Tellus
SERIES A
DYNAMIC
METEOROLOGY
AND OCEANOGRAPHY
PUBLISHED BY THE INTERNATIONAL METEOROLOGICAL INSTITUTE IN STOCKHOLM

Tellus (2010), 62A, 266–287

TELLUS

Mixed Gaussian-lognormal four-dimensional data assimilation

By S. J. FLETCHER*, *Cooperative Institute for Research in the Atmosphere, Colorado State University, 1375 Campus Delivery, Fort Collins, CO 80523-1375, USA*

(Manuscript received 8 July 2009; in final form 15 February 2010)

ABSTRACT

In this paper, the current methods that are used for deriving the Gaussian cost function in four-dimensional variational (4D VAR) data assimilation, are extended to lognormal and mixed, lognormal and Gaussian, random variables (rv). It is also shown that transforming a lognormal rv into a Gaussian rv to use in a Gaussian based 4D VAR results in the analysis state being similar to a median in lognormal space. An alternative version of the functional approach is derived so that the minimum of this alternative cost function is the mode in lognormal space. However, the current approaches do not allow for a generalized probability density function (pdf) approach, to overcome this a general probability model is derived so that the mode is found for any pdf from Bayesian networks and conditional independence properties. It is shown that the current Gaussian cost function and the lognormal can be found by the probability model method. The paper is finished by comparing the median approach with the modal approach in the Lorenz 1963 model for both weak and strong constraint 4D VAR and show that the mode is a more reliable analysis statistic than the median under certain circumstances.

1. Introduction

One of the advantages of four-dimensional variational (4D VAR) data assimilation (DA) in numerical weather prediction is that it allows for the comparison of observations of the atmosphere in near real time with the numerical model output. The differences between the observations and model outputs, or transformed model output, are stored as innovations to then be used in the minimization of a cost function. This has the advantage of using temporal information about different flows in the atmosphere. By using the adjoints of the tangent linear models of the numerical model and the observation operators it is possible to find new initial conditions at the start of the assimilation window, the time which the numerical model is run over and observations are collected, which are then used to minimize the innovations in the gradient of the cost function and find the best trajectory through the assimilation window.

The cost function that is minimized in 4D VAR, whilst it appears to have be derived from a conditional probability approach, as in the 3D VAR, Lorenc (1986), is actually derived from a weighted least square, or a functional, approach, Lewis and Derber (1985), Le Dimet and Talagrand (1986) and Talagrand (1988). The solution to the minimum of the cost function derived from either of these two approaches is a median, which is the unbiased statistic. For the Gaussian distribution the three descriptive statistics, mean (minimum variance), median (unbiased) and the mode (maximum likelihood) are the same. However, for left skewed distribution, for example the lognormal distribution, the three statistics are not equal, where the mode is always less than or equal to the median and the mean is always greater than or equal to the median.

As the resolution of the synoptic numerical models increase and as more smaller domain, high resolution mesoscale and cloud resolving scale data assimilation is needed, the variables involved do not conform to the continuous Gaussian assumptions which underlies weighted least-squares theory, Clarke and Cooke (2004). It has long been known that there are many positive definite variables, that is, variables which are greater than or equal to zero, Mielke et al. (1977), Miles et al. (2000) and Sengupta et al. (2002), but there are also discontinuous variables, that is, precipitation, which are important variables at these scales but do not satisfy the Gaussian assumption. In Miles et al. (2000) there is a good summary of the distributions of cloud scale variables.

An approach to extend the methods used to derive the Gaussian version of 4D VAR to non-Gaussian random variables is not obvious. One problem lies with which statistic to use to approximate the true state. The weighted least-squares approach for defining 4D VAR first appears in Lewis and Derber (1985) where a weighted least-squares method is combined with

*Corresponding author.
e-mail: fletcher@cira.colostate.edu
DOI: 10.1111/j.1600-0870.2010.00439.x

adjoint techniques to find the minimum of a cost function. The first functional/variational formulation of the observational component of 4D VAR appears in Le Dimet and Talagrand (1986) where the problem is defined for observational errors. The modern version of 4D VAR appears in Talagrand (1988). However, all of these approaches are based upon finding the median, or the unbiased estimator, but for skewed distribution this is not the most desirable statistic.

In a recent series of papers the Bayesian model for the Gaussian version of 3D VAR has been extended, first for lognormal observational error, Fletcher and Zupanski (2006a), (FZ06a hereafter), then to a combination of Gaussian and lognormal observational errors, which we will now refer to as a mixed DA system, formally the hybrid DA system, Fletcher and Zupanski (2006b), (FZ06b hereafter), and finally to lognormal-Gaussian background and observational errors, Fletcher and Zupanski (2007), (FZ07 hereafter). It is in FZ07 that the 3D VAR version of the mixed system is tested with the Lorenz'63 model, Lorenz (1963), where the z component of the model is shown not to be Gaussian distributed, nor is it lognormally distributed. However, it is shown that the primary mode is captured by a lognormal approximation.

For the 4D VAR problem there is no obvious probability model where the definition of a different distribution can be substituted for the multivariate Gaussian and then maximized to find the mode. However, in Van Leeuwen and Evensen (1996) an approach is presented but is based upon the vectorial form of Bayes theorem which is not necessarily the correct approach to model the probabilities involved in 4D VAR. In Section 4, different probability models are presented using a multiple probabilistic event version of Bayes theorem with conditional independence tools.

The remainder of the paper is as follows: In Section 3, a brief summary of the original techniques used to derive the Gaussian version of 4D VAR, which are based upon variational and weighted least-squares techniques, are presented. Also in Section 3 the weighted least squares is extended to lognormal random variables by the transform technique, and for the variational approach by defining a functional similar to the associated with Gaussian random variables. It is shown that the solutions to both approaches are equivalent but that the solution to both is a median and not the mode. This is also presented for the mixed Gaussian-lognormal case with the same conclusions. An alternative functional form is also presented in Section 3 where the associated minimums for both the lognormal case and the mixed case are shown to be the mode of the analysis distribution. In Section 5, the probability model approach is compared with the transform approach, effectively comparing the mode with the median as the analysis state, with the Lorenz'63 model for both the strong and weak constraint versions of 4D VAR. It is shown that under certain circumstances the two are almost identical but that under larger uncertainties then the mode is a more reliable statistic than the median for the strong constraint version, for the weak constraint version this conclusion appears to be valid as well but not to as large a window. This paper is finished with some conclusions and plans for future work. However, in the next section some assumptions that are made in the Gaussian 4D VAR derivation are investigated to see if they can be extended to non-Gaussian 4D VAR.

2. Assumptions

In this section, the assumptions that are made in both Gaussian and lognormal variational data assimilation are examined. The implicit assumptions used in Gaussian data assimilation are addressed and shown not to apply to lognormal random variables. Proof is also provided of the existence of non-Gaussian state variables and hence errors.

2.1. Implicit Gaussian assumptions

In the derivation of Gaussian random variable based 4D VAR the errors are defined as a difference, and as such are Gaussian distributed. However, this is using the implicit assumption that the difference between two independent Gaussian random variables is also a Gaussian random variable, Clarke and Cooke (2004), Therefore, the difference between the true state and the background state have to be independent and Gaussian distributed. This property also has to be true for the observations and the observation operator.

In Fig. 1, we present two samples randomly drawn from different lognormal distributions, plots (a) and (b), where each sample contains 20 000 points. These two samples are independent of each other and therefore satisfy the independence assumption mentioned above. In plot (c) the distribution of the difference between the two samples has been plotted. Also in plot (c) is the best Gaussian approximation that describes the distribution of the difference between the two samples, as determined by DFITTOOL in MATLAB. It is obvious from plot (c) that the difference is not Gaussian distributed. This is an important figure as it demonstrates that we have to be careful in assuming that the error, defined as a difference between two variables, will satisfy the Gaussian assumption.

There are circumstances when the lognormal distribution is very similar, in appearance, to a Gaussian distribution. This occurs when the variance is small. This can be seen by considering the skewness, or the third moment, γ_1, of the lognormal distribution. The definition for the skewness is

$$\gamma_1 = \frac{E\{[X - E(X)]^3\}}{\{E(X^2) - [E(X)]^2\}^{\frac{3}{2}}},$$

where E is the expectation operator and is defined as

$$E(X) = \int_a^b X f(X) \mathrm{d}X,$$

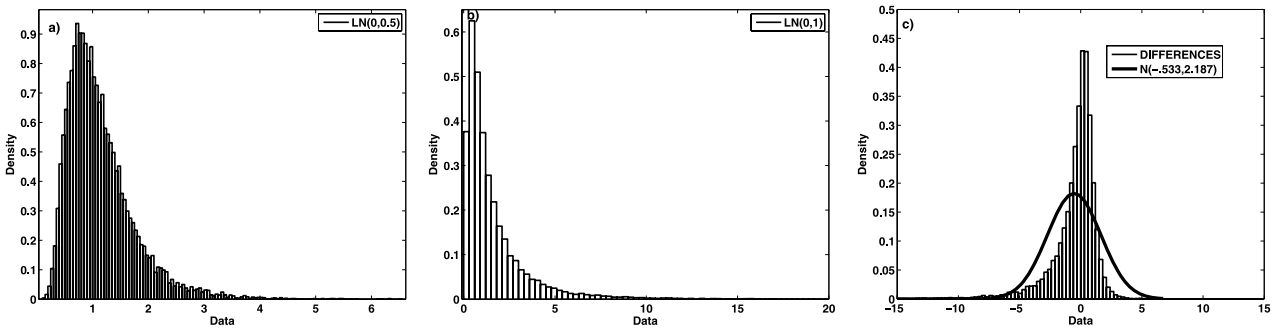

Fig. 1. Plot (a) Lognormal sample from LN(0, 0.5), Plot (b) Lognormal sample from LN(0, 1), Plot (c) Distribution of the difference and the nearest Gaussian with $\mu = -0.533$, $\sigma = 2.187$.

where $f(X)$ is the probability density function for the random variable X.

For the univariate lognormal distribution the skewness is

$$\gamma_1 = (\exp\{\sigma^2\} + 2)\sqrt{(\exp\{\sigma^2\} - 1)}, \tag{1}$$

where $\sigma^2 = E[(\ln X)^2] - [E(\ln X)]^2$.

An important feature to note about (1) is that it is only a function of the variance, σ^2, therefore as $\sigma^2 \to 0$ then $\gamma_1 \to 0$, which implies that the distribution is becoming more symmetric. As a result of the skewness tending to zero then the three descriptive statistics, mean, mode and median, are tending to the same value, as can be seen from there definitions below

$$x_{\text{mod}} = \exp\{\mu - \sigma^2\},$$

$$x_{\text{med}} = \exp\{\mu\},$$

$$x_{\text{mea}} = \exp\left\{\mu + \frac{\sigma^2}{2}\right\}.$$

However, it is not guaranteed that the variance of the underlying lognormal random variables will be small enough for the two distributions to almost be similar for all the different types of flows associated with numerical weather prediction. In Fig. 2, this closeness between the two distributions is demonstrated (plots a and b), also shown is the distribution of the differences between the two samples which is approximately Gaussian.

Therefore, when defining the error as a difference between two variables, if these variables are lognormally distributed then their difference is not distributed as a Gaussian, but is approximately a Gaussian random variable if the variance of the lognormal variables is sufficiently small.

