Tellus AND OCEANOGRAPHY

PUBLISHED BY THE INTERNATIONAL METEOROLOGICAL INSTITUTE IN STOCKHOLM

Measures of observation impact in non-Gaussian data assimilation

By ALISON FOWLER* and PETER JAN VAN LEEUWEN, *Department of Meteorology, University of Reading, Reading RG6 6BB, UK*

(Manuscript received 11 January 2012; in final form 29 March 2012)

ABSTRACT

Non-Gaussian/non-linear data assimilation is becoming an increasingly important area of research in the Geosciences as the resolution and non-linearity of models are increased and more and more non-linear observation operators are being used. In this study, we look at the effect of relaxing the assumption of a Gaussian prior on the impact of observations within the data assimilation system. Three different measures of observation impact are studied: the sensitivity of the posterior mean to the observations, mutual information and relative entropy. The sensitivity of the posterior mean is derived analytically when the prior is modelled by a simplified Gaussian mixture and the observation errors are Gaussian. It is found that the sensitivity is a strong function of the value of the observation and proportional to the posterior variance. Similarly, relative entropy is found to be a strong function of the value of the observation. However, the errors in estimating these two measures using a Gaussian approximation to the prior can differ significantly. This hampers conclusions about the effect of the non-Gaussian prior on observation impact. Mutual information does not depend on the value of the observation and is seen to be close to its Gaussian approximation. These findings are illustrated with the particle filter applied to the Lorenz '63 system. This article is concluded with a discussion of the appropriateness of these measures of observation impact for different situations.

Keywords: mutual information, relative entropy, Lorenz 1963 system, particle filter

1. Introduction

In data assimilation, the aim is to combine observations with a priori information in a way which takes into account a statistical representation of their respective errors. In the Geosciences, the a priori information commonly comes from a sophisticated physical and dynamical model of the phenomena of interest, for example a numerical weather prediction.

Despite the non-linearity of these models, assimilation methods based on linearising the model and assuming Gaussian statistics have proved a powerful tool. Such methods include 4DVar, which is in operational use at many meteorological centres (Rabier, 2005). However, the assimilation is restricted to time and spatial scales where the non-linearity is small (Pires et al., 1996; Evensen, 1997).

At higher resolutions (e.g. convective scales), these geophysical models are highly non-linear, which potentially gives rise to significantly non-Gaussian a priori error distributions. This has led to an increasing interest in the methods of data assimilation, which do not rely on assumptions about the near linearity of the model and Gaussian error distributions.

Two reviews highlighting the recent developments in non-linear (non-Gaussian) data assimilation have been given by van Leeuwen (2009) and Bocquet et al. (2010). Many of the methods discussed in these papers are based on the direct application of Bayes' theorem in which the posterior distribution, $p(\mathbf{x}|\mathbf{y})$, (the probability distribution of the state given the observations), is derived from the multiplication and normalisation of the prior, $p(\mathbf{x})$, with the likelihood, $p(\mathbf{y}|\mathbf{x})$ (the probability distribution of the observations given the state):

$$p(\mathbf{x}|\mathbf{y}) = \frac{p(\mathbf{x})\,p(\mathbf{y}|\mathbf{x})}{\int p(\mathbf{x})\,p(\mathbf{y}|\mathbf{x})\mathrm{d}\mathbf{x}}. \qquad (1)$$

The analysis is then often defined as the mode or mean of the posterior distribution.

The assimilation of observations is expected to give us a better understanding of the state; however, it is clear that some observations are more useful than others.

*Corresponding author.
email: a.m.fowler@reading.ac.uk

Citation: Tellus A 2012, **64**, 17192, http://dx.doi.org/10.3402/tellusa.v64i0.17192

For example, some observations may be more accurate, and others may provide information about a larger part of state space. It is therefore important to understand the impact that individual observations and subsets of observations have on the posterior distribution.

In Gaussian data assimilation (Gaussian prior and likelihood function), many measures of observation impact on the analysis are operationally used, for example:

1. The degrees of freedom for signal (the effective degrees of freedom) (e.g. Fisher, 2003).
2. The sensitivity of the analysis to the observations, utilising the adjoint of the model (Cardinali et al., 2004).
3. The reduction in the analysis error covariances compared with the a priori error covariances. This may be related to the idea of mutual information (the reduction in entropy) as used by Eyre (1990).
4. There has also been an increasing interest in the use of relative entropy (Xu, 2007; Xu et al., 2009), which will be shown in the next section to measure the observations influence on both the analysis and the analysis error covariance.

The aim of this study is to look at the impact a non-Gaussian prior has on the observation impact as measured by three different measures: the sensitivity of the analysis to the observations, mutual information and relative entropy. It is assumed throughout that the observation error has a Gaussian distribution.

In the next section, these three measures will be derived for Gaussian data assimilation. In Section 3, a simple model of the non-Gaussian prior will be presented. The effect that this prior has on the sensitivity of the analysis to the observations will then be studied and compared with the impact on the mutual information and relative entropy. In Section 4.1, these differences will be illustrated with the Lorenz '63 model. Finally, in Section 5, a summary of the findings and conclusions will be presented.

2. Observation impact in Gaussian data assimilation

If the prior and likelihood can be assumed to be Gaussian [given by $N(\mathbf{x}_\mathrm{b}, \mathbf{B})$ and $N(\mathbf{y}, \mathbf{R})$, respectively] and the function mapping from state to observation space is linear (written as matrix $\mathbf{H}$), then the posterior distribution is also Gaussian and can be characterised solely by its mean (the analysis), $\mathbf{x}_\mathrm{a}$, and its error covariance matrix, $\mathbf{P}_\mathrm{a}$:

$$\mathbf{x}_\mathrm{a} = \mathbf{x}_\mathrm{b} + \mathbf{K}(\mathbf{y} - \mathbf{H}\mathbf{x}_\mathrm{b}). \tag{2}$$

Here, $\mathbf{K}$ is known as the Kalman gain matrix, which is a function of $\mathbf{B}$, $\mathbf{R}$ and $\mathbf{H}$:

$$\mathbf{K} = (\mathbf{H}^T\mathbf{R}^{-1}\mathbf{H} + \mathbf{B}^{-1})^{-1}\mathbf{H}^T\mathbf{R}^{-1}. \tag{3}$$

The posterior (analysis error) covariance matrix is given by:

$$\mathbf{P}_a = (\mathbf{I} - \mathbf{K}\mathbf{H})\mathbf{B}. \tag{4}$$

This is independent of the means of the prior and likelihood.

For a derivation of eqs. (2) and (4), refer to, for example, Kalnay (2003).

2.1. *The sensitivity of the analysis to the observations*

The linear relationship between the analysis and the observations, given by eq. (2), allows for a direct interpretation of the analysis sensitivity to the observations in terms of the Kalman gain matrix:

$$\mathbf{S} = \frac{\partial \mathbf{H}\mathbf{x}_\mathrm{a}}{\partial \mathbf{y}} = \mathbf{H}\mathbf{K}. \tag{5}$$

This is a $p \times p$ matrix, where p is the size of the observation space. This allows for the evaluation of the impact of individual observations on the analysis projected onto observation space. In this context, $\mathbf{S}$ was studied by Cardinali et al. (2004). From eqs. (5) and (3), it can be seen that accurate independent observations of features, which we have little prior knowledge of, have the greatest impact on the analysis because the Kalman gain will be large (Cardinali et al., 2004).

The trace of $\mathbf{S}$ gives the degrees of freedom for signal, d_s. This evaluates the expected fit of the analysis to $\mathbf{x}_\mathrm{b}$ normalised by the error covariance matrix, $\mathbf{B}$. That is $d_\mathrm{s} = E[(\mathbf{x}_\mathrm{a} - \mathbf{x}_\mathrm{b})^T\mathbf{B}^{-1}(\mathbf{x}_\mathrm{a} - \mathbf{x}_\mathrm{b})] = \mathrm{trace}(\mathbf{S})$ (Rodgers, 2000). The diagonal elements of the sensitivity matrix are bounded by 0 and 1 if $\mathbf{R}$ is diagonal, and therefore d_s lies between 0 and p. The closer d_s is to p the greater the observation impact.

2.2. *Mutual information*

Mutual information measures the reduction in entropy when an observation is made, that is the difference between entropy in the prior and the posterior. In information theory, entropy is a measure of the uncertainty associated with a random variable. For a probability distribution $p(\chi)$, entropy can be defined as $-\int p(\chi)\ln p(\chi)\mathrm{d}\chi$. The entropy of a conditional probability distribution, $p(\chi|\mathbf{z})$, is defined as $-\int\int p(\chi, \mathbf{z})\ln p(\chi|\mathbf{z})\mathrm{d}\chi\mathrm{d}\mathbf{z}$. This is the expected entropy of χ when conditioning with $\mathbf{z}$. (Note that given these definitions, entropy is dependent on the choice of units for the variable χ.)

When $p(\chi)$ is a Gaussian, the entropy associated with χ depends only on its covariance matrix, $\mathbf{C}_\chi$. The entropy in this case is given by (1/2)ln $[(2\pi e)^n|\mathbf{C}_x|]$, where n is the size of the vector χ and |*| denotes the determinant (Rodgers, 2000). Mutual information for a Gaussian prior and posterior is therefore given by:

$$\mathrm{MI} = \frac{1}{2}\ln|\mathbf{B}\mathbf{P}_\mathrm{a}^{-1}|. \tag{6}$$

Mutual information for Gaussian data assimilation is therefore a measure of the difference in the determinant of the prior and posterior covariance matrices. Hence, it can be interpreted as the difference between a measure of the hypervolumes enclosed by iso-probability surfaces of the prior and posterior (Tarantola, 2005).

