimate of the Shannon information content (21d) can be given in terms of eigenvalues as follows:

$$\frac{1}{2} \ln \det \mathbf{B}_0 \mathbf{A}_0^{-1} = \frac{1}{2} \ln \prod_{i=1}^{\mathrm{N_{ens}}} \left(\frac{\lambda^{\mathrm{B}}_i}{\lambda^{\mathrm{A}}_i} \right) = \frac{1}{2} \sum_{i=1}^{\mathrm{N_{ens}}} \ln \left(\frac{\lambda^{\mathrm{B}}_i}{\lambda^{\mathrm{A}}_i} \right) \ . \tag{34}$$

Similarly, the part (21b) of the DFS metric can be estimated by

$$\frac{1}{2}\text{trace}\left(\mathbf{B}_0^{-1/2}\mathbf{A}_0\mathbf{B}_0^{-1/2}\right) = \frac{1}{2}\sum_{i=1}^{\text{N}_{\text{ens}}}\left(\frac{\lambda_i^{\text{A}}}{\lambda_i^{\text{B}}}\right) . \tag{35}$$

4.4.1 Computational procedure for estimating the Shannon information

One constructs two ensembles of initial conditions, one from the background distribution, and one from the analysis distribution. The nonzero eigenvalues of the corresponding ensemble covariances are computed using (33). These eigenvalues are used to estimate the Shannon information content via (34) and the DFS information content via (35). The computational procedure is direct - no additional model runs are necessary. However, for a small number of ensemble members, the ensemble covariance eigenvalues may poorly represent the eigenvalues of the true covariances. In this case the resulting estimates of the Shannon or DFS information content are expected to be inaccurate.

4.5 Estimation of the Signal information content

In this section we assume a linear system with linear observation operators and Gaussian uncertainties as discussed in Appendix B. The analysis state obtained using all the available information is $\mathbf{x}_0^{\text{A}}$. Consider one particular observation $\mathbf{y}_\ell$, remove it from the set of data, and repeat the data assimilation. Let $\mathbf{x}_0^{\text{C}}$ be the analysis state when the data assimilation is carried out *without the observation* $\mathbf{y}_\ell$.

We use the notation of Appendix B. Furthermore, denote the contribution of observation ℓ to the right hand side and to the 4D-Var system matrix (B4) by

$$\mathbf{b}_\ell = \mathbf{M}_\ell^T\mathbf{H}_\ell^T\mathbf{R}_\ell^{-1}\left(\mathbf{y}_\ell - \mathbf{H}_\ell\mathbf{M}_\ell\mathbf{x}_0^{\text{B}}\right) , \quad \mathbf{D}_\ell = \mathbf{M}_\ell^T\mathbf{H}_\ell^T\mathbf{R}_\ell^{-1}\mathbf{H}_\ell\mathbf{M}_\ell .$$

Following equation (B4) the two 4D-Var problems have the following solutions:

$$\mathbf{A}_0^{-1}\cdot\left(\mathbf{x}_0^{\text{A}} - \mathbf{x}_0^{\text{B}}\right) = \mathbf{b} , \quad \left(\mathbf{A}_0^{-1} - \mathbf{D}_\ell\right)\cdot\left(\mathbf{x}_0^{\text{C}} - \mathbf{x}_0^{\text{B}}\right) = \mathbf{b} - \mathbf{b}_\ell .$$

We assume a case where there are many observations such that the contribution of b_ℓ to the total right hand side vector is relatively small, $\mathbf{b} - \mathbf{b}_\ell \approx \mathbf{b}$, and the contribution of D_ℓ to the total inverse covariance is relatively small, $\mathbf{A}_0^{-1} - \mathbf{D}_\ell \approx \mathbf{A}_0^{-1}$. The following approximations are obtained:

$$\mathbf{A}_0^{-1}\cdot\left(\mathbf{x}_0^{\text{C}} - \mathbf{x}_0^{\text{B}}\right) \approx \mathbf{b} - \mathbf{b}_\ell , \quad \mathbf{A}_0^{-1}\cdot\left(\mathbf{x}_0^{\text{A}} - \mathbf{x}_0^{\text{C}}\right) \approx \mathbf{b}_\ell .$$

The difference in the signal part due to the assimilation of observation $\mathbf{y}_\ell$ is