2.2. *Lognormal observations*

In this section observational evidence of non-Gaussian random variables are presented and the implications for the Gaussian error assumption are explained. The observations that are considered are taken from the Atmospheric Radiation Measurement (ARM) at the Southern Great Plains (SGP) site in Oklahoma. The observations were taken with a dual-channel microwave radiometer between 1997 and 2000. The plots in Fig. 3 are of the distribution of column water vapour (CWV) when a boundary layer clouds is present. In each plot the best Gaussian and lognormal fits to the data has been calculated by MATLAB. Each sample contains about 4000 data points.

An important feature to note about the plots is that the Gaussian approximation in three of the four plots (top left and right and bottom right plots) allows for a negative value for a positive definite variable. Another feature to note is the seasonal dependency of the variable. It is clear from the autumn plot (September, October and November, SON) that there is a transition phase from the wet summer months, as seen in the summer

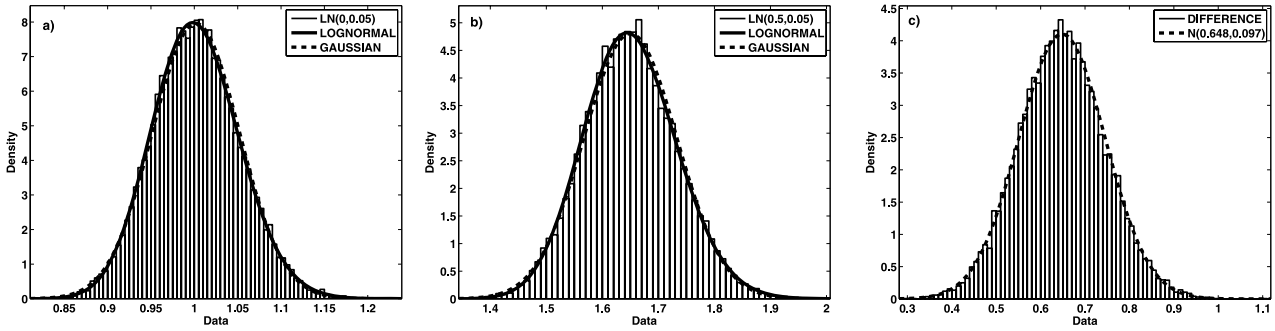

Fig. 2. Plot (a) Lognormal sample from LN(0, 0.05) , Plot (b) Lognormal sample from LN(0.5, 0.05), Plot (c) Distribution of the difference which is approximated by a Gaussian, $\mu = 0.648$, $\sigma = 0.097$.

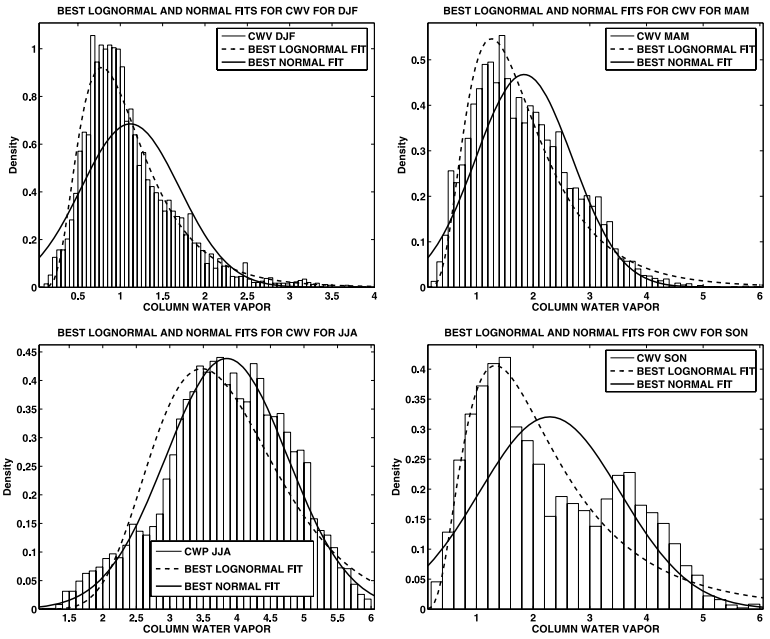

Fig. 3. Histograms of Column Water Vapour (cwv) from the ARM SGP site for the four seasons: Winter (top left-hand panel), Spring (top right-hand panel), Summer (bottom left-hand panel) and Autumn (bottom right-hand panel). In each plot is the best Gaussian and lognormal approximations.

plot (JJA), to the drier winter months (DJF). It can also be seen from Fig. 3 that the Gaussian approximation overestimates the true mode for three out of the four seasons, and in the summer months the lognormal approximation is not that much smaller that the true mode.

The reason for presenting this data is to highlight that if CWV acts similar to a lognormal then due to its calculation as a linear function of humidity, which is a control variable in most synoptic and mesoscale 4D VAR systems, then there is implications for the assimilation of humidity as a Gaussian random variable. CWV is calculated as a sum of humidities multiplied by some scaling constants for each vertical level. It is the sum of the different scaled humidities that plays the significant part. There is an important property of the Gaussian distribution that informs us that relative humidity for certain seasons over Oklahoma is not Gaussian distributed. This property is that for independent Gaussian random variables then the sum of these random variables is also a Gaussian random variable. However, the converse is also true, if you have independent random variables and their sum is a Gaussian random variables then they themselves are Gaussian random variables (Cramèr Theorem).

Therefore, if the sum is not a Gaussian then there exist at least one vertical level where the humidity is not Gaussian. The property also holds for a linear scaling, Clarke and Cooke (2004). At the time of writing there is no known analytical proof that the distribution of the sum of lognormal random variables is lognormal, but numerical proofs of this property do exist, Beaulieu and Rajwani (2004).

It is important to note that as a result of CWV being a linear function of a background state variable then these observations are implying that the state variable that is used to match this data is not always Gaussian distributed. Therefore as a consequence of the sum property we have to consider a non-Gaussian background component, not just for the observational component.

3. Functional form for lognormal/mixed lognormal-Gaussian 4D VAR

This section is comprised of three parts. The first part is a brief overview of the derivation of the Gaussian version of 4D VAR. The second section deals with extending these ideas to lognormal random variables, and also to mixed lognormal-Gaussian random variables (FZ06b) where it is shown that the maximum likelihood state that is found in the extension of the Gaussian framework is a median of the lognormal, or mixed distribution. In the final section a different functional form is presented whose minimum is the mode of the lognormal (or mixed) distribution.

3.1. Overview of the derivation of Gaussian 4D VAR

The earliest form of the mathematical ideas behind the modern 4D VAR appears in Lewis and Derber (1985). The approach described in Lewis and Derber (1985) is to find the minimum of the cost function

$$J(\mathbf{x}(t_0)) = \frac{1}{2}\sum_{i=0}^{t_a} \langle \mathbf{W}(t_i)[\mathbf{x}(t_i) - \mathbf{x_b}(t_i]), \quad \mathbf{x}(t_i) - \mathbf{x_b}(t_i)\rangle, \tag{2}$$

where t_a is the analysis time, $\mathbf{W}$ is a *weight* matrix which can be changed depending on known accuracies, the expression $\langle \cdot , \cdot \rangle$ is the inner product operator, $\mathbf{x}_b$ is the output from a numerical model which has been started by some set of initial conditions,

$\mathbf{x}_{b,0}$ and $\mathbf{x}$ is the analyses that has come from a simpler version of data assimilation, that is, optimal interpolation (OI). The problem is to seek the initial conditions that minimizes the weighted squared differences between the original analysis from the OI scheme at several times and the coincident solutions to the numerical model. Note: in later formulations $\mathbf{W}$ becomes the inverse of the background error covariance matrix, $\mathbf{B}$.

The minimum of (2) is found through its gradient, ∇J, with respect to the initial conditions. To find the minimum of (2) an adjoint approach is used. This approach starts by considering the first order change to (2) from a small perturbation, $\mathbf{x}'(t_0)$, about the initial conditions $\mathbf{x}(t_0)$. Therefore, J' is the first order change in the functional and is related to the directional derivative in $\mathbf{x}'(t_0)$ by

$$J' = \langle \nabla J, \mathbf{x}'(t_0) \rangle. \tag{3}$$

Substituting the information from (2) into (3) and introducing a linearized perturbation model, the reader is referred to Lewis and Derber (1985) for more details about this, and by using the property of adjoints,

$$\langle \mathbf{x}, \mathbf{A}\mathbf{y} \rangle = \langle \mathbf{A}^T\mathbf{x}, \mathbf{y} \rangle, \tag{4}$$

then the gradient can be expressed as

$$\nabla J = \sum_{i=0}^{t_a} \left(\left[\prod_{k=0}^{K_i} \mathbf{D}^T(t_0 + k\Delta t) \right] \hat{\mathbf{x}}_i(t_i) \right), \tag{5}$$

where $\mathbf{D}$ is the matrix containing the coefficient of the discrete approximation to the linearized perturbation equation and K_i represents the total number of time steps from t_0 to t_i, where $i = 1, 2, \ldots, K$ where K is the total number of time steps to t_a, and $\hat{\mathbf{x}}_i(t_i) = \mathbf{W}(t_i)(\mathbf{x}(t_i) - \mathbf{x_b}(t_i))$.

The extension of the adjoint techniques from Lewis and Derber (1985) to an observational base approach appears in Le Dimet and Talagrand (1986). The approach in Le Dimet and Talagrand (1986) is based upon calculus of variations techniques and optimal control theory. The starting point in Le Dimet and Talagrand (1986) is the state vector, $\mathbf{x}$, that is defined in time by the equation

$$\frac{\mathrm{d}\mathbf{x}}{\mathrm{d}t} = \boldsymbol{\mathcal{M}}(\mathbf{x}), \tag{6}$$

where $\boldsymbol{\mathcal{M}}$ is a continuous non-linear model operator acting on $\mathbf{x}$.

It is assumed that there are sets of observations at different times, denoted by $\mathbf{y}(t_1), \mathbf{y}(t_2), \ldots, \mathbf{y}(t_a)$. Given these observations, it is possible to define a functional in terms of the observations as

$$J(\mathbf{x}(t)) = \sum_{i=1}^{t_a} \langle \mathbf{x}(t_i) - \mathbf{y}(t_i), \mathbf{x}(t_i) - \mathbf{y}(t_i) \rangle. \tag{7}$$

However, there is the problem that the observations do not exactly match the state vector and therefore the gradient of (7) can not be assumed to be exactly zero.

As in Lewis and Derber (1985) it is the initial conditions to (6), such that (7) is minimized, that are required. To obtain this goal the first variation of (7) with respect to a small perturbation to the initial conditions, $\delta\mathbf{x}(t_0)$, is considered. Following the techniques summarized above with the continuous perturbation equation given by

$$\frac{\mathrm{d}\delta\mathbf{x}}{\mathrm{d}t} = \mathbf{A}(t)\delta\mathbf{x}, \tag{8}$$

where (8) is started from initial conditions $\delta\mathbf{x}(t_0)$ and

$$A(t) = \frac{\partial \boldsymbol{\mathcal{M}}[\mathbf{x}(t)]}{\partial \mathbf{x}(t)},$$

is the Jacobian matrix of the model equations, then the gradient is given by

$$\nabla J = 2 \sum_{i=0}^{t_a} \mathbf{M}^T(t_i, t_0)[\mathbf{x}(t_i) - \mathbf{y}(t_i)], \tag{9}$$

where the fact that (8) is a linear equation for $\delta\mathbf{x}$ has been used and therefore the solution at $t = t_i$ depends linearly on $\delta\mathbf{x}(t_0)$. The linear operator in (9) is referred to as the *resolvent* between t_0 and t_i and is denoted by $\mathbf{M}(t, t_0)$ and $\mathbf{M}^T(t_i, t_0)$ is its adjoint.