Mutual information can be rewritten in terms of the eigenvalues of the sensitivity matrix, $\mathbf{S}$, presented in the last section. Firstly, note that $\mathbf{B}\mathbf{P}_\mathrm{a}^{-1} = (\mathbf{I}_n - \mathbf{H}\mathbf{K})^{-1}$ using eq. (4) and the determinant of $(\mathbf{I}_n - \mathbf{H}\mathbf{K})$ is equal to the determinant of $(\mathbf{I}_p - \mathbf{K}\mathbf{H})$, where $\mathbf{I}_n$ and $\mathbf{I}_p$ are the identity matrix of dimension $n \times n$ and $p \times p$, respectively. This leads to:

$$\mathrm{MI} = -\frac{1}{2}\sum_{i=1}^{r}\ln|1-\lambda_i|, \tag{7}$$

where λ_i is the ith eigenvalue of $\mathbf{S}$ (ordered in descending magnitude) and $r \le \min(n,p)$ is the rank of $\mathbf{S}$. This links mutual information to the sensitivity of the analysis to the observations and hence to the degrees of freedom for signal [trace ($\mathbf{S}$)]. It is a scalar interpretation of the observation impact, and therefore the impact of individual observations may not be easily quantified. However, mutual information can be shown to be additive with successive observations (see Appendix A.1).

2.3. *Relative entropy*

Relative entropy measures the gain in information of the posterior relative to the prior.

$$\mathrm{RE} = \int p(\mathbf{x}|\mathbf{y})\ln\frac{p(\mathbf{x}|\mathbf{y})}{p(\mathbf{x})}\mathrm{d}\mathbf{x}. \tag{8}$$

Relative entropy can be thought of as a measure of the 'distance' between $p(\mathbf{x}|\mathbf{y})$ and $p(\mathbf{x})$. However, it is not a true distance because it is not symmetric (Cover and Thomas, 1991).

When both the prior and posterior are Gaussian, relative entropy is given by (Bishop 2006):

$$\mathrm{RE} = \frac{1}{2}(\mathbf{x}_\mathrm{a}-\mathbf{x}_\mathrm{b})^T\mathbf{B}^{-1}(\mathbf{x}_\mathrm{a}-\mathbf{x}_\mathrm{b}) + \frac{1}{2}\ln|\mathbf{B}\mathbf{P}_\mathrm{a}^{-1}| + \frac{1}{2}\mathrm{trace}(\mathbf{B}^{-1}\mathbf{P}_\mathrm{a}) - \frac{1}{2}n, \tag{9}$$

The first term is known as the signal term, which measures the change in the mean of the distribution. The rest is known as the dispersion term, which measures the change in the covariances. This can be shown to equal MI minus half the degrees of freedom for signal given by the trace of the sensitivity matrix. The dispersion term can therefore be written in terms of the eigenvalues of the sensitivity matrix whilst the signal term depends on the value of the observations and the prior mean.

The dependence of relative entropy on both the mean and variance of the posterior makes it an attractive measure, as it gives a more complete description of the observation impact. It can also be shown that relative entropy is invariant under a general non-linear change of variables (Kleeman, 2011). In Section 3.3, it is seen that mutual information may be written as a measure of the relative entropy of $p(\mathbf{x}, \mathbf{y})$ with respect to $p(\mathbf{x})p(\mathbf{y})$. Writing mutual information in this way shows that if the same non-linear transformation is applied to $\mathbf{x}$ and $\mathbf{y}$ then mutual information is also left invariant.

A comparison of the degrees of freedom for signal, mutual information and relative entropy for Gaussian assimilation was performed by Xu et al. (2009). It was concluded that in application to the optimal radar scan configurement there was little difference in which measure was used. In the next section, we shall look at how a non-Gaussian prior affects these three measures.

3. Observation impact in non-Gaussian data assimilation

From the previous section, it is seen that in Gaussian data assimilation the impact of the observations as measured by the three different measures is dependent on the ratio of the prescribed error variances of the prior and observations as well as their means for relative entropy. When the prior is non-Gaussian, the additional structure in the prior will be shown to be important for calculating the impact of the observations.

A study performed by Bocquet (2008) compared the information content of observations when the prior is assumed to be Gaussian and Bernoulli for the inverse modelling of a pollutant source. For this case study, a tracer gas was released from a point source over Northern France; observations of the gas were then made at locations across Europe. The measures of observation impact included the analysis error variance, the sensitivity matrix, mutual information and the degrees of freedom for signal. It was found that the more realistic non-Gaussian prior, which took into account the positivity of the released mass, allowed for observations far from the source to have a far greater impact on the analysis, giving a more accurate retrieval.

In this work, it is intended to use a much simpler idealised setup to understand the difference between the analysis sensitivity, mutual information and relative entropy when the prior is no longer Gaussian. An initially Gaussian prior may become non-Gaussian due to a non-linear forecast model. In particular, the model dynamics may lead to a skewed or multimodal distribution. The particle filter (PF) is an example of an assimilation scheme, which tries to represent the non-linear evolution of the prior (van Leeuwen, 2009). This will be used in Section 4.1 to illustrate the effect of the prior structure on the observation impact. Firstly, we shall look at the case when the prior can be modelled as a Gaussian mixture.

3.1. *Problem setup*

A Gaussian mixture allows for the representation of a wide range of non-Gaussian prior distributions. It is given by:

$$p(\mathbf{x}) = \sum_{i=1}^{N} w_i[(2\pi)^n|\mathbf{B}_i|]^{-(1/2)} \times \exp\left[-\frac{1}{2}(\mathbf{x}-\mathbf{x}_{\mathrm{b}i})^T\mathbf{B}_i^{-1}(\mathbf{x}-\mathbf{x}_{\mathrm{b}i})\right], \tag{10}$$

where $\Sigma_{i=1}^{N} w_i = 1$ (Bishop, 2006) and $\mathbf{x}_{\mathrm{b}i}$ and $\mathbf{B}_i$ are the mean and covariance of the *i*th Gaussian component, respectively.

In this study, $p(x)$ is simplified to a two-component Gaussian mixture (i.e. $N=2$) in one dimension (i.e. $n=1$). To reduce the number of parameters describing the prior further, the variances of the two component Gaussians are equal. This allows the prior to be described by four free parameters:

(1) w, the weight given to the first Gaussian, leaving the weight given to the second Gaussian as $1-w$;
(2) μ_1, the mean of the first Gaussian;
(3) μ_2, the mean of the second Gaussian; and
(4) σ^2, the variance of both Gaussian components.

Although restrictive, a large range of non-Gaussian priors can be modelled by this mixture, see, for example, Fig. 5.

The likelihood function is then taken to be Gaussian with mean μ_y, interpreted as the measured value, and variance $k\sigma^2$, where k is a scalar. This is equivalent to assuming we have direct observations of x.

Using Bayes' theorem, eq. (1), this implies that the posterior distribution is also a two-component Gaussian distribution with updated parameters given by $\tilde{w}$, $\tilde{\mu}_1$, $\tilde{\mu}_2$ and $\tilde{\sigma}^2$:

$$\tilde{w} = \frac{we^{-a_1}}{we^{-a_1} + (1-w)e^{-a_2}}, \tag{11}$$

where $a_i = [(\mu_y - \mu_i)^2)/(2(1+k)\sigma^2]$:

$$\tilde{\mu}_i = \frac{\mu_y + k\mu_i}{1+k}, \tag{12}$$

for $i=1$, 2:

$$\tilde{\sigma}^2 = \frac{k\sigma^2}{1+k}.$$

3.2. *Analysis sensitivity to the observations*

Here, we define the analysis as the mean of the posterior. In many cases. this will not correspond to the mode, as the posterior may be bi-modal [or multimodal in the case of the more general prior described by eq. (10)]. This makes the mode more difficult to uniquely define, and the mode will have infinite sensitivity to observations when the mode transfers from one peak to another.

When the prior is non-Gaussian, the simple linear relationship between the mean of the posterior and the observations [as seen in eq. (2)] breaks down. This can be seen in the following.

The mean of the posterior, the analysis, is given by:

$$\mu_a = \tilde{w}\tilde{\mu}_1 + (1-\tilde{w})\tilde{\mu}_2.$$

Recall that $\tilde{w}$ and $\tilde{\mu}_i$ are a function of μ_y given by eqs. (11) and (12), respectively. The sensitivity of the analysis may be computed as:

$$S = \frac{\partial\mu_a}{\partial\mu_y} = \frac{1}{k+1} + (\tilde{\mu}_1 - \tilde{\mu}_2)\frac{\partial\tilde{w}}{\partial\mu_y}.$$

With a little manipulation this can be written in terms of the parameters describing the prior:

$$S = \frac{1}{k+1} + \frac{kw(1-w)(\mu_1-\mu_2)^2\exp^{-a_1-a_2}}{(1+k)^2\sigma^2[w\exp^{-a_1} + (1-w)\exp^{-a_2}]^2}. \tag{13}$$

From eq. (13) it is seen that S is a function of the observation value due to the appearance of the exponent a_i.

An illustration of S as a function of μ_y is given in Fig. 1 for $k=2$, $\sigma^2=1$, $w=0.25$, $\mu_1=-1.5$, $\mu_2=1.5$. On the left, the prior given by these parameters is plotted; it is both negatively skewed and bimodal. On the right, S is given.

S is seen to be a symmetric function about a maximum at $\mu_y=\mu_0$ (solid grey vertical line). Away from $\mu_y=\mu_0$ the sensitivity asymptotes to [1/(k+1)] [(1/3) in this case].

The value of μ_y for which the analysis shows the greatest sensitivity, μ_0, is given in terms of the parameters describing the prior and likelihood as:

$$\mu_0 = \frac{1}{2(\mu_1-\mu_2)}\left[\mu_1^2 - \mu_2^2 - 2(1+k)\sigma^2\ln\left(\frac{w}{1-w}\right)\right]. \tag{14}$$

This is found by solving $(\partial^2\mu_{\mathrm{a}}/\partial\mu_y^2) = 0$ for μ_y.