$$
\begin{aligned}
\mathcal{I}_{\mathbf{y}_\ell}^{\text{Signal}} &= \frac{1}{2}\left(\mathbf{x}_0^{\text{A}}-\mathbf{x}_0^{\text{B}}\right)^T\mathbf{B}_0^{-1}\left(\mathbf{x}_0^{\text{A}}-\mathbf{x}_0^{\text{B}}\right)-\frac{1}{2}\left(\mathbf{x}_0^{\text{C}}-\mathbf{x}_0^{\text{B}}\right)^T\mathbf{B}_0^{-1}\left(\mathbf{x}_0^{\text{C}}-\mathbf{x}_0^{\text{B}}\right)\\
&= \frac{1}{2}\left(\mathbf{x}_0^{\text{A}}-\mathbf{x}_0^{\text{B}}\right)^T\mathbf{B}_0^{-1}\left(\mathbf{x}_0^{\text{A}}-\mathbf{x}_0^{\text{B}}\right)-\frac{1}{2}\left(\mathbf{x}_0^{\text{C}}-\mathbf{x}_0^{\text{B}}\right)^T\mathbf{B}_0^{-1}\left(\mathbf{x}_0^{\text{A}}-\mathbf{x}_0^{\text{B}}\right)\\
&\quad+\frac{1}{2}\left(\mathbf{x}_0^{\text{A}}-\mathbf{x}_0^{\text{B}}\right)^T\mathbf{B}_0^{-1}\left(\mathbf{x}_0^{\text{C}}-\mathbf{x}_0^{\text{B}}\right)-\frac{1}{2}\left(\mathbf{x}_0^{\text{C}}-\mathbf{x}_0^{\text{B}}\right)^T\mathbf{B}_0^{-1}\left(\mathbf{x}_0^{\text{C}}-\mathbf{x}_0^{\text{B}}\right)\\
&= \frac{1}{2}\left(\mathbf{x}_0^{\text{A}}-\mathbf{x}_0^{\text{C}}\right)^T\mathbf{B}_0^{-1}\left(\mathbf{x}_0^{\text{A}}-\mathbf{x}_0^{\text{B}}\right)+\frac{1}{2}\left(\mathbf{x}_0^{\text{A}}-\mathbf{x}_0^{\text{C}}\right)^T\mathbf{B}_0^{-1}\left(\mathbf{x}_0^{\text{C}}-\mathbf{x}_0^{\text{B}}\right)\\
&= \frac{1}{2}\left(\mathbf{A}_0^{-1}(\mathbf{x}_0^{\text{A}}-\mathbf{x}_0^{\text{C}})\right)^T\mathbf{A}_0\mathbf{B}_0^{-1}\mathbf{A}_0\left(\mathbf{A}_0^{-1}(\mathbf{x}_0^{\text{A}}-\mathbf{x}_0^{\text{B}})+\mathbf{A}_0^{-1}(\mathbf{x}_0^{\text{C}}-\mathbf{x}_0^{\text{B}})\right)\\
&\approx \frac{1}{2}(\mathbf{b}_\ell)^T\mathbf{A}_0\mathbf{B}_0^{-1}\mathbf{A}_0(2\mathbf{b}-\mathbf{b}_\ell)\\
&\approx \mathbf{b}_\ell^T\mathbf{A}_0\mathbf{B}_0^{-1}\mathbf{A}_0\mathbf{b}\\
&= \left(\mathbf{y}_\ell-\mathbf{H}_\ell\mathbf{M}_\ell\mathbf{x}_0^{\text{B}}\right)^T\mathbf{R}_\ell^{-1}\mathbf{H}_\ell\mathbf{M}_\ell\mathbf{A}_0\mathbf{B}_0^{-1}\left(\mathbf{x}_0^{\text{A}}-\mathbf{x}_0^{\text{B}}\right)\ .
\end{aligned}
$$

Let

$$
\begin{aligned}
&\widetilde{\mathbf{x}}_0^A=\mathbf{A}_0\mathbf{B}_0^{-1}\mathbf{x}_0^{\text{A}},\quad \widetilde{\mathbf{x}}_0^B=\mathbf{A}_0\mathbf{B}_0^{-1}\mathbf{x}_0^{\text{B}}, \qquad (36)\\
&\mathbf{H}_\ell\mathbf{M}_\ell\mathbf{A}_0\mathbf{B}_0^{-1}\left(\mathbf{x}_0^{\text{A}}-\mathbf{x}_0^{\text{B}}\right)\approx\mathbf{H}_\ell\widetilde{\mathbf{x}}_\ell^A-\mathbf{H}_\ell\widetilde{\mathbf{x}}_\ell^B\ .
\end{aligned}
$$

The contribution of measurement $\mathbf{y}_\ell$ to the signal information can therefore be approximated as:

$$
\begin{aligned}
\mathcal{I}_{\mathbf{y}_\ell}^{\text{Signal}} &\approx \left(\mathbf{y}_\ell-\mathcal{H}_\ell\left(\mathbf{x}_\ell^B\right)\right)^T\mathbf{R}_\ell^{-1}\left(\mathcal{H}_\ell\left(\widetilde{\mathbf{x}}_\ell^A\right)-\mathcal{H}_\ell\left(\widetilde{\mathbf{x}}_\ell^B\right)\right) \qquad (37a)\\
&\approx \left(\mathbf{y}_\ell-\mathcal{H}_\ell\left(\mathbf{x}_\ell^B\right)\right)^T\mathbf{R}_\ell^{-1}\left(\mathcal{H}_\ell\left(\mathbf{x}_\ell^A\right)-\mathcal{H}_\ell\left(\mathbf{x}_\ell^B\right)\right) \qquad (37b)
\end{aligned}
$$

where the last approximation is rather coarse.