The more advance observation operator version of (7) appears in Talagrand (1988) where the cost function is defined by

$$J[\mathbf{x}(t_i)] = \frac{1}{2} \sum_{i=1}^{t_a} \langle \mathbf{y} - \mathbf{h}_i[\mathbf{x}(t_i)], \mathbf{y}(t_i) - \mathbf{h}_i[\mathbf{x}(t_i)] \rangle. \tag{10}$$

Given that the observation operator $\mathbf{h}_i(\cdot)$ is a function of the model state at t_i then a discrete version of the dynamic model equations is required, which is

$$\mathbf{x}_{i+1} = \mathbf{x}_i + \Delta t \boldsymbol{\mathcal{M}}(\mathbf{x}_i). \tag{11}$$

To find the minimum of (10), the first variations of (10) and (11) with respect to the initial conditions to (11) are required. By using the techniques summarized above as well as the adjoint property (4) results in

$$\nabla_{\mathbf{x}_0} J = \sum_{i=1}^{t_a} \left(\mathbf{I} + \Delta t \mathbf{M}_0^T\right) \left(\mathbf{I} + \Delta t \mathbf{M}_{0,1}^T\right) \ldots \left(\mathbf{I} + \Delta t \mathbf{M}_{0,i-1}^T\right) \times \mathbf{H}_i^T(\mathbf{h}(\mathbf{x}_i) - \mathbf{y}_i), \tag{12}$$

where

$$\mathbf{H}_i = \frac{\partial \mathbf{h}(\mathbf{x}_i)}{\partial \mathbf{x}_i} \quad \text{and} \quad \mathbf{M}_i = \frac{\partial \boldsymbol{\mathcal{M}}_{0,i}(\mathbf{x}_0)}{\partial \mathbf{x}_0},$$

are the tangent linear models of the observation operator and the non-linear model, respectively.

In the three approaches summarized above the final expressions are independent of the variances of the observational errors, nor do they contain a background error term. The observational error is introduced in Courtier and Talagrand (1990). The background component is speculated upon in Thépaut and Courtier (1991) and then applied in Thépaut et al. (1993) as an extra

constraint in the functional formulation, which leads to

$$J(\mathbf{x}_0) = \frac{1}{2}(\mathbf{x}_0 - \mathbf{x}_{b,0})^T \mathbf{B}_0^{-1}(\mathbf{x}_0 - \mathbf{x}_{b,0}) + \frac{1}{2}\sum_{i=1}^{N_o}(\mathbf{y}_i - \mathbf{h}_i(\boldsymbol{\mathcal{M}}_{0,i}(\mathbf{x}_0)))^T \times \mathbf{R}_i^{-1}(\mathbf{y}_i - \mathbf{h}_i(\boldsymbol{\mathcal{M}}_{0,i}(\mathbf{x}_0))), \quad (13)$$

where N_o is the total number of sets of observations.

3.2. Weighted least-squares approach to lognormal and mixed distributed errors

3.2.1. Lognormal background errors. One of the main assumptions in a least-squares minimization approach is that the residuals (errors) are Gaussian distributed, Clarke and Cooke (2004). Given this restriction it may appear that the least-squares approach cannot be extended to lognormal variables, but there exist a property of the lognormal distribution that enables this approach to be considered. This property is that the natural logarithm of a lognormal random variables is a Gaussian random variable. Therefore, if the random variable $\mathbf{x}$ is distributed $\mathbf{x} \sim LN(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ then $\mathbf{X} = \ln \mathbf{x} \sim G(\boldsymbol{\mu}, \boldsymbol{\Sigma})$.

As a result of the transform linking lognormal random variables with Gaussian random variables then it is possible to define a functional of the form

$$J(\mathbf{X}) = \frac{1}{2}\left\langle \mathbf{B}_0^{-1}(\mathbf{X}_0 - \mathbf{X}_{b,0}), (\mathbf{X}_0 - \mathbf{X}_{b,0})\right\rangle + \frac{1}{2}\sum_{i=1}^{N_o}\left\langle \mathbf{R}_i^{-1}(\mathbf{y}_i - \mathbf{h}_i(\mathcal{M}_{0,i}(\mathbf{x}_0))), (\mathbf{y}_i - \mathbf{h}_i(\mathcal{M}_{0,i}(\mathbf{x}_0)))\right\rangle. \quad (14)$$

Following a similar variational approach as outline in Section 3.1 and using the adjoint properties, the gradient of (14) can easily be shown to be

$$\nabla_{\mathbf{X}_0} J = \mathbf{B}_0^{-1}(\mathbf{X}_0 - \mathbf{X}_{b,0}) - \mathbf{W}_B^T \sum_{i=1}^{N_o} \mathbf{M}_i^T \mathbf{H}_i^T \mathbf{R}_i^{-1}(\mathbf{y}_i - \mathbf{h}_i(\boldsymbol{\mathcal{M}}_{0,i}(\mathbf{x}_0))), \quad (15)$$

where

$$\mathbf{W}_B \equiv \frac{\partial \mathbf{x}_0}{\partial \mathbf{X}_0}.$$

Therefore the non-linear solution to (15) when set to zero, for $\mathbf{x}_0$ is

$$\mathbf{x}_0 = \mathbf{x}_{b,0} \exp\left\{\mathbf{B}_0 \mathbf{W}_B^T \sum_{i=1}^{N_o} \mathbf{M}_i^T \mathbf{H}_i^T \mathbf{R}_i^{-1}\{\mathbf{y}_i - \mathbf{h}_i[\boldsymbol{\mathcal{M}}_{0,i}(\mathbf{x}_0)]\}\right\}. \quad (16)$$

An important property of (16) is that this is also the solution to the minimum of the functional

$$J(\mathbf{x}_0) = \frac{1}{2}\left\langle \mathbf{B}_0^{-1}(\ln \mathbf{x}_0 - \ln \mathbf{x}_{b,0}), (\ln \mathbf{x}_0 - \ln \mathbf{x}_{b,0})\right\rangle + \frac{1}{2}\sum_{i=1}^{N_o}\left\langle \mathbf{R}_i^{-1}(\mathbf{y}_i - \mathbf{h}_i(\boldsymbol{\mathcal{M}}_{0,i}(\mathbf{x}_0))), (\mathbf{y}_i - \mathbf{h}_i(\boldsymbol{\mathcal{M}}_{0,i}(\mathbf{x}_0)))\right\rangle, \quad (17)$$

with respect to $\mathbf{x}_0$. This is important as the minimum of (17) is a median, and not the mode, in lognormal space. To prove that (17) is a median we have to use the fact that the most common median of the multivariate lognormal distribution is $\exp\{\boldsymbol{\mu}\}$. This is an extension of the univariate case given in FZ06a. It can be shown that the minimum of the functional

$$J(\mathbf{x}_0) \equiv \frac{1}{2}\langle \boldsymbol{\Sigma}^{-1}(\ln \mathbf{x} - \boldsymbol{\mu}), (\ln \mathbf{x} - \boldsymbol{\mu})\rangle, \quad (18)$$

is $\exp\{\boldsymbol{\mu}\}$ which is a median of the multivariate lognormal distribution. Equation (18) arises from considering the log-likelihood approach to the multivariate lognormal distribution

$$LN(\mathbf{x}, \boldsymbol{\mu}, \boldsymbol{\Sigma}) = (2\pi)^{-\frac{N_s}{2}} |\boldsymbol{\Sigma}|^{-\frac{1}{2}} \prod_{i=1}^{N_s}\left(\frac{1}{x_i}\right) \times \exp\{(\ln \mathbf{x} - \boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1}(\ln \mathbf{x} - \boldsymbol{\mu})\}, \quad (19)$$

where $\boldsymbol{\mu} = E(\ln \mathbf{x})$ and Σ is the covariance matrix for $\ln \mathbf{x}$ **not** $\mathbf{x}$ and $N_s = \dim(\mathbf{x})$ with the product term omitted.

The mode of the multivariate lognormal distribution can be found by minimizing the cost function

$$J(\mathbf{x}_0) = \frac{1}{2}\langle \boldsymbol{\Sigma}^{-1}(\ln \mathbf{x} - \boldsymbol{\mu}), (\ln \mathbf{x} - \boldsymbol{\mu})\rangle + \langle \mathbf{1}_{N_s}, (\ln \mathbf{x} - \boldsymbol{\mu})\rangle, \quad (20)$$

whose solution is $\exp\{\boldsymbol{\mu} - \boldsymbol{\Sigma}\mathbf{1}\} \equiv \exp\{\boldsymbol{\mu}\}\exp\{-\boldsymbol{\Sigma}\mathbf{1}\}$ where $\mathbf{1}$ is a vector of 1s of dimension N_s. Therefore the mode is the product of the median with a decaying exponential with respect to variance. The significance of this result is that the median approach is equivalent to the transform method that is used in operational systems to overcome non-Gaussian random variables, Polavarapu et al. (2005).

3.2.2. Lognormal observational errors. The same transform that is used for the background component can also be used for the observation component. Therefore the functional for the transformed observations errors is

$$J_o(\mathbf{x}_0) = \frac{1}{2}\sum_{i=1}^{N_o}\left\langle \mathbf{R}_i^{-1} \ln \frac{\mathbf{y}_i}{\mathbf{h}_i(\boldsymbol{\mathcal{M}}_{0,i}(\mathbf{x}_0))}, \ln \frac{\mathbf{y}_i}{\mathbf{h}_i(\boldsymbol{\mathcal{M}}_{0,i}(\mathbf{x}_0))}\right\rangle, \quad (21)$$

where the ratio in the logarithm is componentwise.

To find the minimum of (21) the same first variation techniques are used here but now the composite chain rule

$$\frac{\mathrm{d}f(g(h(u(\mathbf{x}))))}{\mathrm{d}\mathbf{x}} = \frac{\mathrm{d}f(g(h(u(\mathbf{x}))))}{\mathrm{d}g}\frac{\mathrm{d}g(h(u(\mathbf{x})))}{\mathrm{d}h}\frac{\mathrm{d}h(u(\mathbf{x}))}{\mathrm{d}u} \times \frac{\mathrm{d}u(\mathbf{x})}{\mathrm{d}\mathbf{x}}. \tag{22}$$

is used.