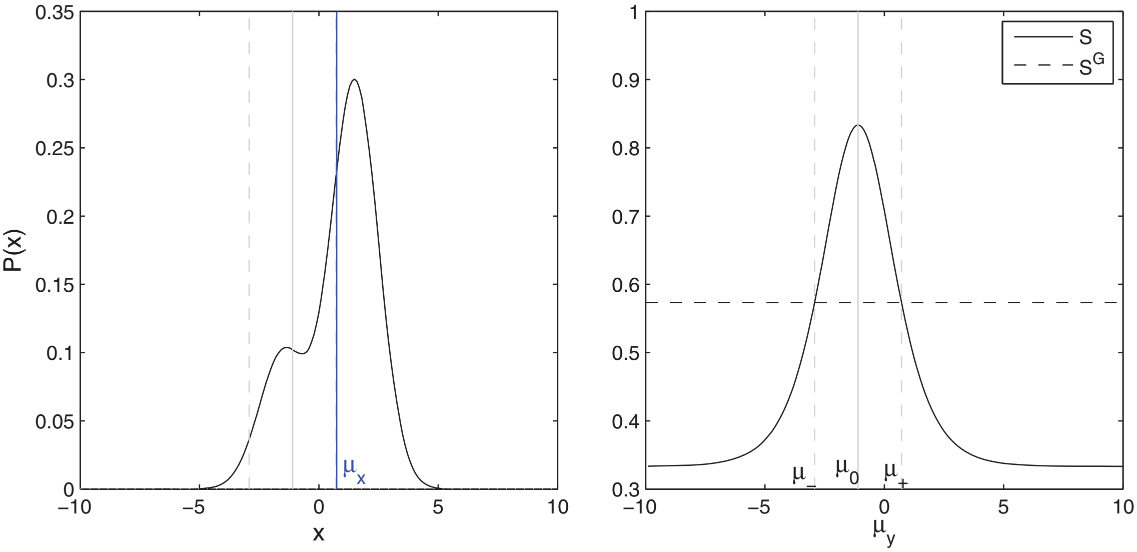

Fig. 1. Left: the prior distribution. The vertical blue line shows the prior mean, μ_x. Right: $(\partial\mu_a/\partial\mu_y)$ (solid) and the Gaussian approximation (dashed) for $k=2$, $\sigma^2=1$, $w=0.25$, $\mu_1=-1.5$, $\mu_2=-1.5$. μ_-, μ_0 and μ_+ explained within the text.

It can be shown that for this value of μ_y the Gaussian components in the posterior have equal weight, that is $w\,e^{-a_1}=(1-w)e^{-a_2}$ from eq. (11). Therefore, the posterior is symmetric and the analysis (given by the mean) is in the well between the two Gaussian components of the posterior. At this point, a small change in the observed value will dramatically change the shape of the posterior and the value of its mean. Note that this is also true for the posterior mode, which has infinite sensitivity in the region where observations give a symmetric posterior distribution.

An implication of this is that for a fixed prior and observation error variance, the observation that has maximum impact on the analysis value also gives the largest analysis error variance. In fact it can be proved:

$$S=\frac{\partial\mu_\mathrm{a}}{\partial\mu_y}=\frac{\sigma^2_{x|y}}{\sigma^2_y}, \tag{15}$$

for Gaussian mixture priors of any order and Gaussian likelihoods, see Appendix A.2. Note that S is unbounded as $[(\mu_1-\mu_2)^2/\sigma^2]$ increases. Therefore, $\sigma^2_{x|y}$ may be much greater than σ^2_y when the prior describes two highly probable but distinct regimes.

Also plotted in Fig. 1 is the sensitivity of the analysis to the observations when the prior is approximated by a Gaussian distribution, S^G, (dashed line). This is not a function of the value of the observation as shown in Section 2.1. For this case:

$$S^G=\frac{\sigma^2_x}{\sigma^2_x+\sigma^2_y} \tag{16}$$

where σ^2_x is the variance of the prior, $\sigma^2_x=\sigma^2+w(1-w)(\mu_1-\mu_2)^2$. Substituting this into eq. (16) gives:

$$S^G=\frac{1}{k+1}+\frac{kw(1-w)(\mu_1-\mu_2)^2}{(1+k)^2\sigma^2+(1+k)w(1-w)(\mu_1-\mu_2)^2}.$$

This is bounded by $[1/(k+1)]$ and 1. Note that throughout this paper $*^G$ refers to the value of * derived when approximating the prior as a Gaussian.

From Fig. 1 we see that when the full prior is used to assimilate the observation the analysis may be both more or less sensitive to the observation than when the prior is approximated by a Gaussian. The degree to which the sensitivity is affected depends on the value of the observation. When $\mu_-<\mu_y<\mu_+$ (marked on Fig. 1 by the dashed grey vertical lines) the Gaussian approximation results in an analysis, which is less sensitive to the observations than when the full prior is used and vice versa when the observation is outside of this region. The dependence of μ_- and μ_+ on the parameters of the prior and likelihood and the magnitude of the disagreement between S and S^G will be discussed further in the next section.

3.3. *Comparison to mutual information and relative entropy*

Recall that mutual information is given by the prior entropy minus the conditional entropy:

$$\mathrm{MI}=-\int p(x)\ln[p(x)]\mathrm{d}x+\int\int p(x,y)\ln[p(x|y)]\mathrm{d}x\mathrm{d}y \tag{17}$$

Despite the dependence of the posterior error variance on the observations, the conditional entropy is independent of the value of the observations. Therefore, mutual information, unlike the sensitivity of the posterior mean to the observations, is independent of the value of the observations. This is shown in Fig. 2 where the measures are normalised by their Gaussian approximations. The analysis sensitivity is given in black and the mutual information is given in red. In this example, the Gaussian approximation to mutual information is about 102% of the true value, a very small error.

Also plotted in Fig. 2 is the relative entropy normalised by its Gaussian approximation, RE^G. As seen in Section 2.3, relative entropy combines the effect of the observations on both the position and shape of the posterior distribution. This explains the asymmetry in $\mathrm{RE}/\mathrm{RE}^G$ as:

(1) The error in the effect of the observations on the shape of the posterior is to some extent measured by the error in $\sigma^2_{x|y}$ and hence S [see eq. (15)]. This is symmetric about $\mu_y = \mu_0$ (solid grey vertical line in Figs. 1 and 2).

(2) However, the error in the position of the posterior is given by the error in the squared difference between the prior mean and the analysis $(\mu_a - \mu_x)^2$. This is zero at $\mu_y = \mu_x$ (solid blue vertical line), but while $(\mu_a^G - \mu_x)^2$ is a quadratic function of μ_y this is distorted in the non-Gaussian case because of the dependence of $(\partial\mu_a/\partial\mu_y)$ on μ_y.

Using relative entropy to measure how the observation impact changes when the full prior is used may now lead to different conclusions than when the analysis sensitivity is used.

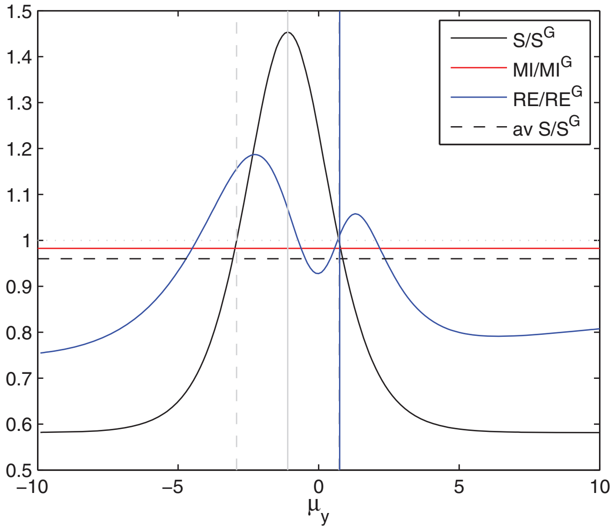

Fig. 2. S (black), MI (red), RE (blue) all normalised by their Gaussian approximations. For the same parameters in Fig. 1. The black dashed line shows $\int p(y)S\,\mathrm{d}y$ normalised by its Gaussian approximations.

- Firstly, in this case, the range of values for the error in relative entropy when a Gaussian prior is assumed is smaller.
- Secondly, the observation values, which have a bigger (lesser) impact when the full prior is used, can differ. For example, in Fig. 2, if the observation value were -5, then the relative entropy would agree with the Gaussian approximation but the Gaussian approximation to the analysis sensitivity would be almost twice its real value. Similarly if the observation value were -3, the Gaussian approximation to the relative entropy would be about (5/6) times its real value, whilst the error in the Gaussian approximation to the analysis sensitivity would be approximately 0.

The large variation of these two measures as a function of the observation makes their interpretation more difficult for a single experiment. For some applications such as the design of new observation systems it may be more useful to look at the average impact.

The expected value of relative entropy can be shown to be equal to mutual information. This can be shown by writing mutual information in its equivalent form:

$$\mathrm{MI} = \int\int p(\mathbf{x},\mathbf{y}) \ln\left[\frac{p(\mathbf{x},\mathbf{y})}{p(\mathbf{x})\,p(\mathbf{y})}\right] \mathrm{d}\mathbf{x}\mathrm{d}\mathbf{y}. \tag{18}$$

Here mutual information is interpreted as how ‘close’ two variables are to being independent, that is the error in approximating $p(\mathbf{x},\ \mathbf{y})$ by $p(\mathbf{x})\,p(\mathbf{y})$ (Cover and Thomas, 1991). In this form MI can be seen to be $\int p(y)\mathrm{RE}\,\mathrm{d}y$ [see eq. (8)]. The marginal distribution, $p(y)$, is given by:

$$p(y) = \sum_{i=1}^{N} w_i A_i e^{-a_i}, \tag{19}$$

where

$$A_i = \frac{1}{\sqrt{2\pi\left(\sigma_i^2 + \sigma_y^2\right)}}.$$

In Fig. 2, we can compare mutual information (red) to the expected analysis sensitivity (black dashed), $\int p(y)S\,\mathrm{d}y$. It is seen that for this case mutual information is marginally closer to its Gaussian approximation than $\int p(y)S\,\mathrm{d}y$ is. In both cases, the Gaussian approximation to the prior overestimates the average impact of the observations, as this assumes the prior has less information (less structure) than in fact it does.

The relationship between the three measures of observation impact shown in Fig. 2 extends to a wider range of prior distributions described by the simple two-component Gaussian mixture. In Fig. 3, the analysis sensitivity and relative entropy are plotted for a range of values of μ_2–μ_1 when $w = (1/2)$ (top row) and w when $\mu_2 - \mu_1 = 3$ (bottom row).

For this range of prior distributions, the magnitude of the error in the Gaussian approximation to the relative entropy is smaller than the error in the analysis sensitivity. The tilt in the error fields seen for both measures as w is varied (bottom row) follows from the equation for μ_0, eq. (14).