4.5.1 Computational procedure for estimating the Signal information

Two modified initial conditions are computed by (36). (If this is not feasible, the background and the analysis initial conditions can be used, at the price of a larger approximation error). The model is run from the modified analysis and the "synthetic observations" $\mathcal{H}_\ell\left(\widetilde{\mathbf{x}}_\ell^A\right)$ are recorded. The model is run again from the modified background and the "synthetic observations" $\mathcal{H}_\ell\left(\widetilde{\mathbf{x}}_\ell^B\right)$ are also recorded (this run is not necessary if one uses (37b)). Finally, the model is run from the background state, and the estimates (37a) or (37b) are evaluated for each data point $\mathbf{y}_\ell$.

5 Numerical Experiments

In this section we apply the estimation methodology developed in Section 4 to the Lorenz-96 model (Lorenz, 1996), a highly nonlinear test case.

In (Singh et al., 2012b) we report results with a linear test case with Gaussian uncertainties, where the numerical results confirm the accuracy of theoretical estimates. In the companion paper (Singh et al., 2012a) we consider a 4D-Var data assimilation study with a global chemical transport model. The data assimilation experiment focuses on ozone. Ozone is an important constituent of stratosphere which absorbs the high energy UV-B and UV-C rays, thus preventing the disintegration of DNA molecules and supporting the existence of life. However, ozone present in mid to low troposphere is a pollutant, a powerful oxidizing agent leading to destruction of tissues, damaging fibers and creating breathing problems. We estimate the DFS information content of satellite ozone column retrievals at different times.

The Lorenz 96 system (Lorenz, 1996) is described by the following set of equations:

$$\frac{dx_j}{dt} = -x_{j-1}(x_{j-2} - x_{j+1} - x_j) + F\ ,\ j = 1,\ldots n\ , \tag{38}$$

with $n = 40$ and periodic boundary conditions ($x_i = x_{n+i}$ for any i). The forcing term is $F = 8.0$.

We start with the state $x_i(t_{-10}) = 1 + 0.1\mathrm{mod}(i,5)$, $i = 1,\cdots,n$, and integrate it forward for 10 time units to obtain the reference ("true") state at t_0, x_0^{ref}.

A static background covariance matrix $\mathbf{B}_{t_0}$ is constructed as follows. First define the covariance matrix $\widehat{\mathbf{B}}_0$ using

$$\left(\widehat{\mathbf{B}}_0\right)_{ij} = \sigma_i \cdot \sigma_j \cdot \exp\left(-\frac{d^2}{L^2}\right)\ ,\ i,j = 1,\ldots,n\ , \tag{39}$$

where the standard deviation for the state variable i is $\sigma_i = 0.03 x_i^{\mathrm{ref}}(t_0)$, and the correlation distance is set to $L = 4$. We account for the periodic boundary conditions in that $d = \min\{|i-j|, n-|i-j|\}$. The covariance matrix is obtained via

$$\mathbf{B}_0 = \alpha\,\mathbf{I}_{n\times n} + (1-\alpha)\,\widehat{\mathbf{B}}_0\ ,\quad \alpha = 0.1\ .$$

This construction ensures a nonsingular $\mathbf{B}_0$, as required by the 4D-Var algorithm.

The background initial state is obtained from the reference solution, plus a random perturbation consistent with the background error statistics:

$$\mathbf{x}_0^{\mathrm{B}} = \mathbf{x}_0^{\mathrm{ref}} + \mathbf{B}_0^{1/2} \cdot \xi\ ,\quad \xi \in \mathcal{N}(0,1)^n\ . \tag{40}$$

The simulation time interval is $[0, 0.3]$ time units. The reference trajectory is generated using Matlab's ode45 integrator with tight tolerances (relative error tolerance RelTol=1.e-9 and absolute error tolerance

AbsTol=1.e-9). The model consists of a numerical integration using "the original" fourth order Runge-Kutta method with a time step $\Delta t = 0.015$ time units. The model error is truncation error associated with the numerical integration using a relatively large step size.

The vector of observations has $m = 34$ components at each time. The observation operator $\mathcal{H} \in \mathbb{R}^{34\times 40}$ is linear and captures a subset of 30 model states plus 4 linear combinations of model states as follows

$$\mathcal{H}\cdot\mathbf{x} = \left[\underbrace{x_1,x_3,\cdots,x_{19}}_{\text{odd-numbered}} \;,\; \underbrace{x_{21},x_{22}\cdots x_{39},x_{40}}_{\text{all states}} \;,\; \underbrace{\sum_{i=1}^{10} x_i, \sum_{i=1}^{20} x_i, \sum_{i=21}^{40} x_i, \sum_{i=31}^{40} x_i}_{\text{linear combinations}} \right]^T . \tag{41}$$

The odd-numbered states among the first 20 ($\{1,3,5,\cdots,19\}$) and all of the last 20 states ($\{21,22,\cdots,39,40\}$) are directly observed. Also observed are the sum of the first 10 states, the sum of the first 20 states, the sum of the last 20 states, and the sum of the last 10 states. The sums of the last 10 and 20 states are redundant observations which can be recovered from the observations of individual states $\{21,\cdots,40\}$. The sums of the first 10 and 20 states bring additional information about the even-numbered states $\{2,4,\cdots,16,18\}$ which are not directly observed.