Equation (22) is required to calculate the Taylor series expansion of $\ln \frac{\mathbf{y}_i}{\mathbf{h}_i[\boldsymbol{\mathcal{M}}(\mathbf{x}_0)]}$. Therefore the expressions for the functions in (22) are

$$\begin{array}{ll} f = \ln g & \frac{\mathrm{d}f}{\mathrm{d}g} = \frac{1}{g} \\ g = \frac{\mathbf{y}_i}{h} & \frac{\mathrm{d}g}{\mathrm{d}h} = -\frac{\mathbf{y}_i}{h^2} \\ h = \mathbf{h}_i(u) & \frac{\mathrm{d}h}{\mathrm{d}u} = \mathbf{H}_i \\ u = \boldsymbol{\mathcal{M}}_{0,i}(\mathbf{x}_0) & \frac{\mathrm{d}u}{\mathrm{d}\mathbf{x}_0} = \mathbf{M}_i(\mathbf{x}_0) \end{array},$$

which results in

$$\ln \frac{\mathbf{y}_i}{\mathbf{h}_i(\boldsymbol{\mathcal{M}}_{0,i}(\mathbf{x}_0 + \delta\mathbf{x}_0))} \approx \ln \frac{\mathbf{y}_i}{\mathbf{h}_i(\boldsymbol{\mathcal{M}}_{0,i}(\mathbf{x}_0))} - \mathbf{W}_{o,i}\mathbf{H}_i(\boldsymbol{\mathcal{M}}_i(\mathbf{x}_0))\mathbf{M}_i(\mathbf{x}_0)\delta\mathbf{x}_0, \tag{23}$$

where

$$\mathbf{W}_{o,i} \equiv \frac{\partial \ln \mathbf{h}_i(\boldsymbol{\mathcal{M}}_i(\mathbf{x}_0))}{\partial \mathbf{x}_i} = diag\left\{\frac{1}{\mathbf{h}_i(\boldsymbol{\mathcal{M}}_i(\mathbf{x}_0))}\right\}. \tag{24}$$

Using the same arguments for adjoints from earlier results, then the gradient with respect to $\mathbf{X}_0$ is

$$\nabla_{\mathbf{x}_0} J = -\sum_{i=1}^{N_o} \mathbf{M}_i^T \mathbf{H}_i^T \mathbf{W}_{o,i}^T \mathbf{R}_i^{-1} \ln \frac{\mathbf{y}_i}{\mathbf{h}_i(\boldsymbol{\mathcal{M}}_{0,i}(\mathbf{x}_0))}. \tag{25}$$

The solution to (25) when set to zero is a non-linear function of $\mathbf{x}_0$ due to $\mathbf{h}_i(\boldsymbol{\mathcal{M}}_{0,i}(\mathbf{x}_0))$. Therefore combined the two gradients (25) and (16) and setting them to zero results in the solution

$$\mathbf{x}_0 = \mathbf{x}_{b,0} \exp\left\{-\mathbf{B}_0\mathbf{W}_B^T \sum_{i=1}^{N_o} \mathbf{M}_i^T \mathbf{H}_i^T \mathbf{W}_{o,i}^T \mathbf{R}_i^{-1} \ln \frac{\mathbf{y}_i}{\mathbf{h}_i(\boldsymbol{\mathcal{M}}_{0,i}(\mathbf{x}_0))}\right\}, \tag{26}$$

which is combining the median state of the background error distribution with the median state of the observational error distribution.

3.2.3. Mixed distributed background and observational errors. In FZ06b a multivariate probability density function (pdf) is defined which describes the probability of a random vector where some components are Gaussian distributed and others that are lognormally distributed. Given this distribution the associated 3D VAR cost function is also derived in FZ06b for the case when there are observational errors with both Gaussian and lognormal components. This approach is extended to the background error case in a 3D VAR framework in FZ07.

The definition of a mixed distribution variate is

$$\mathbf{x} \equiv \begin{pmatrix} \mathbf{x}_p \\ \mathbf{x}_q \end{pmatrix}, \quad \text{where} \quad \begin{array}{l} \mathbf{x}_p \sim G(\boldsymbol{\mu}_p, \boldsymbol{\Sigma}_p) \\ \mathbf{x}_q \sim LN(\boldsymbol{\mu}_q, \boldsymbol{\Sigma}_q) \end{array}, \tag{27}$$

and p is the total number of Gaussian variates and q is the total number of lognormal variates. As with the lognormal cases it is possible to use the transform approach for the lognormal components of a mixed distributed random variable. However, following the arguments above again leads to a median and not the mode, therefore an alternative formulation is required for both the lognormal and mixed distributed random variables.

3.3. Functional approach for lognormal/mixed distribution 4D VAR

In the previous subsection it was shown that if the Gaussian method is extended to lognormal and mixed distributed random variables through the natural logarithm transform, then the state that is the minimum of the cost function is the median. However, for lognormal, and mixed, random variables it is possible to define a functional whose minimum is the mode.

3.3.1. Lognormal background errors. In the last subsection the definition of the multivariate lognormal distribution was presented, (19). It was shown that the product term in (19) is what makes the difference between the mode and the median. Therefore, a functional defined for lognormal errors, and also mixed distributed errors, must contain a term, which resembles the product term when the negative logarithm has been applied. The equivalent functional for lognormal background errors is

$$J(\mathbf{x}) = \frac{1}{2}\left\langle \mathbf{B}_0^{-1}(\ln \mathbf{x}_0 - \ln \mathbf{x}_{b,0}) + \mathbf{2}_{N_s}, (\ln \mathbf{x}_0 - \ln \mathbf{x}_{b,0})\right\rangle + \frac{1}{2}\sum_{i=1}^{N_o}\left\langle \mathbf{R}_i^{-1}(\mathbf{y}_i - \mathbf{h}_i(\boldsymbol{\mathcal{M}}_{0,i}(\mathbf{x}_0))), (\mathbf{y}_i - \mathbf{h}_i(\boldsymbol{\mathcal{M}}_{0,i}(\mathbf{x}_0)))\right\rangle, \tag{28}$$

where $\mathbf{2}$ is a vectors of 2 s of dimension N_s.

Using the methods as outlined earlier for the first variation and the properties of adjoint, then the gradient of (28) is

$$\nabla_{\mathbf{x}_0} J = \mathbf{W}_B^{-T}\left[\mathbf{B}_0^{-1}(\ln \mathbf{x}_0 - \ln \mathbf{x}_{b,0}) + \mathbf{1}\right] - \sum_{i=1}^{N_o} \mathbf{M}_i^T \mathbf{H}_i^T \mathbf{R}_i^{-1}(\mathbf{y}_i - \mathbf{h}_i(\boldsymbol{\mathcal{M}}_{0,i}(\mathbf{x}_0))), \tag{29}$$

where the superscript $-T$ represents the inverse of the transpose of the matrix and

$$\mathbf{W}_B^{-1} \equiv \frac{\partial \ln \mathbf{x}_0}{\partial \mathbf{x}_0} \equiv \frac{\partial \mathbf{X}_0}{\partial \mathbf{x}_0}.$$

Setting (29) to zero, rearranging and taking the exponential results in the non-linear solution to (28) as

$$\mathbf{x}_0 = \mathbf{x}_{b,0} \exp\{-\mathbf{B}_0\mathbf{1}\} \times \exp\left\{\mathbf{W}_B^T \sum_{i=1}^{N_o} \mathbf{M}_i^T \mathbf{H}_i^T \mathbf{R}_i^{-1}(\mathbf{y}_i - \mathbf{h}_i(\boldsymbol{\mathcal{M}}_{0,i}(\mathbf{x}_0)))\right\}. \tag{30}$$

The main difference between (30) and (16) is the term $\exp\{-\mathbf{B}_0^T\mathbf{1}\}$ which makes (30) resemble the mode of a multivariate lognormal distribution.

3.3.2. Lognormal observational errors. In the case when the observational errors are lognormally distributed then the functional for the mode is

$$J(\mathbf{x}_0) = \frac{1}{2}\sum_{i=1}^{N_o}\left\langle \mathbf{R}_i^{-1}\left(\ln\frac{\mathbf{y}_i}{\mathbf{h}_i(\boldsymbol{\mathcal{M}}_{0,i}(\mathbf{x}_0))} + \mathbf{R}_i\mathbf{2}_{N_i}\right), \ln\frac{\mathbf{y}_i}{\mathbf{h}_i(\boldsymbol{\mathcal{M}}(\mathbf{x}_0))}\right\rangle, \tag{31}$$

where $\mathbf{2}_{N_i}$ is a column vector of 2s of dimension N_i where N_i is the number of observations at $t = t_i$.

Following the derivations as for the other functional from earlier, and using the properties of adjoints, results in the gradient for (31) as

$$\nabla J_{\mathbf{x}_0} = -\sum_{i=1}^{N_o}\mathbf{M}_i^T\mathbf{H}_i^T\mathbf{W}_{o,i}^T\mathbf{R}_i^{-1}\left(\ln\frac{\mathbf{y}_i}{\mathbf{h}_i(\boldsymbol{\mathcal{M}}_i(\mathbf{x}_0))} + \mathbf{R}_i\mathbf{1}_{N_i}\right). \tag{32}$$

The solution of (32), when combined with the lognormal background component of (28), is

$$\mathbf{x}_0 = \mathbf{x}_{b,0}\exp\{\mathbf{B}_0\mathbf{I}_{N_s}\}\exp\left\{-\mathbf{W}_B^T\sum_{i=1}^{N_o}\mathbf{M}_i^T\mathbf{H}_i^T\mathbf{W}_{o,i}^T\mathbf{R}_i^{-1} \times\left(\ln\frac{\mathbf{y}_i}{\mathbf{h}_i\,(\boldsymbol{\mathcal{M}}_i(\mathbf{x}_0))} + \mathbf{R}_i\mathbf{1}_{N_i}\right)\right\}, \tag{33}$$

which is the mode, as detailed in FZ07 but extended to all of the innovations.

3.3.3. Mixed distributed background errors. To obtain the mode of the mixed distribution then the functional is defined as

$$J(\mathbf{x}) = \frac{1}{2}\left\langle \mathbf{B}_0^{-1}\begin{pmatrix}\mathbf{x}_{p,0}-\mathbf{x}_{p,b,0}\\ \ln\mathbf{x}_{q,0}-\ln\mathbf{x}_{q,b,0}\end{pmatrix} + 2\begin{pmatrix}\mathbf{0}_p\\ \mathbf{1}_q\end{pmatrix}, \begin{pmatrix}\mathbf{x}_{p,0}-\mathbf{x}_{p,b,0}\\ \ln\mathbf{x}_{q,0}-\ln\mathbf{x}_{q,b,0}\end{pmatrix}\right\rangle + \frac{1}{2}\sum_{i=1}^{N_o}\left\langle \mathbf{R}_i^{-1}(\mathbf{y}_i-\mathbf{h}_i(\boldsymbol{\mathcal{M}}_{0,i}(\mathbf{x}_0))), (\mathbf{y}_i-\mathbf{h}_i(\boldsymbol{\mathcal{M}}_{0,i}(\mathbf{x}_0)))\right\rangle. \tag{34}$$

The gradient and the non-linear solution of (34) can easily be shown to be (29) and (30) with $\mathbf{1}$ replaced by $(\mathbf{0}_p\ \mathbf{1}_q)^T$, respectively. It can also easily be verified that the solution is the mode of the hybrid distribution.

3.3.4. Mixed distributed observational errors. The functional for hybrid observation errors is very similar to (31) but now the vector $\mathbf{1}_{N_i}$ is replaced with $(\mathbf{0}_{p,i}\ \mathbf{1}_{q,i})^T$, therefore the functional is defined as

$$J(\mathbf{x}_0) = \frac{1}{2}\sum_{i=1}^{N_o}\left\langle \mathbf{R}_i^{-1}\left(\ln\frac{\mathbf{y}_i}{\mathbf{h}_i(\boldsymbol{\mathcal{M}}_{0,i}(\mathbf{x}_0))} + 2\mathbf{R}_i\begin{pmatrix}\mathbf{0}_{p,i}\\ \mathbf{1}_{q,i}\end{pmatrix}\right), \ln\frac{\mathbf{y}_i}{\mathbf{h}_i(\boldsymbol{\mathcal{M}}(\mathbf{x}_0))}\right\rangle. \tag{35}$$

The gradient and the non-linear solutions of (35), combined with the mixed distributed background error functional, is the same as (32) and (33) but with $\mathbf{2}_i$ replaced with $(\mathbf{0}_{p,i}\ \mathbf{1}_{q,i})^T$, respectively.