It is seen in Fig. 3 that the magnitude of the error in the Gaussian approximation increases as the prior becomes more non-Gaussian, that is more bi-modal ($\mu_2 - \mu_1$ increases) and more skewed [$|w - (1/2)|$ increases], as one would expect.

The range of values for which the sensitivity is underestimated (red contours), given by $\mu_- - \mu_+$, is only weakly a function of the non-Gaussianity of the prior. It is more greatly influenced by the variance of the likelihood. As the error variance of the observations increases the magnitude of S decreases (as would be expected, poor observations have a weaker impact). However, the range of values of μ_y for which S is greater than S^G also increases.

In Fig. 4, mutual information normalised by its Gaussian approximation (left) can be compared with the average analysis sensitivity normalised by its Gaussian approximation (right) for a range of values of $\mu_2 - \mu_1$ (y-axis) and w (x-axis).

As expected from Fig. 3, in which S/S_G was seen to be generally larger than $\mathrm{RE}/\mathrm{RE}_G$, the error in the Gaussian approximation of MI is less than the error in the Gaussian approximation of $\int p(y) S\, \mathrm{d}y$. In both measures, the Gaussian approximation always overestimates the observation impact but only by a marginal amount, increasing as the prior becomes more skewed and more bimodal. The peak in the skewness, calculated as $\sigma_x^{-3} \int (x - \mu_x)^3 p(x) \mathrm{d}x$, is given by the grey lines in Fig. 4.

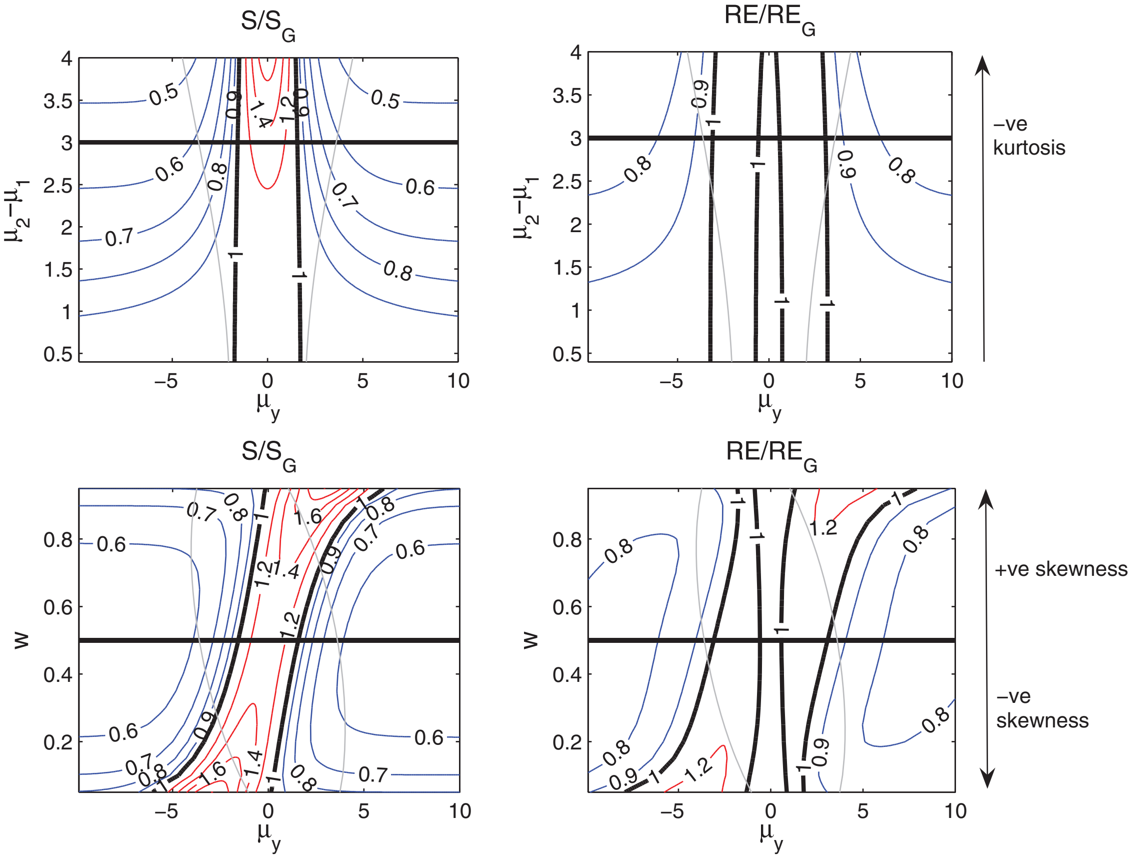

Fig. 3. Contour plots of S (left) and RE (right) all normalised by their Gaussian approximations. These are given as a function of μ_y and $\mu_2 - \mu_1$ when $w = (1/2)$ (top row) and w when $\mu_2 - \mu_1 = 3$ (bottom row). $\sigma^2 = 1$, $k = 2$ as in Figs. 1 and 2. The grey lines mark $\mu_y = \mu_x \pm 2\sigma_x$, where μ_x and σ_x are the mean and *SD* of the prior, respectively.

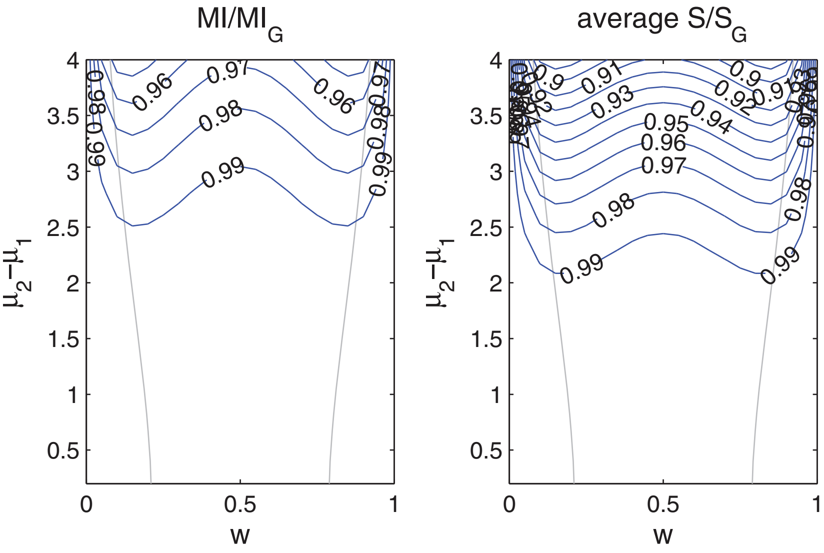

Fig. 4. Contour plots of MI (left) and $\int p(y)S\,\mathrm{d}y$ (right) all normalised by their Gaussian approximations. These are given as a function of $\mu_2 - \mu_1$ (y-axis) and w (x-axis). $\sigma^2 = 1$, $k = 2$. The grey lines mark the peak in the skewness (both positive and negative) of the prior.

4. Calculating observation impact in the PF

4.1. Illustration using the Lorenz '63 model

The effect of the non-Gaussian prior on the measures of observation impact is now illustrated using the low-dimensional Lorenz 1963 model (Lorenz, 1963), given by:

$$\begin{aligned} \frac{\mathrm{d}\chi_1}{\mathrm{d}t} &= \sigma(\chi_2 - \chi_1) \\ \frac{\mathrm{d}\chi_2}{\mathrm{d}t} &= -\chi_1\chi_3 + \rho\chi_1 - \chi_2 \\ \frac{\mathrm{d}\chi_3}{\mathrm{d}t} &= \chi_1\chi_2 - \beta\chi_3. \end{aligned} \tag{20}$$

Using the following parameters, $\sigma = 10$, $\rho = 28$ and $\beta = 8/3$, eq. (20) gives rise to a chaotic system with a strange attractor. The solution is seen to orbit around two equilibrium points giving two 'regimes'.

We can represent a prior distribution of the state, $\mathbf{x} = (\chi_1, \chi_2, \chi_3)^T$, by a large number, N_p, of weighted 'particles' (e.g. van Leeuwen, 2009):

$$p(\mathbf{x}^t) \approx \sum_{i=1}^{N_p} w_i^{t-1}\delta(\mathbf{x} - \mathbf{x}_i^t) \tag{21}$$

where $\delta(*)$ is the Dirac delta function, i is the particle index and t is the observation time index. The number of particles, N_p, used in this work is 10 000 to avoid sampling issues.

At the initial time, all weights, w_i^0, are equal to $(1/N_p)$. In this work, the initial prior is taken to be Gaussian with covariance matrix given by:

$$\mathbf{B} = \begin{pmatrix} 1 & \frac{1}{2} & \frac{1}{4} \\ \frac{1}{2} & 1 & \frac{1}{2} \\ \frac{1}{4} & \frac{1}{2} & 1 \end{pmatrix},$$

and therefore the initial particles are drawn from $N(\mathbf{x}_\mathrm{T}, \mathbf{B})$, where $\mathbf{x}_\mathrm{T}$ is the known truth at initial time.

This initial prior distribution is then evolved forward in time to the first observation by propagating each particle forward using a fourth-order Runge-Kutta discretisation of the equations given by eq. (20) with a time step of 0.01. Observations, $\mathbf{y}$, are relatively sparse, made at every 50 time steps, and of only the χ_1 variable. The observation error is assumed to be Gaussian with mean zero and a large error variance, $\sigma_\mathrm{y}^2 = 10$. At the time of the first observation the non-linearity of the Lorenz '63 system gives rise to a new prior distribution of χ_1, which is no longer Gaussian.

At the time of the observation, the weights of each particle (which were initially all equal) are updated using:

$$w_i^t = \frac{w_i^{t-1}p(y^t|\mathbf{x}_i^t)}{\sum_{j=1}^{N_p} w_j^{t-1}p\left(y^t|\mathbf{x}_j^t\right)}.$$

The particles with updated weights now represent the posterior distribution, which is given by:

$$p(\mathbf{x}^t|\mathbf{y}^{1:t}) \approx \sum_i^{N_p} w_i^t\delta(\mathbf{x} - \mathbf{x}_i^t). \tag{22}$$

The particles are then propagated forward to the next observation, and the weights are again updated. It is desirable to have a large number of particles with non-negligible weight, so that the posterior distribution is accurately represented. A common problem with this standard PF for a limited sample size is filter divergence, when over time all the weight falls onto only a few particles.