Observations are taken every 0.03 time units, i.e., there are $N_{\text{obs}} = 10$ uniformly spaced observation times: $t_k = 0.03k$ (time units) for $k = 1,\cdots,N_{\text{obs}}$. Synthetic observation values are generated as follows. First, the reference trajectory is used to obtain perfect observations $\mathbf{y}_k^{\text{ref}} = \mathcal{H}\cdot\mathbf{x}_k^{\text{ref}}$. The vector of standard deviations of observation errors is taken to be 0.5% of the time-averaged reference observations, $\sigma^{\text{obs}} = 0.005\,(\sum_{k=1}^{N_{\text{obs}}} y_k^{\text{ref}})/N_{\text{obs}}$. The observation errors are assumed to be uncorrelated. The observation covariance is the diagonal matrix $\mathbf{R} = \text{diag}_{i=1\cdots n}\left\{(\sigma_i^{\text{obs}})^{-2}\right\}$ and is constant for all observations times.

Synthetic observations are generated by adding Gaussian noise to the reference observations:

$$\mathbf{y}_k = \mathbf{y}_k^{\text{ref}} + \mathbf{R}^{1/2}\cdot\eta_k \;,\quad \eta_k \in \mathcal{N}(0,1)^m \;,\quad k = 1,\cdots,N_{\text{obs}}\,. \tag{42}$$

Our implementation of 4D-Var makes use of the matlab implementation of L-BFGS algorithm provided in Heinkenschloss (2008). L-BFGS (Zhu et al., 1997) is the de facto "gold standard" of gradient-based optimizers used in data assimilation studies.

An ensemble of 1,000 4D-Var optimization runs is performed. Each run uses a different background state generated according to (40), and a different set of observations generated according to (42). The ensemble of optimized initial states samples the analysis distribution (of initial conditions). Ensembles of runs started from these initial states are used to estimate various information content metrics.

From the ensemble of initial conditions we derive the ensemble covariance matrices $\mathbf{B}_e \approx \mathbf{B}_0$ and $\mathbf{A}_e \approx \mathbf{A}_0$. These are used to compute directly estimates of the information metrics. The information content metrics obtained directly, and via the approximation formulas proposed here, are shown in Table 1. The two computational approaches give very close estimates.

Table 1. Information content metrics for the 4D-Var experiment with the Lorenz-96 system. The values obtained by the estimation formulas are close to those obtained by direct calculations.

	Direct		Estimate	
	Equation	Value	Equation	Value
DFS	$n - \mathrm{trace}\left(\mathbf{A}_e^{1/2} \cdot \mathbf{B}_e^{-1} \cdot \mathbf{A}_e^{1/2}\right)$	3.995e+01	(28)	3.978e+01
Fisher	$\mathrm{trace}\left(\mathbf{A}_e^{-1} - \mathbf{B}_e^{-1}\right)$	2.269e+06	(25)	2.207e+06
Signal	$0.5 \cdot \left(\mathbf{x}_0^{\mathrm{A}} - \mathbf{x}_0^{\mathrm{B}}\right)^T \cdot \mathbf{B}_e^{-1} \cdot \left(\mathbf{x}_0^{\mathrm{A}} - \mathbf{x}_0^{\mathrm{B}}\right)$	1.785e+01	(37a)	1.723e+01
Shannon	$0.5 \cdot (\ln\det \mathbf{B}_e - \ln\det \mathbf{A}_e)$	1.637e+02	(34)	1.631e+02