The functionals that have been presented in this section come from being able to know in advance how to write the pdfs of the errors in functional form to find the mode. It may not always be possible to know this form and as such in the next section a probabilistic approach is presented using the multi-event version of Bayes theorem along with conditional independence properties to find a general version of the cost function, which is independent of any specific distribution, whose minimum is the mode. However, this section is finished with a brief explanation of why the mean is not considered as the analysis statistic, for 3D VAR this is addressed in Fletcher and Zupanski (2006a).

3.4. Functional formulation for the mean

So far in the derivations presented the mode and the median have been considered. However, it is possible to solve a functional whose minimum is the mean of a multivariate lognormal distribution. There is a problem with considering the mean as the statistic that a lognormal or mixed data assimilation system is based upon. This problem is the property that this statistic is unbounded with respect to the variance, Fletcher and Zupanski (2006a), A second drawback for this statistic in general is that the mean of a multivariate distribution is calculated as a vector of the mean of each random variable and as such does not allow for the covariance between the variables.

The starting point for the functional is to considered the mean for a univariate lognormal distribution, this is $\exp\{\mu + \frac{\sigma^2}{2}\}$, which can easily be verified as the solution to the integral

$$E(X) = \frac{1}{\sqrt{2\pi}\sigma}\int_0^\infty X\frac{1}{X}\exp\left\{-\frac{1}{2}\frac{(\ln X-\mu)^2}{\sigma^2}\right\}\mathrm{d}X. \tag{36}$$

The multivariate functional form for the mean of a multivariate lognormal distribution is

$$J(\mathbf{x}) = \frac{1}{2}\langle diag(\boldsymbol{\Sigma}^{-1})(\ln\mathbf{x}-\boldsymbol{\mu}) - \mathbf{1}_{N_s}, (\ln\mathbf{x}-\boldsymbol{\mu})\rangle, \tag{37}$$

where by differentiating (37) with respect to $\mathbf{x}$ and setting to zero then the solution is

$$\mathbf{x} = \exp\left\{\boldsymbol{\mu} + \frac{diag(\boldsymbol{\Sigma})\mathbf{1}_{N_s}}{2}\right\}. \tag{38}$$

However, as presented in Fletcher and Zupanski (2006a), the statistic is an unbounded statistic with respect to the variance

and hence it is not a desirable statistic upon which to base a data assimilation system if the random variable is lognormally distributed.

4. Bayesian networks approach to 4D VAR

In this section, two different probability frameworks for 4D VAR, which are based upon a technique in probability theory called Bayesian networks, Pearl (2007), are derived.

Bayesian networks are graphical methods which, as according to Pearl (2007), have three roles:

(i) To provide convenient means of expressing substantive assumptions.

(ii) To facilitate economical representation of joint probability functions.

(iii) To facilitate efficient inference from observations.

In data assimilation, all three of the above properties are desirable in describing how the errors are behaving with respect to the information that is at hand.

The main feature of Bayesian networks is that they enable the removal of non-dependent probability events from the multiple probabilistic event version of Bayes theorem. The process of the removal of the non-dependent events has many names but the two most common names are conditional independence and Markov processes.

In the next subsection some of the theory of Bayesian networks is presented using the perfect model assumption as an example. Given the expressions that are derived, it is shown that the current Gaussian error formulation from weighted least squares is equivalent to the maximum likelihood estimator from this approach.

In Section 4.2, the network for a model error formulation, or weak constraint, version of 4D VAR is presented. Upon substituting the definition of errors into a multivariate Gaussian distribution, the expression that arises is the same as that in Van Leeuwen and Evensen (1996).

4.1. Bayesian networks for perfect model assumption

Bayesian networks are a type of graphical model that consist of nodes, arrows between nodes and probability assignments. These are also referred to as directed graphs. If there is an arrow pointing from node X to node Y then X is said to be the parent of Y. A node without a parent is referred to as a root node.

The advantage of Bayesian networks is that they allow the simplification of large joint distributions, $P(x_1, x_2, \ldots, x_N)$, where there are 2^N outcomes that have to be considered. Bayesian networks allow for the simplification of the joint probability through highlighting which events are connected to each other. This is usually accomplished through graphical representations, referred to as directed acyclic graphs, (DAG), Pearl (2007). The DAG for a small version of strong constraint 4D VAR is presented in Fig. 4.

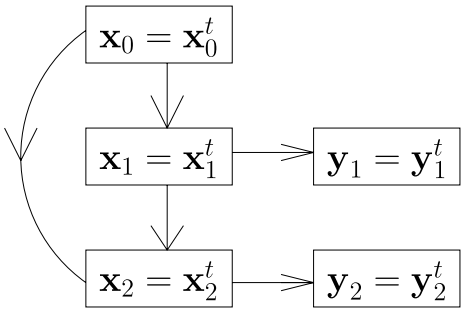

Fig. 4. DAG for strong constraint 4D VAR.

The joint probability density function, which is required to conditioned the probability of all of the observations, is represented by the multiple probabilistic event version of Bayes theorem. The well known two event version of Bayes theorem, which is the basis for 3D VAR, is

$$P(X, Y) = P(Y|X)P(X).$$

The multiple probabilistic event version of Bayes theorem allows the multivariate joint probability to be written in terms of the product of conditional distributions and a marginal distributions, given by

$$P(X_N, X_{N-1}, \ldots, X_2, X_1) = \left(\prod_{i=2}^{N} P(X_i|\hat{X}_{i-1})\right) P(X_1), \tag{39}$$

where $\hat{X}_{i-1} \equiv X_{i-1}, X_{i-2}, \ldots, X_1$.

It is clear from (39) that there are many expressions to consider in the joint probability function and this is where Bayesian networks become useful.

As a way to demonstrate this consider Fig. 4, which is the DAG for strong constraint 4D VAR for a small sample. The joint probability for Fig. 4 is

$$\begin{aligned} P(X_0, X_1, Y_1, X_2, Y_2) = {} & P(X_0)P(X_1|X_0)P(Y_1|X_1, X_0) \\ & \times P(X_2|Y_1, X_1, X_0) \\ & \times P(Y_2|X_2, Y_1, X_1, X_0), \end{aligned} \tag{40}$$

where X_i, $i = 0, 1, 2$ is the random event that $\mathbf{x}_i = \mathbf{x}_i^t$ and Y_i, $i = 1, 2$ is the random event that $\mathbf{y}_i = \mathbf{y}_i^t$ where $= \mathbf{x}_i$ is the model state and $\mathbf{y}_i$ is a set of observations at $t = t_i$ as in the previous section. This is where Bayesian networks become useful as they allow the removal of non-dependencies in (40). This is accomplished by the definition of Markov Parents/conditional independence.

Definition 1. (Markovian Parents)

For every variable X in the DAG, and every set $\boldsymbol{Y}$ of variables such that it does not include the set $\boldsymbol{DE}(X)$ of descendent's of X then X is conditionally independent from $\boldsymbol{Y}$ given the set $\boldsymbol{PA}(X)$

of its parents and therefore

$$P(X|\boldsymbol{PA}(X), \boldsymbol{Y}) = P(X|\boldsymbol{PA}(X)), \tag{41}$$

(Pearl, 2007).

Mathematically, Definition 1 can be written as

$$P(X_1, \ldots, X_n) = \left[\prod_{i=2}^{n} P(X_i|\boldsymbol{PA}(X_i))\right] P(X_1), \tag{42}$$

where (42) can be interpreted as the conditional independent version of Bayes theorem.

Therefore, considering Fig. 4, the Markov parents need to be identified. It is clear from Fig. 4 that X_0 is the root node. For the second event, X_1, then the parent node is X_0. For the first set of observations the parent node is X_1, not the set $\{X_1, X_0\}$. Therefore the joint probability is $P(Y_1|X_1)$. However, when we consider X_2 it is clear that this is dependent on the events X_1 and X_0. By assuming a hidden Markov model the conditional probability of event X_2 becomes $P(X_2|X_1)$.

Combining all the arguments above results in the final joint probability as

$$\begin{aligned} P(X_0, X_1, Y_1, X_2, Y_2) = {} & P(X_0)P(X_1|X_0)P(Y_1|X_1) \\ & \times P(X_2|X_1)P(Y_2|X_2,). \end{aligned} \tag{43}$$

There is one more simplification that can be made to (40) which comes from the perfect model assumption. The perfect model assumption implies that if the initial conditions are the true initial conditions then all of proceeding model states are true, therefore the conditional probability of the future model states is 1. Using this information results in the joint probability as

$$P(X_0, X_1, Y_1, X_2, Y_2) = P(X_0)P(Y_1|X_1)P(Y_2|X_2). \tag{44}$$

The arguments that have been used to derive (44) can be extended to t_a sets of observations which results in

$$P\left(X_0, X_1, Y_1, \ldots, X_{N_o}, Y_{N_o}\right) = P(X_0)\prod_{i=1}^{t_a} P(Y_i|X_i). \tag{45}$$

However, (45) is an expression for the joint pdf but it is the mode of this pdf that is required. To obtain the mode of the pdf defined in (45), then (45) has to be maximized. By using the log-likelihood approach then the mode of (45) is the minimum of the cost function

$$J(\mathbf{x}) = -\ln P(X_0) - \sum_{i=1}^{N_o} \ln P(Y_i|X_i). \tag{46}$$

4.1.1. Equivalence of the weighted least squares and probability models for multivariate Gaussian errors. To show the equivalence between the weighted least-squares approach and the probability model for the multivariate Gaussian case, the definition for multivariate normal background and observational errors are required. These are defined as

$$\begin{aligned} \boldsymbol{\varepsilon}_{b,0} &= \mathbf{x}_0 - \mathbf{x}_{b,0}, \qquad \boldsymbol{\varepsilon}_i^o = \mathbf{y}_i - \mathbf{h}_i(\boldsymbol{\mathcal{M}}_{0,i}(\mathbf{x}_0)) \\ \boldsymbol{\varepsilon}_{b,0} &\sim G(\mathbf{0}, \mathbf{B}) \qquad \boldsymbol{\varepsilon}_i^o \sim G(\mathbf{0}, \mathbf{R}_i), \end{aligned} \tag{47}$$

where G stands for multivariate Gaussian, which is

$$G(\boldsymbol{\mu}, \boldsymbol{\Sigma}) = (2\pi)^{-\frac{N}{2}} |\boldsymbol{\Sigma}|^{-\frac{1}{2}} \exp\left\{-\frac{1}{2}(\mathbf{x} - \boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1}(\mathbf{x} - \boldsymbol{\mu})\right\}, \tag{48}$$

where $\boldsymbol{\Sigma}$ is a covariance matrix and $\boldsymbol{\mu}$ is the vector of the expectations of the components of the random vector, $\mathbf{x}$.

Combining (47) and (48) and substituting them into (46) results in

$$\begin{aligned} J(\mathbf{x}) = {} & \frac{1}{2}(\mathbf{x}_0 - \mathbf{x}_{b,0})^T \mathbf{B}^{-1}(\mathbf{x}_0 - \mathbf{x}_{b,0}) \\ & + \frac{1}{2}\sum_{i=1}^{t_a}(\mathbf{y}_i - \mathbf{h}_i(\boldsymbol{\mathcal{M}}_{0,i}(\mathbf{x}_0)))^T \\ & \times \mathbf{R}_i^{-1}(\mathbf{y}_i - \mathbf{h}_i(\boldsymbol{\mathcal{M}}_{0,i}(\mathbf{x}_0))), \end{aligned} \tag{49}$$

which is same as (13).