In this example, the large sample size ensures that the effective number of particles $[1/(\Sigma_{\mathrm{i}}\mathrm{w}_{\mathrm{i}}^2)]$ is greater than 25 up to the 10th observation time. Note that no resampling of the particles is used.

In Fig. 5, the priors are plotted as histograms for the first 10 observation times, and the observations are given by the blue stars.

The priors are far from Gaussian (black line). In particular at observation times 4 and 9 the prior appears to be bimodal. At each assimilation time, the observed state is represented by the particles. This is in agreement with the large effective number of particles (see definition above).

In Fig. 6a, the analysis (mean of the particles, $\mu_a = \sum_{i=1}^{N_p} w_i \mathbf{x}_i$) of χ_1 (red) is plotted alongside the true trajectory (grey) and the observations (black crosses). The analysis, given by the mean of the particles, gives a fairly good estimate of the truth until the 350th time step (7th observation) when there is a large uncertainty as to the sign of the χ_1 seen in Fig. 5.

For these non-idealised prior distributions we can now calculate the impact of the observations using the three different measures and compare to the observation impact when the priors are approximated by a Gaussian. Firstly, the Gaussian approximations to the observation impacts are calculated using eqs. (5), (6) and (9) in Section 2. $\mathbf{x}_{\mathrm{b}}$ and $\mathbf{B}$ are calculated directly from the particles at the observation time using the weights before they have been updated, for example:

$$\mathbf{x}_{\mathrm{b}}^t = \sum_{i=1}^{N_p} w_i^{t-1}\mathbf{x}_i^t \tag{23}$$

$$\mathbf{B}^t = \sum_{i=1}^{N_p} w_i^{t-1}(\mathbf{x}_i^t - \mathbf{x}_{\mathrm{b}}^t)(\mathbf{x}_i^t - \mathbf{x}_{\mathrm{b}}^t)^{\mathrm{T}}. \tag{24}$$

$\mathbf{x}_{\mathrm{a}}$ and $\mathbf{p}_{\mathrm{a}}$ are then calculated using eqs. (2) and (4), where $\mathbf{H} = (1, 0, 0)$.

In this example, $\mathbf{R} = \sigma_{\mathrm{y}}^2$ is constant and so for the Gaussian approximations to the observation impact only

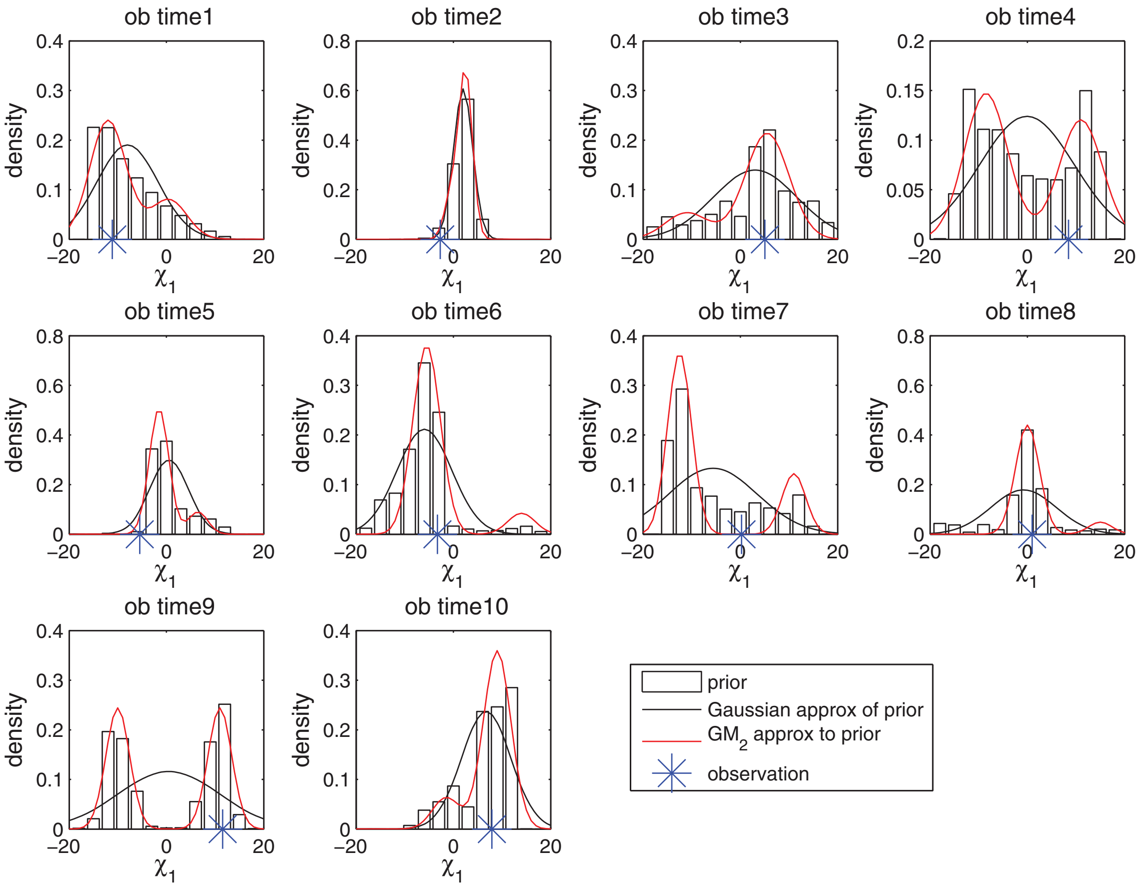

Fig. 5. The evolution of the marginal prior distribution of χ_1: each panel gives a histogram representation of the particles at the time of the observations (bar plots). The blue stars give the value of the observations at each time. The black lines give a Gaussian approximation to the prior distribution and the red lines give a two-component Gaussian mixture fit to the prior distribution with identical variances, as described in Section 3.1.

the spread in the prior distribution is important, given by the variance, this is plotted in Fig. 6b. In Fig. 6c, the Gaussian approximations to the observation impacts are plotted. As expected in all cases the observation impact is large when the prior variance is large (observation times 4, 7 and 9).

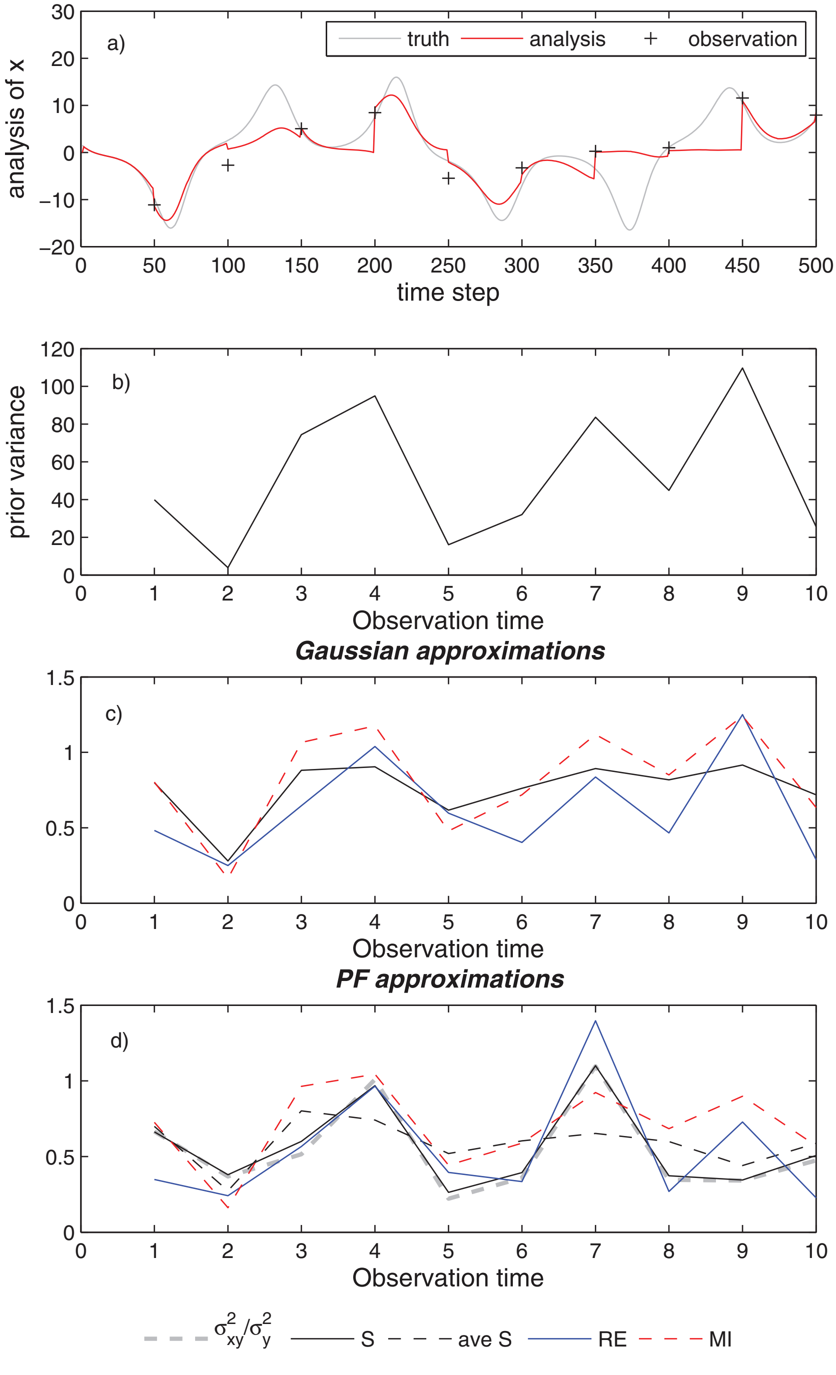

Fig. 6. (a) The analysis (mean of particles) as a function of time (red), the true trajectory (grey) and observations of the truth (black crosses). (b) The prior variance as a function of observation time. (c) Approximations to the analysis sensitivity (black), relative entropy (blue) and mutual information (red dashed) assuming the prior distribution is Gaussian, with mean and covariance calculated from the weighted particles. (d) Approximations to the analysis sensitivity (black), relative entropy (blue) and mutual information (red dashed) calculated directly from the particle representation of the prior and posterior. Also plotted is the expected sensitivity $\int p(y)S\,dy$ (black dashed) and the ratio of the posterior variance to the observation error variance (grey dashed).