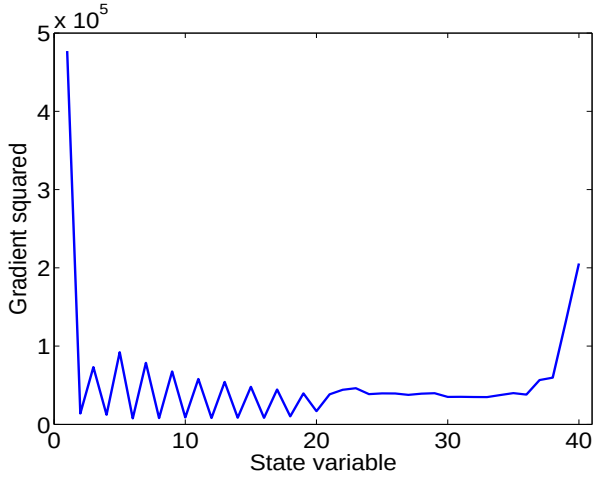


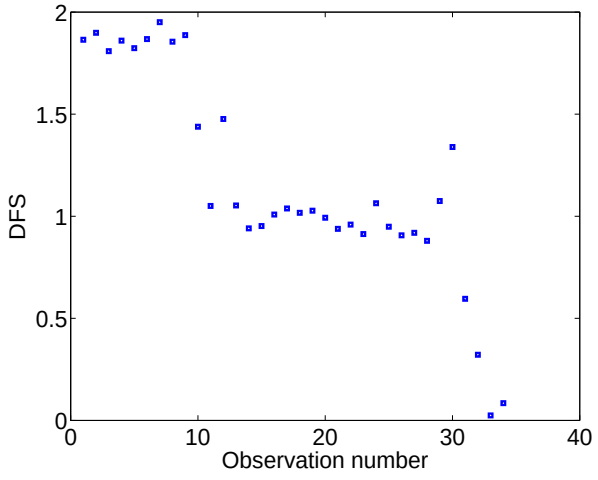
Fig. 1. Estimated Fisher information gain per state variable for the Lorenz-96 model.

The Fisher information is estimated using equation (25). Figure 1 shows each component of the gradient squared $\mathcal{I}_i^{\mathrm{FIM}} \approx \left\langle (\lambda_0)_i^2 \right\rangle^{\mathrm{A}}$. This quantifies the informational benefit that each state $x_i(t_0)$ receives due to data assimilation, as measured by the Fisher information matrix. Among the first twenty states the odd numbered ones (x_1,x_3,$\cdots$,x_{19}) benefit more than the even numbered ones (x_2,x_4,$\cdots$,x_{18}). This correlates well with the structure of the operator (41) which observes directly only the odd numbered states. The last twenty states show about the same information benefit, and this is expected since all of them are directly observed. The end states x_1 and x_{40} show the largest information gain; this cannot be explained based solely on the structure of the observation operator (41).

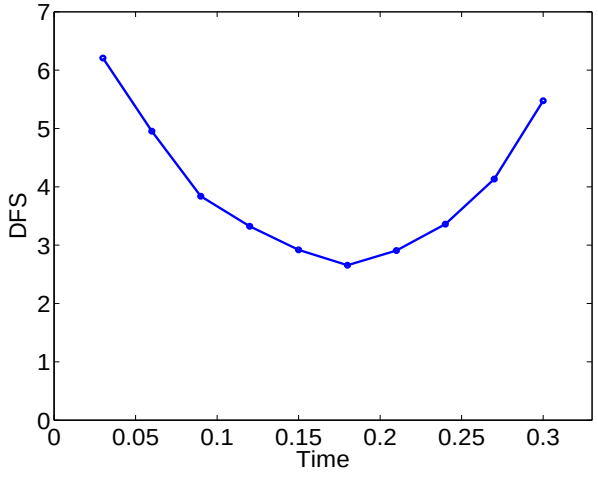
The DFS information is estimated using equation (28). The formula allows to split the DFS information into contributions brought by each observation at each time. Figure 2(a) presents the DFS contributions of each observations (summed up for all observation times). Each of the first 10 observations (of states x_1,x_3,$\cdots$,x_{19}) contributes about two degrees of freedom for signal. We can infer that direct observation of an odd numbered state brings information about its un-observed even numbered neighbors. Each of the next 20 observations (of states x_{21},$\cdots$,x_{40}) contributes a single degree of freedom for signal. This is

expected as each observation in this group measures a single state variable, and all neighboring states are directly observed. Each of the observations 31 (sum of the first 10 states) and 32 (sum of the first 20 states) brings in about half a degree of freedom for signal. Finally, the last two observations (sums of the last 10 and of the last 20 states) bring in almost zero degrees of freedom for signal. This is expected as the information is redundant.

Figure 2(b) presents the DFS contribution of each observation time (summed up for all observations). The time points near the beginning and near the end of the assimilation window bring larger contributions of over 5 degrees of freedom for signal. The points near the middle of the assimilation window have smaller contributions of under 3 degrees of freedom for signal.



(a) DFS per observation



(b) DFS per time point

Fig. 2. Estimated degrees-of-freedom-for-signal information metrics for the Lorenz-96 model.

The signal information is estimated using equation (37a). Figure 3 shows the signal contribution of all observations at different times. The observations near the beginning of the assimilation window contribute the most, while those near the middle of the assimilation window contribute the least.

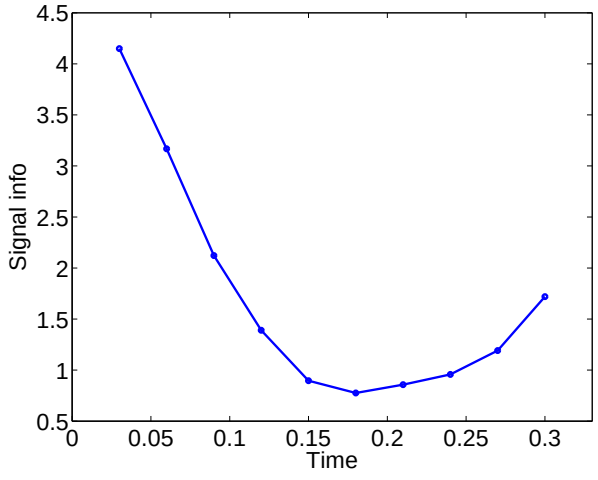


Fig. 3. Estimated signal information contribution per observation time for the Lorenz-96 model.