4.2. Bayesian network for weak constraint 4D VAR

The weak constraint approach, as outlined in Sasaki (1970), is often considered the more realistic approach to data assimilation, as it is well known that the versions of the governing equations used in the numerical weather prediction systems are not the exact equations. Many simplifications are made by scale analysis to the continuous equations, Pedlosky (1987), which then enables the discrete approximations to be simpler. However, more errors are introduced when the numerical approximations are applied.

Model errors are a major problem for data assimilation methods as the strong constraint approach ignores them. However, in Daley (1992) the model error is shown to be a significant term. In this section a first order Markov variable (chain) is considered as a model error term, this is equivalent to assuming that the current model state is only conditionally dependent on the model state at the previous model time. The DAG for model error approach is presented in Fig. 5. Therefore, the conditional independent version of Bayes theorem for Fig. 5 is

$$\begin{aligned} & P(X_0, X_1, X_2, Y_1, X_3, X_4, Y_2) = P(X_0)P(X_1|X_0) \\ & P(X_2|X_1)P(Y_1|X_2)P(X_3|X_2)P(X_4|X_3)P(Y_2|X_4) \end{aligned} \tag{50}$$

which generalizes to

$$\begin{aligned} & P(X_0, X_1, X_2, Y_1, \ldots, X_{N-1}, Y_{N_o}, X_N) \\ & \quad = P(X_0)\prod_{i=1}^{N} P(X_i|X_{i-1})\prod_{j=1}^{t_a} P(Y_j|X_i), \end{aligned} \tag{51}$$

where N is the total number of model evaluations in the assimilation window.

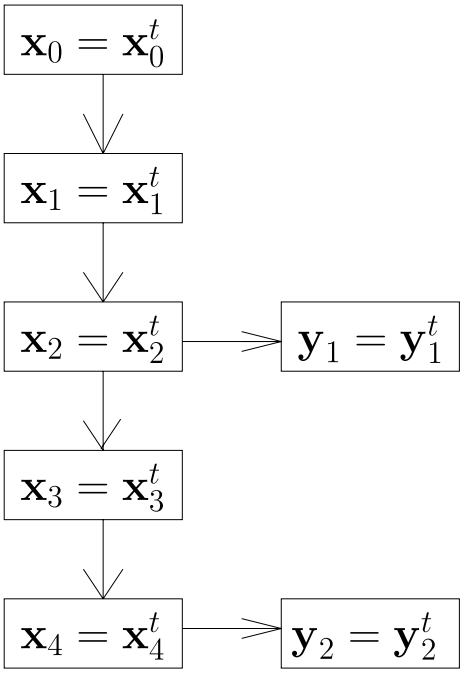

Fig. 5. DAG for weak constraint/imperfect model 4D VAR.

As with the strong constraint formulations, it is the maximum likelihood state that is required, therefore, by using the log-likelihood method, the cost function is

$$J(\mathbf{x}) = -\ln P(X_0) - \sum_{i=1}^{N} \ln P(Y_i|X_i) - \sum_{i=1}^{N_o} \ln P(X_i|X_{i-1}). \tag{52}$$

4.2.1. Gaussian model errors. If the model errors are Gaussian distributed then they are defined as by

$$\boldsymbol{\varepsilon}_m = \mathbf{x}_i - \boldsymbol{\mathcal{M}}_{i-1,i}(\mathbf{x}_{b,i-1}). \tag{53}$$

Therefore, using the error definitions given in (47) and (53) and then substituting these into (48) results in

$$\begin{aligned} J(\mathbf{x}_0) = {} & \frac{1}{2}(\mathbf{x}_0 - \mathbf{x}_{b,0})^T \mathbf{B}^{-1}(\mathbf{x}_0 - \mathbf{x}_{b,0}) \\ & + \frac{1}{2}\sum_{j=1}^{N_o}(\mathbf{y}_j - \mathbf{h}_j(\boldsymbol{\mathcal{M}}_{0,j}(\mathbf{x}_0)))^T \\ & \times \mathbf{R}_j^{-1}(\mathbf{y}_j - \mathbf{h}_j(\boldsymbol{\mathcal{M}}_{0,j}(\mathbf{x}_0))) \\ & + \sum_{i=1}^{N}(\mathbf{x}_i - \boldsymbol{\mathcal{M}}_{i-1,i}(\mathbf{x}_{b,i-1}))^T \\ & \times \mathbf{Q}_i^{-1}(\mathbf{x}_i - \boldsymbol{\mathcal{M}}_{i-1,i}(\mathbf{x}_{b,i-1})), \end{aligned} \tag{54}$$

which is the same as the expression in Van Leeuwen and Evensen (1996), where $\mathbf{Q}$ is the model error covariance matrix.

4.3. Probability model approach for lognormal observational, background and model errors

In this section, (52) and (46) are evaluated with lognormally distributed background, observational and model errors. The first step is to define the error structure for lognormally distributed errors, which due to the geometric nature of the lognormal distributions, these definitions are in the form of ratios, Cohn (1997) as follows

$$\boldsymbol{\varepsilon}_0(i) = \frac{\mathbf{x}_0(i)}{\mathbf{x}_{b,0}(i)}, \qquad i = 1, 2, \ldots, N \tag{55}$$

$$\boldsymbol{\varepsilon}^o(j,k) = \frac{\mathbf{y}(j,k)}{[\mathbf{h}(\boldsymbol{\mathcal{M}}_{0,j}(\mathbf{x}_0))]_{j,k}}, \quad j = 1, 2, \ldots, J$$

$$k = 1, 2, \ldots, K_j \tag{56}$$

$$\boldsymbol{\varepsilon}_m(i,l) = \frac{\mathbf{x}(i,l)}{\mathbf{x}_b(i,l)}, \qquad l = 1, 2, \ldots, t_a, \tag{57}$$

where $\mathbf{x}_b(:,l) = \boldsymbol{\mathcal{M}}_{l-1,l}(\mathbf{x}_{b,l-1})$, J is the total number of observations, K_j is the total number of observations in set j and the superscript o represents observation for the remainder of the paper.

Substituting (55), (56) and (57) into (19) and then into (52) results in

$$\begin{aligned} J(\mathbf{x}) = {} & \frac{1}{2}(\ln \mathbf{x}_0 - \ln \mathbf{x}_{b,0})^T \mathbf{B}^{-1}(\ln \mathbf{x}_0 - \ln \mathbf{x}_{b,0}) \\ & + \frac{1}{2}\sum_{k=1}^{K}(\ln \mathbf{y}_i - \ln \mathbf{h}_i(\boldsymbol{\mathcal{M}}_{0,i}(\mathbf{x}_0)))^T \\ & \times \mathbf{R}_i^{-1}(\ln \mathbf{y}_i - \ln \mathbf{h}_i(\boldsymbol{\mathcal{M}}_{0,i}(\mathbf{x}_0))) \\ & + \frac{1}{2}\sum_{l=1}^{t_a}(\ln \mathbf{x}_l - \ln \boldsymbol{\mathcal{M}}_{l-1,l}(\mathbf{x}_{b,l-1}))^T \\ & \times \mathbf{Q}_l^{-1}(\ln \mathbf{x}_l - \ln \boldsymbol{\mathcal{M}}_{l-1,l}(\mathbf{x}_{b,l-1})) \\ & + (\ln \mathbf{x}_0 - \ln \mathbf{x}_{b,0})^T \mathbf{1}_{N_s} \\ & + \sum_{i=1}^{N}(\ln \mathbf{y}_i - \ln \mathbf{h}_i(\boldsymbol{\mathcal{M}}_{0,i}(\mathbf{x}_0)))^T \mathbf{1}_{N_i} \\ & + \sum_{l=1}^{t_a}(\ln \mathbf{x}_l - \ln \boldsymbol{\mathcal{M}}_{l-1,l}(\mathbf{x}_{b,l-1}))^T \mathbf{1}_{N_s}, \end{aligned} \tag{58}$$

whose minimum is the mode of the analysis distribution.

It can easily be verified that applying (55) with (19) into (46) gives the same expression as (31) combined with lognormal background error component of (28).

5. Numerical experiments

In this section, the probability approach is compared with the transform method using the Lorenz'63 model, Lorenz (1963) for both the strong and weak constraint versions of the mixed distribution version of 4D VAR. Therefore in these experiments we are comparing the median with the mode as an analysis state. In the probability approach the mixed distribution version of the

4D VAR is used where two components of the model are Gaussian distributed and the third is lognormal. The cost function for the strong constraint mixed distribution 4D VAR is presented in Section 5.2. The weak constraint version is presented in Section 5.6.

In FZ07 the 3D VAR version of the mixed distribution system is demonstrated to outperform the transform approach more consistently for different observational errors sizes, as well as for different length of times between the observations. It is also shown in FZ07 that the z component of the Lorenz'63 model is not Gaussian distributed. The z component is also shown not to be lognormally distributed, however, the global mode is captured by a lognormal approximation, whereas the Gaussian approximation missed both modes due to the symmetry property. Therefore, to demonstrate the 4D version of the mixed distribution data assimilation technique the same model with the same assumptions as in FZ07 is used.

5.1. Numerical model

The Lorenz'63 model is a 3-D non-linear model consisting of ordinary differential equations, used to represent finite amplitude convection. These ordinary differential equations are

$$\dot{x} = -\sigma(x + y), \tag{59}$$

$$\dot{y} = -xz + \rho x - y, \tag{60}$$

$$\dot{z} = xy - \beta z, \tag{61}$$

where the superscript dot represents the time derivative, whilst σ, ρ and β are constants, usually taken to be $10, 28$ and $\frac{8}{3}$, respectively. These choices for the three constant generates a chaotic system, where chaotic means that the solution does not stay on the left lobe of the attractor but transitions between the two lobes of the attractors with no definite condition for the solution to change between the two lobes of the attractor.

This system of equations are often used to test new techniques in data assimilation methods. The chaotic nature of the system makes it possible to obtain quite different answers from two different initial conditions that differ at a few decimal places.

The numerical approximation that is used to approximate (59)–(61) is the modified Euler scheme with a time step of $h = 0.01$. To complete the numerical setup for the experiment the tangent linear model of the modified Euler approximation to (59)–(61) and its adjoint are required.

5.2. Strong constraint cost functions and their gradients

The two cost functions used in the experiments are

$$J_{mx}(\mathbf{x}_0) = \frac{1}{2}\boldsymbol{\varepsilon}_{b,0}^T\mathbf{B}^{-1}\boldsymbol{\varepsilon}_{b,0} + \frac{1}{2}\sum_{i=1}^{N_o}\boldsymbol{\varepsilon}_i^{oT}\mathbf{R}_i^{-1}\boldsymbol{\varepsilon}_i^o + (\ln z_0 - \ln z_{b,0}) + \sum_{i=1}^{N_o}\left(\ln z_i^o - \ln \boldsymbol{\mathcal{M}}_{0,i}(z_0)\right), \tag{62}$$

$$J_{tr}(\mathbf{x}_0) = \frac{1}{2}\hat{\boldsymbol{\varepsilon}}_{b,0}^T\mathbf{B}^{-1}\hat{\boldsymbol{\varepsilon}}_{b,0} + \frac{1}{2}\sum_{i=1}^{N_o}\hat{\boldsymbol{\varepsilon}}_i^{oT}\mathbf{R}_i^{-1}\hat{\boldsymbol{\varepsilon}}_i^o, \tag{63}$$

where

$$\boldsymbol{\varepsilon}_{b,0} = \begin{pmatrix} x_0 - x_{b,0} \\ y_0 - y_{b,0} \\ \ln z_0 - \ln z_{b,0} \end{pmatrix} \quad \boldsymbol{\varepsilon}_i^o = \begin{pmatrix} x_i^o - \boldsymbol{\mathcal{M}}_{0,i}(x_0) \\ y_i^o - \boldsymbol{\mathcal{M}}_{0,i}(y_0) \\ \ln z_i^0 - \ln \boldsymbol{\mathcal{M}}_{0,i}(z_0) \end{pmatrix},$$

$$\hat{\boldsymbol{\varepsilon}}_{b,0} = \begin{pmatrix} x_0 - x_{b,0} \\ y_0 - y_{b,0} \\ Z_0 - Z_{b,0} \end{pmatrix} \quad \hat{\boldsymbol{\varepsilon}}_i^o = \begin{pmatrix} x_i^o - \boldsymbol{\mathcal{M}}_{0,i}(x_0) \\ y_i^o - \boldsymbol{\mathcal{M}}_{0,i}(y_0) \\ Z_i^0 - Z_i \end{pmatrix},$$

$$\mathbf{x}^T = (x \quad y \quad z), \qquad \mathbf{X}^T = (x \quad y \quad \ln z), \tag{64}$$

where $Z = \ln z$.