The strong dependence of each measure on the ratio of the observation error variance to the prior error variance means that each measure is seen to be in good agreement, as was also seen by Xu et al. (2009). Even relative entropy (blue) is dominated by the change in the prior variance rather than the value of the observation. This is not necessarily expected, as for Gaussian statistics relative entropy is a quadratic function of the observed value [see eq. (9)]. As such the consistency of relative entropy with the other measures may change dramatically depending on the underlying model used. The results given in Kleeman (2002) illustrate this for the application of quantifying predictability.

In order to calculate the sensitivity of the analysis to the observations in the full non-Gaussian case, small perturbations to the observations, $\Delta\mu_y$, are made at each observation time; the analysis sensitivity is then approximated by the change in the analysis of χ_1, $\Delta\mu_\mathrm{a}$, that is:

$$\frac{\partial\mu_\mathrm{a}}{\partial\mu_y} \approx \frac{\Delta\mu_\mathrm{a}}{\Delta\mu_y}. \tag{25}$$

However, using eq. (15) we may also approximate the sensitivity using the spread in the particles using the updated weights, and therefore:

$$\frac{\partial\mu_\mathrm{a}}{\partial\mu_y} \approx \frac{1}{\sigma_\mathrm{y}^2}\sum_{i=1}^{N_p} w_i^t\left[(\chi_1)_i^t - (\mu_{\chi_1})_\mathrm{a}^t\right]^2, \tag{26}$$

where $\left(\mu_{\chi_1}\right)_\mathrm{a}^t$ is the mean value of χ_1 in the updated particles at observation time t, that is $\left(\mu_{\chi_1}\right)_\mathrm{a}^t = \sum_{i=1}^{N_p} w_i^t(\chi_1)^t$. The agreement between these two approximations to the sensitivity is seen to be good in Fig. 6d, comparing the black solid line ($\Delta\mu_a/\Delta\mu_y$) to the grey-dashed line $\left(\sigma_{x|y}^2/\sigma_y^2\right)$.

The relative entropy may be calculated directly from the weights:

$$\mathrm{RE}^t \approx \sum_{i=1}^{N_p} w_i^t \ln\frac{w_i^t}{w_i^{t-1}} \tag{27}$$

and is given by the blue line in Fig. 6d.

Mutual information is approximated using quadrature by:

$$\mathrm{MI}^t \approx \sum_{j=1}^{P} \mathrm{RE}^t\left(y_j^t\right) p\left(y_j^t\right)\Delta y \tag{28}$$

where

$$p\left(y_j^t\right) = \sum_{i=1}^{N_p} w_i^{t-1} P\left(y_j|\mathbf{x}_i^t\right)$$
$$\text{for } y_j = -20, -20+\Delta y, \ldots, 20-\Delta y, 20.$$

Δy was taken to be 1.

For the non-Gaussian representation of the prior the observation impacts are seen to roughly follow the pattern of higher observation impact when the prior variance is large (Fig. 6d). However, there are some notable differences between these and the Gaussian approximations to these measures:

(1) The sensitivity of the analysis to the observation is more variable, fluctuating from about 0.2 to 1. In particular, the sensitivity is much greater at observation time 7 and much less at time 9. This leads to cases when the analysis is less sensitive to observations despite a larger prior variance, for example comparing the sensitivity at observation time 9 to time 8.

(2) The Gaussian approximation to the sensitivity is more comparable to the full sensitivity averaged over observations, $\int Sp(y)\,\mathrm{d}y$ (black-dashed line in Fig. 6d). Although again the relationship between this measure and the prior variance does break down at observation time 9.

(3) For relative entropy and mutual information, the relationship between high observation impact and large prior variance appears to hold more robustly, although there are still discrepancies for relative entropy. For example, relative entropy is greater at observation time 7 than 9 despite the prior variance being greater at observation time 9.

It is clearer to see the differences between the Gaussian and PF approximations by plotting their ratio, as given in Fig. 7a.

As also seen in Section 3, the Gaussian approximation always overestimates the mutual information (red-dashed line) and the averaged sensitivity (black-dashed line). The error in these averaged measures roughly increases with time, as the Gaussian approximation to the prior becomes increasingly poor (see Fig. 7b). In particular, at observation time 9, S^G is approximately twice $\int Sp(y)\,\mathrm{d}y$. The error in the Gaussian approximation to the sensitivity and relative entropy is seen to fluctuate. As already shown in Section 3, these errors are expected to depend strongly on the exact value of the observation.

In Fig. 8, the four measures are plotted as a function of the observation value and then normalised by their Gaussian approximation (similar to Fig. 2) for observation time 7 (when the error in relative entropy is largest) and time 9 (when the sensitivity is much smaller than expected given the increase in the prior variance at this time; see Fig. 6d).

At each of these times there is indeed a large range of values for the normalised S and RE as a function of observation value.

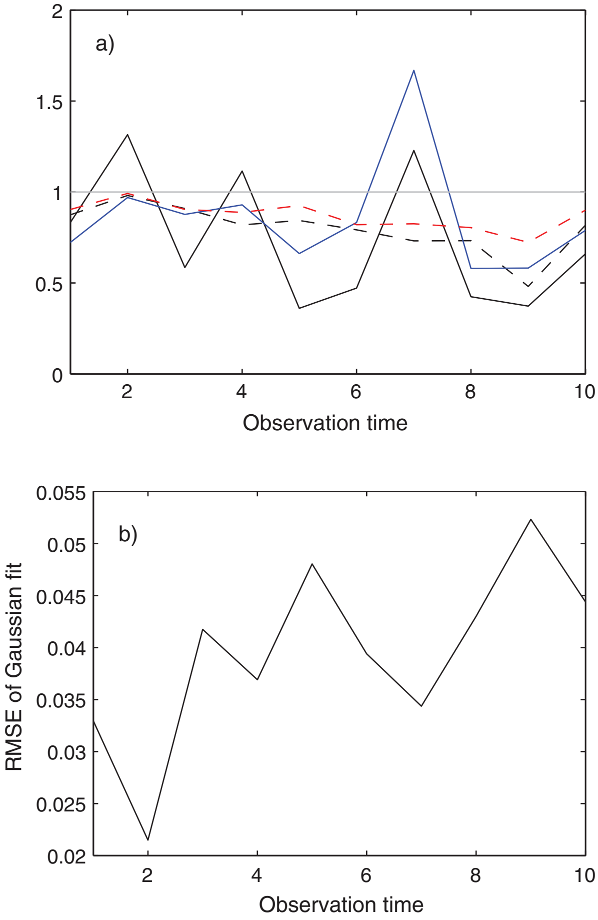

Fig. 7. Top: the PF approximation to the observation impact divided by the Gaussian approximation for each measure. Line colours as in Fig. 6d. Bottom: the RMSE in the Gaussian approximation to the prior distribution as a function of observation time.

The error in the Gaussian approximation of S has the largest range at observation time 9 when the prior is clearly bimodal (see Fig. 5), ranging from approximately three times too large to about four times too small. For the realisation of the observation used (represented by the vertical black dash-dot line in Fig. 8a), the sensitivity is much smaller than would be expected.

At observation time 7, the root mean square error (RMSE) of the Gaussian fit is better (see Fig. 7), and the Gaussian approximation to S has a smaller error range. In this case, the Gaussian approximation to S ranges from being approximately six times too large to about 1.5 times too small. The observed value at this time (vertical black dash-dot line in Fig. 8b) leads to the sensitivity being underestimated.

For the normalised relative entropy, the range of values is more uniform across the observation times. Ranging from approximately 0.6 to 1.8 at observation time 9 and 0.6–1.7 at observation time 7. However, the realisation of the observations at each of the times results in a very different error in the Gaussian approximation.

Fig. 8 can be understood to some extent using the theory for a prior given by a Gaussian mixture developed in Section 3. The goodness of fit of the simplified Gaussian mixture to the priors for these cases is summarised in Table 1; also see red lines in Fig. 5.

At observation time 9 the simplified Gaussian mixture fit to the prior has the smallest RMSE. Therefore, as expected in Section 3, the sensitivity is symmetrical about a maximum; however, it does not tend to $\sigma^2/\left(\sigma^2+\sigma_y^2\right)$, which in this case is 0.375, as expected, continuing to decrease to a much smaller value. At the observation time 7, the two-component Gaussian mixture fit is poorer, and the normalised S is no longer symmetrical as a function of observation value.

5. Conclusions and discussion

The aim of this work has been to give a detailed study of the effect of a non-Gaussian prior on the impact of the observations on the analysis. For simplicity this has been restricted to one dimension.

A non-Gaussian prior was modelled as a two-component Gaussian mixture with equal variance, allowing for skewness and bi-modality. Describing the prior in this way allowed for the sensitivity of the mean of the posterior, our analysis, to the observation to be derived analytically.

The sensitivity of the analysis was shown to be a strong function of the value of the observations and equal to the posterior variance divided by the observation error variance. This result extends to the case of all smooth priors that can be described as a Gaussian mixture and Gaussian likelihoods. This means that an observation for which the analysis is very sensitive may not necessarily be a good observation in terms of minimising the posterior variance. The difference between relative entropy and its Gaussian approximation was also shown to be a strong function of the observation value. Mutual information, however, is independent of the value of observation used and was seen to be in good agreement with its Gaussian approximation.