The estimates of the signal contribution of each observation are not producing relevant results. They are shown in Figure 4. The approximation formula that separates the signal contributions for each observation seems to be too inaccurate.

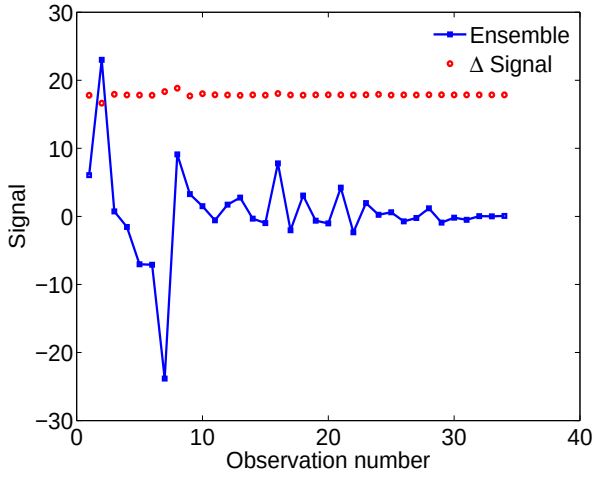


Fig. 4. Estimated signal information contribution per observation for the Lorenz-96 model.

6 Conclusions

This paper discusses a characterization of the information content of observations in the context of four dimensional variational (4D-Var) data assimilation framework. The ability to characterize the usefulness of different data points is important for analyzing the effectiveness of the assimilation system, for data pruning, and for the design of future sensor systems.

Several metrics from information theory are used to quantify the information content of data, including the trace of the Fisher information matrix, the Shannon information, the relative entropy, the signal information, and the degrees of freedom for signal. In the Gaussian case the signal information measures the benefit of data assimilation in terms of adjusting the mean of the distribution. Fisher, Shannon, and DFS all measure the benefit of data assimilation in terms of decreasing the (co-)variance of the error. Relative entropy offers a combination of metrics.

The analysis is carried out under the assumptions that errors have Gaussian distributions and that the model dynamics is linear. The analysis reveals that the information content of observations is intimately related to the statistics of the variational cost function and its gradient. These statistics are obtained with respect to the analysis probability distribution. The theoretical results lead to a new computational procedure to estimate the information content of various observations in the context of 4D-Var. After data assimilation is complete, and ensemble of simulations is run with the initial conditions drawn from the posterior probability distribution. Mean values of the adjoint norms are used to estimate the trace of the Fisher information matrix. The mean value of the observation part of the cost function, minus its value for the analysis, is used to estimate the DFS information content. Scaled dot products between the background innovation and the difference between the background and the analysis innovations provide estimates of the signal information content. The estimates require expected values with respect to the posterior distribution. A detailed discussion on how these can be obtained is given in the companion paper (Singh et al., 2012a).

The information content estimation approach is illustrated on a nonlinear test problem. In the companion paper (Singh et al., 2012a) we report results with a 4D-Var ozone data assimilation study with a global chemical transport model, where we estimate the DFS information content of satellite ozone column retrievals.

The assumptions and approximations made during the analysis and computations impact the accuracy of the information content estimates. While the analysis assumes normal error distributions and a linear dynamics, it is desirable to apply the methodology to nonlinear systems and arbitrary uncertainty distributions. The analysis distribution is not explicitly available, samples are taken from distributions that only approximate the analysis under certain assumptions. Finally, relatively small ensembles lead to relatively large sampling errors. Future effort will focus on quantifying the impact that each of these

issues (nonlinearity, non-normality, approximate posterior distributions, and small samples) has on the accuracy of the information content estimates.

Acknowledgements

This work has been supported in part by NASA through the ROSES-2005 AIST project, by NSF through the awards NSF CCF–0635194, NSF OCI–0904397, NSF CCF–0916493, and NSF DMS–0915047, and by the Computational Science Laboratory at Virginia Tech.

Appendix A Properties of random quadratic functions

In the paper we use the following useful property of random quadratic functions.

Let $\mathbf{Q} = \mathbf{Q}^T$ be a symmetric positive semidefinite matrix and ζ a random vector with $\mathbb{E}[\zeta] = \mu$ and $\mathrm{cov}[\zeta] = \mathbf{C}$. Then the quadratic function $\zeta^T \mathbf{Q} \zeta$ has the following statistics:

$$\mathbb{E}\left[\zeta^T \cdot \mathbf{Q} \cdot \zeta\right] = \mathrm{trace}(\mathbf{QC}) + \mu^T \cdot \mathbf{Q} \cdot \mu \ , \tag{A1a}$$

$$\mathrm{var}\left[\zeta^T \cdot \mathbf{Q} \cdot \zeta\right] = \mathrm{trace}(\mathbf{QCQC}) + 4\mu^T \cdot \mathbf{QCQ} \cdot \mu \ . \tag{A1b}$$