An important feature to note here is the control vector that the two cost functions are to be minimized with respect to; for (62) the minimization is with respect to $\mathbf{x}$, which consists of the three variables in the Lorenz model. However, for (63) the control vector is $\mathbf{X}$, which contains the transformed state $\ln z$.

The gradients of (62) and (63) are

$$\nabla_{\mathbf{x}_0} J_{mx} = \mathbf{W}_b^T\left(\mathbf{B}_0^{-1}\boldsymbol{\varepsilon}_{b,0} - \begin{pmatrix} 0 \\ 0 \\ 1 \end{pmatrix}\right) + \sum_{i=1}^{N_o}\mathbf{M}_{0,i}^T\left(\mathbf{W}_{o,i}^T\left(\mathbf{R}_i^{-1}\boldsymbol{\varepsilon}_i^o + \begin{pmatrix} 0 \\ 0 \\ 1 \end{pmatrix}\right)\right), \tag{65}$$

$$\nabla_{\mathbf{x}_0} J_{tr} = \mathbf{B}_0^{-1}\hat{\boldsymbol{\varepsilon}}_{b,0} + \mathbf{W}_b^{-T}\sum_{i=1}^{N_o}\mathbf{M}_i^T\mathbf{R}_i^{-1}\hat{\boldsymbol{\varepsilon}}_i^o, \tag{66}$$

where

$$\mathbf{W}_b = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & z_0^{-1} \end{pmatrix}, \qquad \mathbf{W}_{o,i} = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & (\boldsymbol{\mathcal{M}}_{0,i}(z_0))^{-1} \end{pmatrix}.$$

The important feature to note about (65) and (66) is the increment that the tangent linear model is applied to. For (65) $\delta\mathbf{x}$ is

$$\delta\mathbf{x}_i = \begin{pmatrix} \sigma_{o,x}^{-2}\left(\boldsymbol{\mathcal{M}}_{0,i}(x_0) - x_i^o\right) \\ \sigma_{o,y}^{-2}\left(\boldsymbol{\mathcal{M}}_{0,i}\,(y_0) - y_i^o\right) \\ \left(\sigma_{o,z}^{-2}\left(\ln \boldsymbol{\mathcal{M}}_{0,i}(z_0) - \ln z_i^o\right) + 1\right)(\boldsymbol{\mathcal{M}}_{0,i}(z_0))^{-1} \end{pmatrix}, \tag{67}$$

whilst for (63) $\delta\mathbf{x}$ is

$$\delta\mathbf{x}_i = \begin{pmatrix} \sigma_{o,x}^{-2}\left(\boldsymbol{\mathcal{M}}_{0,i}(x_0) - x_i^o\right) \\ \sigma_{o,y}^{-2}\left(\boldsymbol{\mathcal{M}}_{0,i}(y_0) - y_i^o\right) \\ \sigma_{o,z}^{-2}\left(\ln \boldsymbol{\mathcal{M}}_{0,i}(z_0) - \ln z_i^o\right)(\boldsymbol{\mathcal{M}}_{0,i}(z_0))^{-1} \end{pmatrix}. \tag{68}$$

To minimize the cost functions, the MATLAB routine FMINSEARCH, which uses a Nelder–Mead simplex direct search

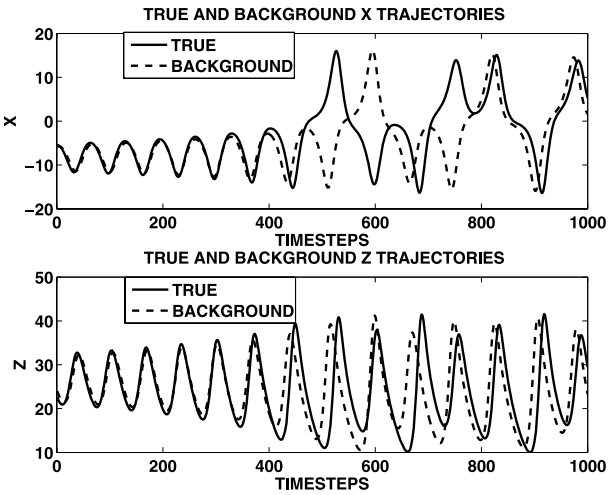

Fig. 6. Plot of the true and background x and z trajectories for the 1000 time steps.

method to find the minimum, is used. The routine is given the cost function and the gradient to evaluate.

The background error covariance matrices are allowed to evolve through the assimilation experiment. The generation of the matrices are as follows: The initial covariance matrix is calculated as the average of the differences between the true solution of a background solution (whose initial conditions are $x_{b,0} = -5.9$, $y_{b,0} = -5.0$ and $z_{b,0} = 24.0$, whereas the true solutions starts from $x_{t,0} = -5.4458$, $y_{t,0} = -5.4841$ and $z_{t,0} = 22.5606$) over the length of the first assimilation window. As the matrix is only a 3×3 matrix the exact inverse matrix is used. The true and the background trajectories can be seen in Fig. 6.

The **B** matrices for the remaining assimilation windows are calculated as follows:

$$\mathbf{B}_0^{(1)} = \frac{1}{S}\sum_{i=1}^{S}\left\langle \mathbf{x}_b - \mathbf{x}_{mx}, (\mathbf{x}_b - \mathbf{x}_{mx})^T \right\rangle_i ,$$

$$\mathbf{B}_0^{(2)} = \frac{1}{S}\sum_{i=1}^{S}\left\langle \mathbf{x}_{mx}^{(1)} - \mathbf{x}_{mx}^{(2)}, \left(\mathbf{x}_{mx}^{(1)} - \mathbf{x}_{mx}^{(2)}\right)^T \right\rangle_i ,$$

$$\vdots$$

$$\mathbf{B}_0^{(k)} = \frac{1}{S}\sum_{i=1}^{S}\left\langle \mathbf{x}_{mx}^{(k-1)} - \mathbf{x}_{mx}^{(k)}, \left(\mathbf{x}_{mx}^{(k-1)} - \mathbf{x}_{mx}^{(k)}\right)^T \right\rangle_i , \quad (69)$$

where S is the total number of time steps in the assimilation window, and $k = 2, 3, \ldots, K$ where K is the total number of assimilation windows. Therefore, the **B** matrices are calculated as the differences between the model run starting with the solution at the end of the last assimilation window and the solution coming from the minimization for the current assimilation window. The same method is used to generate the **B** matrix for the transform approach. It should be noted that for the z component of **x** it is $\ln z$ that is evaluated in (69), not z.

The observations used in the experiments are generated from adding perturbations that are generated from a Gaussian random generator in MATLAB given a value for σ^2. For the lognormal observations the property of the Gaussian distribution that if $x\ G(\mu, \sigma^2)$ then $\ln x\ LN(\mu, \sigma^2)$ and hence a Gaussian perturbation is generated given σ^2, then the exponential of that perturbation multiplies the z component of the true model run, the same way as in FZ07. Results are presented from experiments where different assimilation window lengths and different observational errors sizes are used.

5.3. Experiment 1

In all of the results that are shown, the total number of time steps that the assimilation is performed on is 1000. Each of the experiments have either different assimilation window lengths, number of observations or different observational error sizes.

The first set of results, shown in Figs 7 and 8, have assimilation window lengths of 100 time steps, but have different observational errors of $\sigma^2 = 0.25$ and 1.5. There are 10 observations in each window.

The results presented in all of the figures in this section are of the departures from the true solution for each analysis solution. For the first experiment in Fig. 7, it can be seen that both methods perform quite well, this is not unexpected considering how close the lognormal distribution is to the Gaussian distribution for small variances, but also given the volume of observations and the shortness of the assimilation windows then the two scheme should be quite similar.

When the observational error is increased to 1.5 (Fig. 8), but the number of observations is kept at 10 per window, then there again appears not to be too much difference between the two approaches. The purpose of this experiment is to see the impact of more lognormal structure to the observational error but given a lot of data and small windows is there much difference between the two approaches, and it appears that there is not.

5.4. Experiment 2

For the second experiment, the number of observations in each assimilation window is still 10, but the window lengths are extended to 500 time steps. The observational errors are still 0.25 and 1.5 for the two sets of plots presented for this experiment.

It can be seen in Fig. 9, for both the x and z components, that in the first window, the two solutions are almost identical. However, the important feature of these results is in the second assimilation window. Here it can be seen that the transform approach does not capture the correct initial conditions for the next window while it appears that the model approaches

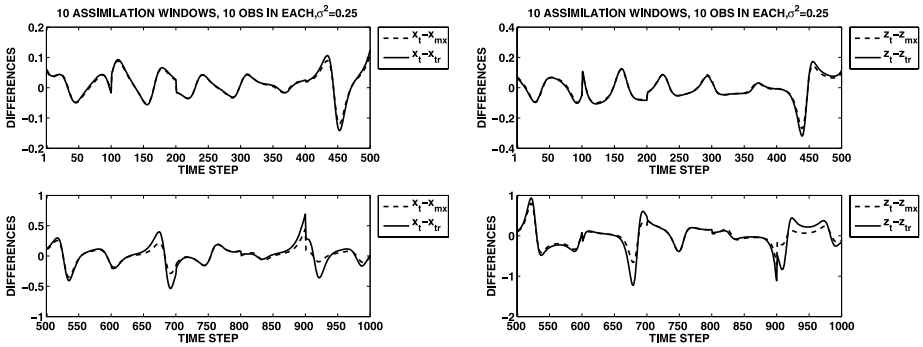

Fig. 7. Plots of the departures from the true solution by both the mixed approach (dashed line) and the transform approach (solid line) for x and z component of the model for $N_o = 10$, $\sigma^2 = 0.25$, $S = 10$.

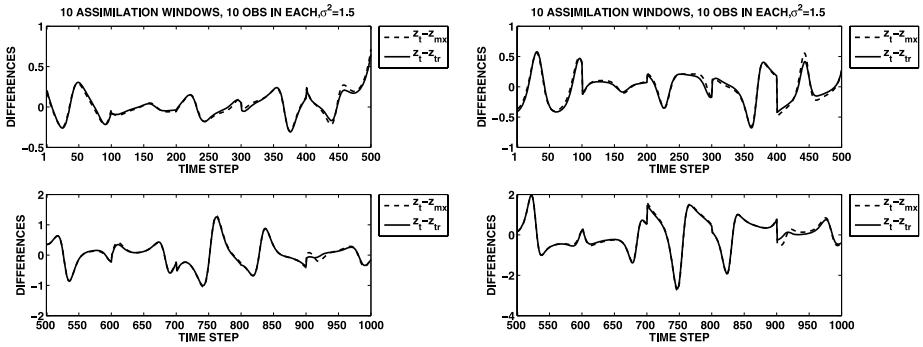

Fig. 8. Plots of the departures from the true solution by both the mixed approach (dashed line) and the transform approach (solid line) for x and z component of the model for $N_o = 10$, $\sigma^2 = 1.5$, $S = 10$.