Applying these measures of observation impact to the PF technique for solving the Lorenz ’63 equations, it is seen that the non-Gaussian prior breaks down the agreement between these measures of observation impact because of their strong dependence on the value of the observation. Averaging these measures over observation space was shown to bring them closer to their Gaussian approximations. However, the Gaussian approximation was always seen to overestimate the averaged values. This

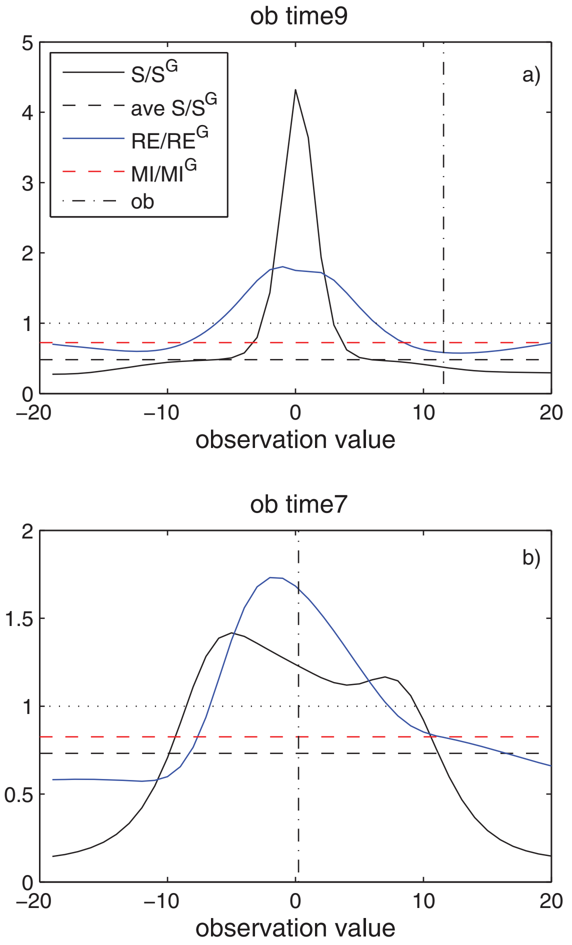

Fig. 8. S (black), MI (red dashed), RE (blue) and $\int p(y)S\,\mathrm{d}y$ (black dashed) as a function of the observation value all normalised by their Gaussian approximations. The black dashed-dot vertical line gives the realisation of the observation assimilated by the PF. Top: observation time 9. Bottom: observation time 7.

is because the Gaussian approximation to the prior underestimates the information in the prior and therefore in an averaged sense overestimates the impact of the observations.

In conclusion when calculating the observation impact in a non-Gaussian assimilation system it is important to give careful consideration to what you wish to measure:

(1) To understand the potential of new observing systems an average value of observation impact, such as mutual information, may be more useful. This can be approximated using a Gaussian assumption to the prior, giving a small overestimate when the prior is non-Gaussian.

(2) However, it can be argued that the relative entropy gives the most complete measure of observation impact and may be more useful when given a particular realisation of an observation that cannot be repeated.

(3) The sensitivity of the posterior mean to observations is a less useful measure of observation impact due to its being inversely proportional to the reduction in the posterior variance, which is often an objective in data assimilation.

Table 1. Parameters describing the simplified Gaussian mixture with two components fit to the prior at the given observation times. The last column summarises the fit as the RMSE

Observation time	w	μ_1	μ_2	σ^2	RMSE
7	0.75	-12.5	11	6	1.87×10^{-2}
9	0.5	-10	11	6	9.8×10^{-3}

In practice, when the model is non-linear, the full prior and posterior PDFs are never known and can only be sampled by expensive techniques such as the PF. A limited ensemble size makes an accurate measurement of relative entropy difficult (Haven et al., 2005). In the work of Majda et al. (2002), lower bounds are given for relative entropy using a maximum entropy approximation to the PDFs using the sample moments. The aim of Haven et al. (2005) was to give a minimum relative entropy estimate with a statistical level of confidence implied by the sample. Similar difficulty is also faced when calculating mutual information when the observation space is large (Hoffmann et al., 2006).

All conclusions made are applicable for Gaussian observation errors. Non-Gaussian observation errors due to a non-linear map between observation and state space or non-Gaussian measurement errors are left for future work.

6. Acknowledgements

This work has been funded by the National Centre for Earth Observation part of the Natural Environment Research Council, project code H5043718. The authors would also like to thank the three anonymous reviewers.

7. Appendix

A.1. Proof that MI is additive

Mutual information has the attractive quality that it is additive with successive observations. For example if at time $t+1$, a new set of observations, $\mathbf{y}^{\mathrm{new}}$, are made such that the total set of observations up to this time are given by the vector, $\mathbf{y}^{t+1} = (\mathbf{y}^{t}, \mathbf{y}^{\mathrm{new}})$. Then the mutual information given the total set of observations, MI^{t+1}, is equal to the sum of the mutual information given the previous set of observations, MI^t, and the mutual information given the new observations, $\mathrm{MI}^{\mathrm{new}}$:

$$\mathrm{MI}^{t+1} = \mathrm{MI}^t + \mathrm{MI}^{\mathrm{new}} \tag{29}$$

Proof. From the definition of mutual information [eq. (17)] we have:

$$\begin{aligned}\mathrm{MI}^{t+1} = &-\int P(\mathbf{x})\ln P(\mathbf{x})\mathrm{d}\mathbf{x}\\ &+\int\int P(\mathbf{x},\mathbf{y}^{t+1})\ln P(\mathbf{x}|\mathbf{y}^{t+1})\mathrm{d}\mathbf{x}\mathrm{d}\mathbf{y}.\end{aligned} \tag{30}$$

This may be expanded using Bayes' theorem to give:

$$\begin{aligned}\mathrm{MI}^{t+1} = &-\int P(\mathbf{x})\ln P(\mathbf{x})\mathrm{d}\mathbf{x}\\ &+\int\int P(\mathbf{x},\mathbf{y}^t,\mathbf{y}^{\mathrm{new}})\ln\left[\frac{P(\mathbf{y}^{\mathrm{new}}|\mathbf{x})P(\mathbf{y}^t|\mathbf{x})P(\mathbf{x})}{P(\mathbf{y}^{\mathrm{new}})P(\mathbf{y}^t)}\right]\mathrm{d}\mathbf{x}\mathrm{d}\mathbf{y}.\end{aligned} \tag{31}$$

The log term may then be separated:

$$\begin{aligned}\mathrm{MI}^{t+1} = &-\int P(\mathbf{x})\ln P(\mathbf{x})\mathrm{d}\mathbf{x}\\ &+\int\int P(\mathbf{x},\mathbf{y}^t,\mathbf{y}^{\mathrm{new}})\ln P(\mathbf{y}^{\mathrm{new}}|\mathbf{x})\mathrm{d}\mathbf{x}\mathrm{d}\mathbf{y}\\ &+\int\int P(\mathbf{x},\mathbf{y}^t,\mathbf{y}^{\mathrm{new}})\ln P(\mathbf{y}^t|\mathbf{x})\mathrm{d}\mathbf{x}\mathrm{d}\mathbf{y}\\ &+\int\int P(\mathbf{x},\mathbf{y}^t,\mathbf{y}^{\mathrm{new}})\ln P(\mathbf{x})\mathrm{d}\mathbf{x}\mathrm{d}\mathbf{y}\\ &-\int\int P(\mathbf{x},\mathbf{y}^t,\mathbf{y}^{\mathrm{new}})\ln P(\mathbf{y}^{\mathrm{new}})\mathrm{d}\mathbf{x}\mathrm{d}\mathbf{y}\\ &-\int\int P(\mathbf{x},\mathbf{y}^t,\mathbf{y}^{\mathrm{new}})\ln P(\mathbf{y}^t)\mathrm{d}\mathbf{x}\mathrm{d}\mathbf{y}.\end{aligned} \tag{32}$$

This can be simplified using the identity: $\int\int P(\mathbf{a},\mathbf{b})\ln P(\mathbf{b})\mathrm{d}\mathbf{a}\mathrm{d}\mathbf{b} = \int P(\mathbf{b})\ln P(\mathbf{b})\mathrm{d}\mathbf{b}$:

$$\begin{aligned}\mathrm{MI}^{t+1} = &\int\int P(\mathbf{x},\mathbf{y}^{\mathrm{new}})\ln P(\mathbf{y}^{\mathrm{new}}|\mathbf{x})\mathrm{d}\mathbf{x}\mathrm{d}\mathbf{y}^{\mathrm{new}}\\ &+\int\int P(\mathbf{x},\mathbf{y}^t)\ln P(\mathbf{y}^t|\mathbf{x})\mathrm{d}\mathbf{x}\mathrm{d}\mathbf{y}^t\\ &-\int P(\mathbf{y}^{\mathrm{new}})\ln P(\mathbf{y}^{\mathrm{new}})\mathrm{d}\mathbf{y}^{\mathrm{new}}\\ &-\int P(\mathbf{y}^t)\ln P(\mathbf{y}^t)\mathrm{d}\mathbf{y}^t.\end{aligned} \tag{33}$$

This is equal to $\mathrm{MI}^t+\mathrm{MI}^{\mathrm{new}}$ using (Cover and Thomas, 1991):

$$\begin{aligned}&\int\int P(\mathbf{x},\mathbf{y})\ln P(\mathbf{y}|\mathbf{x})\mathrm{d}\mathbf{x}\mathrm{d}\mathbf{y} - \int P(\mathbf{y})\ln P(\mathbf{y})\mathrm{d}\mathbf{y}\\ &= \int\int P(\mathbf{x},\mathbf{y})\ln P(\mathbf{x}|\mathbf{y})\mathrm{d}\mathbf{x}\mathrm{d}\mathbf{y} - \int P(\mathbf{x})\ln P(\mathbf{x})\mathrm{d}\mathbf{x}.\end{aligned} \tag{34}$$

Q.E.D.