If $\mathbf{x} \in \mathcal{N}\left(\mathbf{x}_0^{\mathrm{A}}, \mathbf{A}_0\right)$ then $\mathbf{x} - \mathbf{x}_0^{\mathrm{A}} \in \mathcal{N}\left(0, \mathbf{A}_0\right)$ and

$$\mathbb{E}^{\mathrm{A}}\left[\frac{1}{2}\left(\mathbf{x} - \mathbf{x}_0^{\mathrm{A}}\right)^T \mathbf{A}_0^{-1}\left(\mathbf{x} - \mathbf{x}_0^{\mathrm{A}}\right)\right] = 0 + \frac{1}{2}\mathrm{trace}\left(\mathbf{A}_0^{-1}\mathbf{A}_0\right) = \frac{n}{2} \ . \tag{A2}$$

Similarly, $\mathbf{x} - \mathbf{x}_0^{\mathrm{B}} \in \mathcal{N}\left(\mathbf{x}_0^{\mathrm{A}} - \mathbf{x}_0^{\mathrm{B}}, \mathbf{A}_0\right)$ and

$$\mathbb{E}^{\mathrm{A}}\left[\frac{1}{2}\left(\mathbf{x} - \mathbf{x}_0^{\mathrm{B}}\right)^T \mathbf{B}_0^{-1}\left(\mathbf{x} - \mathbf{x}_0^{\mathrm{B}}\right)\right] = \frac{1}{2}\left(\mathbf{x}_0^{\mathrm{A}} - \mathbf{x}_0^{\mathrm{B}}\right)^T \mathbf{B}_0^{-1}\left(\mathbf{x}_0^{\mathrm{A}} - \mathbf{x}_0^{\mathrm{B}}\right) + \frac{1}{2}\mathrm{trace}\left(\mathbf{B}_0^{-1}\mathbf{A}_0\right) \ . \tag{A3}$$

Appendix B 4D-Var data assimilation with linear models, linear observation operators, and Gaussian errors

In this section we consider the case where the model dynamics is linear

$$\mathcal{M}_{t_0 \to t_i}(\mathbf{x}_0) = \mathbf{M}_i \mathbf{x}_0, \tag{B1}$$

and the observation operator is also linear,

$$\mathcal{H}(\mathbf{x}_i) = \mathbf{H}_i \mathbf{x}_i. \tag{B2}$$

In addition, we assume that the background errors and the observation errors are both normally distributed. In this case the 4D-Var cost function is:

$$\mathcal{J}^B(\mathbf{x}_0) = \frac{1}{2}\left(\mathbf{x}_0 - \mathbf{x}_0^{\mathrm{B}}\right)^T \mathbf{B}^{-1}\left(\mathbf{x}_0 - \mathbf{x}_0^{\mathrm{B}}\right)$$

$$\mathcal{J}^{\mathrm{obs}}(\mathbf{x}_0) = \sum_{i=0}^{N} \mathcal{J}_i^{\mathrm{obs}}(\mathbf{x}_0)$$

$$\begin{aligned}\mathcal{J}_i^{\mathrm{obs}}(\mathbf{x}_0) &= \frac{1}{2}\left(\mathbf{H}_i\mathbf{x}_i - \mathbf{y}_i\right)^T \mathbf{R}_i^{-1}\left(\mathbf{H}_i\mathbf{x}_i - \mathbf{y}_i\right) \\ &= \frac{1}{2}\left(\mathbf{H}_i\mathbf{M}_i\mathbf{x}_0 - \mathbf{y}_i\right)^T \mathbf{R}_i^{-1}\left(\mathbf{H}_i\mathbf{M}_i\mathbf{x}_0 - \mathbf{y}_i\right)\end{aligned}$$

The posterior distribution is Gaussian $\mathcal{P}^A(\mathbf{x}_0) = \mathcal{N}\left(\mathbf{x}_0^{\mathrm{A}}, \mathbf{A}_0\right)$. The posterior covariance matrix $\mathbf{A}_0$ satisfies

$$\mathbf{A}_0^{-1} = \mathbf{B}_0^{-1} + \sum_{i=0}^{N} \mathbf{M}_i^T \mathbf{H}_i^T \mathbf{R}_i^{-1} \mathbf{H}_i \mathbf{M}_i, \tag{B3}$$

and the analysis initial condition $\mathbf{x}_0^{\mathrm{A}}$ obtained by solving the linear system

$$\mathbf{A}_0^{-1} \cdot \left(\mathbf{x}_0^{\mathrm{A}} - \mathbf{x}_0^{\mathrm{B}}\right) = \sum_{i=0}^{N} \mathbf{M}_i^T \mathbf{H}_i^T \mathbf{R}_i^{-1}\left(\mathbf{y}_i - \mathbf{H}_i\mathbf{M}_i\mathbf{x}_0^{\mathrm{B}}\right). \tag{B4}$$

References

Abramov, R. V. and Majda, A. J.: Quantifying uncertainty for non-Gaussian ensemble in complex systems. *SIAM Journal on Scientific Computing*; Vol. 26, No. 2, pp. 411–447, 2004.