A.2. The sensitivity of the analysis to the observations for Gaussian mixtures of arbitrary order

It can be proved that when the likelihood is Gaussian $N\left(\mu_y,\sigma_y^2\right)$ and the prior is a Gaussian mixture that the sensitivity of the mean of the posterior to the observations is equal to the analysis error variance divided by the observation error variance. That is:

$$\frac{\partial\mu_a}{\partial\mu_y} = \frac{\sigma_{x|y}^2}{\sigma_y^2}.$$

Proof. Let the prior, $p(x)$, be given by a Gaussian mixture:

$$p(x) = \sum_{i=1}^{N} w_i\left(2\pi\sigma_i^2\right)^{-(1/2)}\exp\left[-\frac{(x-\mu_i)^2}{2\sigma_i^2}\right],$$

where $\sum_{i=1}^{N} w_i = 1$. From Bayes' theorem the posterior may be similarly expressed as:

$$p(x|y) = \sum_{i=1}^{N} \tilde{w}_i\left(2\pi\tilde{\sigma}_i^2\right)^{-(1/2)}\exp\left[-\frac{(x-\tilde{\mu}_i)^2}{2\tilde{\sigma}_i^2}\right].$$

The updated weights are given by:

$$\tilde{w}_i = \frac{w_i A_i e^{-a_i}}{\sum_{j=1}^{N} w_j A_j e^{-a_j}}, \tag{35}$$

where

$$A_i = \frac{1}{\sqrt{2\pi\left(\sigma_i^2+\sigma_y^2\right)}} \text{ and } a_i = \frac{(\mu_y-\mu_i)^2}{2\left(\sigma_y^2+\sigma_i^2\right)}$$

The updated means are given by:

$$\tilde{\mu}_i = \frac{\mu_y\sigma_i^2+\mu_i\sigma_y^2}{\sigma_i^2+\sigma_y^2}. \tag{36}$$

The updated variances are given by:

$$\tilde{\sigma}_i^2 = \frac{\sigma_i^2\sigma_y^2}{\sigma_i^2+\sigma_y^2}.$$

The mean of the posterior is given by:

$$\mu_a = \sum_{i=1}^{N}\tilde{w}_i\tilde{\mu}_i. \tag{37}$$

Differentiating eq. (37) with respect to the observation value gives:

$$\frac{\partial\mu_{\mathrm{a}}}{\partial\mu_y} = \sum_{i=1}^{N}\left(\tilde{w}_i\frac{\partial\tilde{\mu}_i}{\partial\mu_y} + \frac{\partial\tilde{w}_i}{\partial\mu_y}\tilde{\mu}_i\right) \tag{38}$$

From eq. (36) $(\partial\tilde{\mu}_i/\partial\mu_y)$ can be seen to be $\left[\sigma_i^2/\left(\sigma_i^2+\sigma_y^2\right)\right]$, which is equal to $(\tilde{\sigma}_i^2/\sigma_y^2)$.

The second term requires more manipulation. Let $w_i A_i e^{-a_i}$ in eq. (35) be $\hat{w}_i$ then:

$$\tilde{w}_i = \frac{\hat{w}_i}{\sum_j \hat{w}_j}$$

and

$$\frac{\partial \tilde{w}_i}{\partial \mu_y} = \frac{\hat{w}_i' \sum_{j=1}^N \hat{w}_j - \hat{w}_i \sum_{j=1}^N \hat{w}_j'}{\left(\sum_{j=1}^N \hat{w}_j\right)^2} \tag{39}$$

where $\hat{w}_i' = (\partial \hat{w}_i / \partial \mu_y)$. We can rewrite eq. (39) using $\hat{w}_i' = -(\partial a_i / \partial \mu_y)\hat{w}_i = -a_i' \hat{w}_i$:

$$\begin{aligned} \frac{\partial \tilde{w}_i}{\partial \mu_y} &= \frac{\sum_{j=1}^N \left(\hat{w}_i' \hat{w}_j - \hat{w}_i \hat{w}_j'\right)}{\left(\sum_{j=1}^N \hat{w}_j\right)^2} \\ &= \frac{\hat{w}_i \sum_{j=1}^N w_j \left(a_j' \hat{\ } - a_i'\right)}{\left(\sum_{j=1}^N \hat{w}_j\right)^2} \\ &= \tilde{w}_i \sum_{j=1}^N \tilde{w}_j \left(a_j' - a_i'\right). \end{aligned} \tag{40}$$

Substitute $(\partial \tilde{\mu}_i / \partial \mu_y) = \left(\tilde{\sigma}_i^2 / \sigma_y^2\right)$ and eq. (40) into eq. (38):

$$\frac{\partial \mu_\mathrm{a}}{\partial \mu_y} = \sum_{i=1}^N \frac{\tilde{w}_i \tilde{\sigma}_i^2}{\sigma_y^2} + \sum_{i=1}^N \tilde{\mu}_i \tilde{w}_i \sum_{j=1}^N \tilde{w}_j \left(a_j' - a_i'\right).$$

The variance of the posterior is given by:

$$\sigma_{x|y}^2 = \int (x - \mu_\mathrm{a})^2 p(x|y) \mathrm{d}x = \sum_{i=1}^N \tilde{w}_i \tilde{\sigma}_i^2 + \sum_{i=1}^N \tilde{w}_i (\tilde{\mu}_i - \mu_\mathrm{a})^2.$$

The second term can be rewritten as:

$$\begin{aligned} \sum_{i=1}^N \tilde{w}_i (\tilde{\mu}_i - \mu_\mathrm{a})^2 &= \sum_{i=1}^N \tilde{w}_i \tilde{\mu}_i^2 - \left(\sum_{i=1}^N \tilde{w}_i \tilde{\mu}_i\right)^2 \\ &= \sum_{i=1}^N \tilde{w}_i \tilde{\mu}_i \left(\tilde{\mu}_i - \sum_{j=1}^N \tilde{w}_j \tilde{\mu}_j\right) \\ &= \sum_{i=1}^N \tilde{w}_i \tilde{\mu}_i \sum_{j=1}^N \tilde{w}_j (\tilde{\mu}_i - \tilde{\mu}_j). \end{aligned}$$

Therefore $(\partial \mu_\mathrm{a} / \partial \mu_y) = \left(\sigma_{x|y}^2 / \sigma_y^2\right)$ holds if:

$$\begin{aligned} &\sum_{i=1}^N \frac{\tilde{w}_i \tilde{\sigma}_i^2}{\sigma_y^2} + \sum_{i=1}^N \tilde{\mu}_i \tilde{w}_i \sum_{j=1}^N \tilde{w}_j \left(a_j' - a_i'\right) \\ &\quad = \frac{1}{\sigma_y^2} \left[\sum_{i=1}^N \tilde{w}_i \tilde{\sigma}_i^2 + \sum_{i=1}^N \tilde{w}_i \tilde{\mu}_i \sum_{j=1}^N \tilde{w}_j (\tilde{\mu}_i - \tilde{\mu}_j)\right]. \end{aligned}$$

Or equivalently:

$$\sum_{i=1}^N \tilde{\mu}_i \tilde{w}_i \sum_{j=1}^N \tilde{w}_j \left(a_j' - a_i' - \frac{\tilde{\mu}_i - \tilde{\mu}_j}{\sigma_y^2}\right) = 0.$$

This is true because $a_j' - a_i' - [(\tilde{\mu}_i - \tilde{\mu}_j)/\sigma_y^2] = 0$ for all i, j, since: $a_i' + \frac{\tilde{\mu}_i}{\sigma_y^2} = \frac{\mu_y}{\sigma_y^2}$.

Q.E.D.

References

Bishop, C. 2006. *Pattern Recognition and Machine Learning*. Springer, New York.

Bocquet, M. 2008. Inverse modelling of atmospheric tracers: non-Gaussian methods and second-order sensitivity analysis. *Nonlin. Process. Geophys.* **15**, 127–143.

Bocquet, M., Pires, C. A. and Wu, L. 2010. Beyond Gaussian statistical modeling in geophysical data assimilation. *Mon. Weather Rev.* **138**, 2997–3023.

Kalnay, E. 2003. *Atmospheric modeling, data assimilation and predictability*. Cambridge University Press, Cambridge.

Cardinali, C., Pezzulli, S. and Andersson, E. 2004. Influence-matrix diagnostics of a data assimilation system. *Q. J. R. Met. Soc.* **130**, 2767–2786. DOI: 10.1256/qj.03.205.

Cover, T. M. and Thomas, J. A. 1991. *Elements of Information Theory (Wiley series in Telecommunications)*. John Wiley and Sons, New York.

Evensen, G. 1997. Advanced data assimilation for strongly non-linear dynamics. *Mon. Weather Rev.* **125**, 1342–1354.

Eyre, J. E. 1990. The information content of data from satellite sounding systems: A simulation study. *Q. J. R. Met. Soc.* **116**, 401–434. DOI: 551.501.7:551.507.362.2.

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

Haven, K., Majda, A. and Abramov, R. 2005. Quantifying predictability through information theory: small sample estimation in a non-Gaussian framework. *J. Comput. Phys.* **206**, 334–362.

Hoffmann, G. M., Waslander, S. L. and Tomlin, C. J. 2006. Mutual information methods with particle filters for mobile sensor network control. In: Proceedings of the 45th IEEE Conference on Decision & Control, conference held 13–15 Dec, 2006 in San Diego, CA. Publisher is Institute for Electrical and Electronic Engineers (IEEE), USA. pp. 1019–1024.

Kleeman, R. 2002. Measuring dynamical prediction utility using relative entropy. *J. Atmos. Sci.* **59**, 2057–2072.

Kleeman, R. 2011. Information theory and dynamical system predictability. *Entopy*, **13**, 612–649. DOI: 10.3390/e13030612.

Lorenz, E. N. 1963. Deterministic nonperiodic flow. *J. Atmos. Sci.* **20**, 130–141.

Majda, A., Kleeman, R. and Cai, D. 2002. A mathematical framework for quantifying predictability through relative entropy. *Methods Appl. Anal.* **9**, 425–444.

Pires, C. A., Vautard, R. and Talagrand, O. 1996. On extending the limits of variational assimilation in nonlinear chaotic systems. *Tellus* **48A**, 96–121.

Rabier, F. 2005. Overview of global data assimilation developments in numerical weather-prediction centres. *Q. J. R. Met. Soc.* **131**, 3215–3233.

Rodgers, C. D. 2000. *Inverse Methods for Atmospheric Sounding*. World Scientific Publishing, Singapore.

Tarantola, A. 2005. *Inverse Problem Theory*. SIAM, Philadelphia.

van Leeuwen, P. J. 2009. Particle filtering in geophysical systems. *Mon. Weather Rev*. **137**, 4089–4114.

Xu, Q. 2007. Measuring information content from observations for data assimilation: relative entropy versus Shannon entropy difference. *Tellus* **59A**, 198–209.

Xu, Q., Wei, L. and Healy, S. 2009. Measuring information content from observations for data assimilation: connection between different measures and application to radar scan design. *Tellus* **61A**, 144–153.