Bartlett, M. S.: An introduction to stochastic processes, with special reference to methods and applications. *Cambridge University Press*, 1962.

Cardinali, C., Pezzulli, S., Andersson, E.: Influence-matrix diagnostic of data assimilation system. *Quarterly Journal of the Royal Meteorological Society*; Vol. 130, pp. 2767-2786, 2004.

Cheng, H., Jardak, M., Alexe, M. and Sandu, A.: A hybrid approach to estimating error covariances in variational data assimilation. *Tellus A.* Vol. 62, No. 3, pp. 288–297, 2010.

Courtier, P., and Talagrand, O.: Variational assimilation of meteorological observations with the adjoint vorticity equations Part 2: Numerical results. *Quarterly Journal of the Royal Meteorological Society*; Vol. 113, pp. 1329–1347, 1987.

Fisher, R. A.: On the mathematical foundations of theoretical statistics. *Philosophical Transactions of the Royal Society of London*; Series A, Vol. 222, pp. 309–368, URL: http://www.jstor.org/stable/91208, 1922.

Fisher, M.: Estimation of entropy reduction and degrees of freedom for signal for large variational analysis systems. *ECMWF Technical Memoranda*; 397, 2003.

Heinkenschloss, M.: Numerical Solution of Implicitly Constrained Optimization Problems. CAAM Technical Report TR08-05, Department of Computational and Applied Mathematics, Rice University.http://www.caam.rice.edu/~heinken/software/matlab_impl_constr, 2008.

Le Dimet, F. X. and Talagrand, O.: Variational algorithms for analysis and assimilation of meteorological observations: theoretical aspects. *Tellus A*; Vol. 38, pp. 97–110, 1986.

Lions, J. L.: Optimal control of systems governed by partial differential equations. *Springer-Verlag*, 1971.

Lorenz, E.: Predictability: A problem partly solved. *Proceedings of the Seminar on Predictability*; Shinfield Park, Reading, UK, ECMWF, 1996.

Majda, A. J. and Wang, X.: Nonlinear dynamics and statistical theories for basic geophysical flows. *Cambridge University Press*, 2006.

Rabier, F., Fourrie, N., Chafa, D., and Prunet, P.: Channel selection methods for Infrared Atmospheric Sounding Interferometer radiances. *Quarterly Journal of the Royal Meteorological Society*; Vol. 128, pp. 1011–1027, 2002.

Rodgers, C. D.: Information content and optimization of high spectral resolution measurements. *Optical Spectroscopic Techniques and Instrumentation for Atmospheric and Space Research*; SPIE Vol. 2830, pp. 136–147, 1996.

Rodgers, C. D.: Information content and optimization of high spectral resolution measurements. *Advances in Space Research*; Vol. 21, pp. 361–367, 1998.

Rodgers, C. D.: Inverse methods for atmospheric sounding: Theory and Practice. *World Scientific: Singapore*, 2000.

Shannon, C. E. and Weaver, W.: The mathematical theory of communication. *University of Illinois Press, Urbana, IL.*, 1949.

Singh, K., Jardak, M., Sandu, A., Bowman, K., Lee, M.: A Practical Method to Estimate Information Content in the Context of 4D-Var Data Assimilation. II: Application to Global Ozone Assimilation. *Atmospheric Chemistry and Physics*; *submitted*, 2012.

Singh, K., Sandu, A., Jardak, M., Bowman, K., Lee, M.: Information Theoretic Metrics to Characterize Observations in Variational Data Assimilation. *International Conference on Computational Science ICCS 2012*, Workshop on

Atmospheric and Oceanic Computational Science, Omaha, Nebraska, 2012.

Stewart, L. M., Dance, S. L., Nichols, N. K.: Correlated observation errors in data assimilation. *International Journal for Numerical Methods in Fluids*; Vol. 56, pp. 1521–1527, 2008.

Worden, J. R., Bowman, K. W., and Jones D. B. A.: Characterization of atmospheric profile retrievals from Limb Sounding Observations of an inhomogeneous atmosphere. *Journal of Quantitative Spectroscopy & Radiative Transfer*; Vol. 86, (03)00274-7, 2004.

Xu, Q.: Measuring information content from observations for data assimilation: relative entropy versus Shannon entropy difference. *Tellus, A.*; Vol. 59(A), pp. 198–209, 2006.

Zhu, C., Byrd, R. H., and Nocedal., J.: L-BFGS-B: Algorithm 778, FORTRAN routines for large scale bound constrained optimization. *ACM Transactions on Mathematical Software*; Vol. 23, No. 4, pp. 550–560, 1997.

Zupanski, D., Hou, A.Y., Zhang, S.Q.: Applications of information theory in ensemble data assimilation. *Quarterly Journal of the Royal Meteorological Society*; Vol. 133, pp. 1533–1545, 2007.