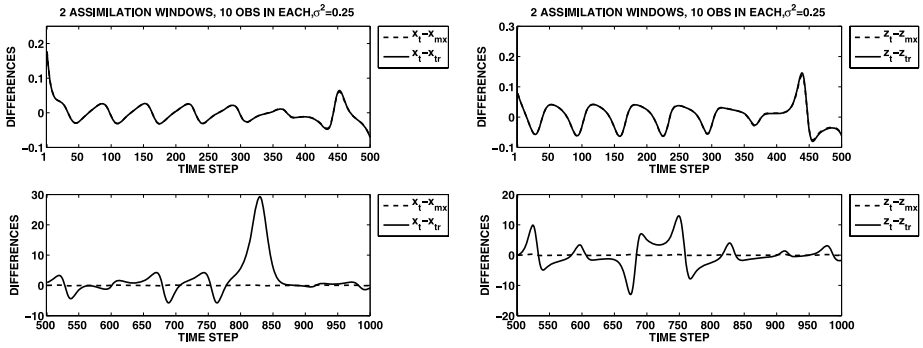

Fig. 9. Plots of the departures from the true solution by both the mixed approach (dashed line) and the transform approach (solid line) for x and z component of the model for $N_o = 10$, $\sigma^2 = 0.25$, $S = 2$.

does capture nearly the true initial conditions for the second window.

In the second part of this experiment the observational error is increased to 1.5, Fig. 10. Again in the first window the two schemes' results are quite similar. However, in the second window the transform approach again fails to find the correct initial conditions. Therefore, these results are suggesting that the median may not be a good consistent statistic with which to perform the analysis with. It can be seen in Fig. 10 that the mixed distribution approach has slightly more noticeable departures from the true solution, indicating that the less accurate observations are having an impact on this approach, but not to the extent as on the transform approach.

5.5. Experiment 3

In the third experiment the two approaches are compared under the situation where the window length is 1000 time steps, there are 20 observations but with observational errors variance of $\sigma^2 = 1.5$, Fig. 11. The important feature of this experiment is to demonstrate that even for long window lengths, the modal approach can obtain the true solution, whilst the transform

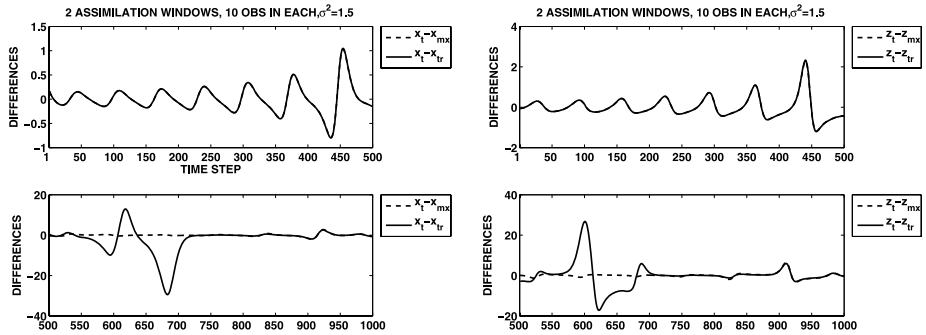

Fig. 10. Plots of the departures from the true solution by both the mixed approach (dashed line) and the transform approach (solid line) for x and z component of the model for $N_o = 10, \sigma^2 = 1.5, S = 2$.

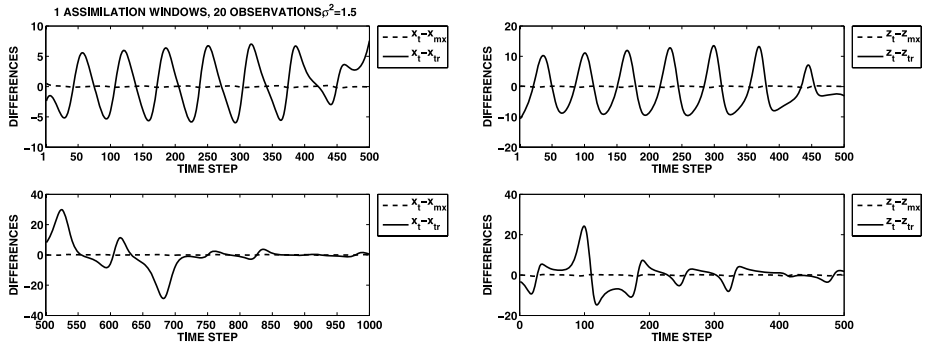

Fig. 11. Plots of the departures from the true solution by both the mixed approach (dashed line) and the transform approach (solid line) for x and z component of the model for $N_o = 20, \sigma^2 = 1.5, S = 1$.

approach does not. Another important fact about this experiment is that although the modal approach did need a few more function evaluations and iterations than the transform approach, it found a set of initial conditions very close to the true initial conditions.

In conclusion, the mixed distribution approach has been a more reliable method for finding the true initial conditions in the experiments given fairly accurate or very inaccurate observations, whilst the transform approach worked well for small windows with fairly accurate observations, it appears not to work as well when there are larger windows with fewer accurate observations.

We now consider the model error/weak constraint version of the mixed distribution system to see if these conclusions are valid in the presence of model error.

5.6. *Weak constraint cost functions and their gradients*

The version of weak constraint 4D VAR that is considered in these experiments is the state augmentation approach as outlined in Zupanski (1997). This approach is where the model error term is considered as a control variable, but where it is assumed that the errors associated with the initial conditions are independent of the model error terms. The assumption of the independence of the initial conditions errors and the model errors enables the **B** and **Q** matrices to be only the size of N_s rather than $2N_s$. In the experiments that are presented in Section 5.8 it is assumed that the model error term is assumed to be constant for each time step but is generated randomly from a lognormal or Gaussian distribution initially. The cost function for the modal approach is given by

$$
\begin{aligned}
J(\mathbf{x}_0, \boldsymbol{\Phi}) = {} & \frac{1}{2}\begin{pmatrix} \mathbf{x}_{0,p} - \mathbf{x}_{b,0,p} \\ \ln \mathbf{x}_{0,q} - \ln \mathbf{x}_{b,0,q} \end{pmatrix}^T \mathbf{B}_0^{-1} \begin{pmatrix} \mathbf{x}_{0,p} - \mathbf{x}_{b,0,p} \\ \ln \mathbf{x}_{0,q} - \ln \mathbf{x}_{b,0,q} \end{pmatrix} \\
& + \begin{pmatrix} \mathbf{x}_{0,p} - \mathbf{x}_{b,0,p} \\ \ln \mathbf{x}_{0,q} - \ln \mathbf{x}_{b,0,q} \end{pmatrix}^T \begin{pmatrix} \mathbf{0}_p \\ \mathbf{1}_q \end{pmatrix} \\
& + \frac{1}{2}\sum_{i=1}^{N_o} \begin{pmatrix} \mathbf{y}_{p_i,i} - \mathbf{y}_{p_i,i,k} \\ \ln \mathbf{y}_{q_i,i} - \ln \mathbf{y}_{q_i,i,k} \end{pmatrix}^T \mathbf{R}_i^{-1} \begin{pmatrix} \mathbf{y}_{p_i,i} - \mathbf{y}_{p_i,i,k} \\ \ln \mathbf{y}_{q_i,i} - \ln \mathbf{y}_{q_i,i,k} \end{pmatrix} \\
& + \begin{pmatrix} \mathbf{y}_{p_i,i} - \mathbf{y}_{p_i,i,k} \\ \ln \mathbf{y}_{q_i,i} - \ln \mathbf{y}_{q_i,i,k} \end{pmatrix}^T \begin{pmatrix} \mathbf{0}_{p_i,i} \\ \mathbf{1}_{q_i,i} \end{pmatrix} \\
& + \frac{1}{2}\begin{pmatrix} \boldsymbol{\Phi}_p \\ \ln \boldsymbol{\Phi}_q \end{pmatrix}^T \mathbf{Q}^{-1} \begin{pmatrix} \boldsymbol{\Phi}_p \\ \ln \boldsymbol{\Phi}_q \end{pmatrix} + \begin{pmatrix} \boldsymbol{\Phi}_p \\ \ln \boldsymbol{\Phi}_q \end{pmatrix}^T \begin{pmatrix} \mathbf{0}_p \\ \mathbf{1}_q \end{pmatrix}, \qquad (70)
\end{aligned}
$$

subject to

$$
\mathbf{x}_{k,p} = \boldsymbol{\mathcal{M}}_k(\mathbf{x}_{k-1}) + \boldsymbol{\Phi}_p, \qquad (71)
$$

$$[\mathbf{x}_{k,q}]_s = [\boldsymbol{\mathcal{M}}_k(\mathbf{x}_{k-1})]_s[\boldsymbol{\Phi}_q]_s, \quad s = p+1, p+2, \ldots, q, \tag{72}$$

$$\mathbf{y}_{i,k} = \mathbf{h}_i(\mathbf{x}_k). \tag{73}$$

For the transform approach the cost function is

$$\begin{aligned} J(\mathbf{x}_0, \boldsymbol{\Phi}) = {} & \frac{1}{2}\begin{pmatrix} \mathbf{x}_{0,p} - \mathbf{x}_{b,0,p} \\ \mathbf{X}_{0,q} - \mathbf{X}_{b,0,q} \end{pmatrix}^T \mathbf{B}_0^{-1} \begin{pmatrix} \mathbf{x}_{0,p} - \mathbf{x}_{b,0,p} \\ \mathbf{X}_{0,q} - \mathbf{X}_{b,0,q} \end{pmatrix} \\ & + \frac{1}{2}\sum_{i=1}^{N_o} \begin{pmatrix} \mathbf{y}_{p_i,i} - \mathbf{y}_{p_i,i,k} \\ \ln \mathbf{y}_{q_i,i} - \ln \mathbf{y}_{q_i,i,k} \end{pmatrix}^T \mathbf{R}_i^{-1} \begin{pmatrix} \mathbf{y}_{p_i,i} - \mathbf{y}_{p_i,i,k} \\ \ln \mathbf{y}_{q_i,i} - \ln \mathbf{y}_{q_i,i,k} \end{pmatrix} \\ & + \begin{pmatrix} \mathbf{y}_{p_i,i} - \mathbf{y}_{p_i,i,k} \\ \ln \mathbf{y}_{q_i,i} - \ln \mathbf{y}_{q_i,i,k} \end{pmatrix}^T \begin{pmatrix} \mathbf{0}_{p_i,i} \\ \mathbf{1}_{q_i,i} \end{pmatrix} \\ & + \frac{1}{2}\begin{pmatrix} \boldsymbol{\Phi}_p \\ \hat{\boldsymbol{\Phi}}_q \end{pmatrix}^T \mathbf{Q}^{-1} \begin{pmatrix} \boldsymbol{\Phi}_p \\ \hat{\boldsymbol{\Phi}}_q \end{pmatrix}, \end{aligned} \tag{74}$$

where $\hat{\boldsymbol{\Phi}} \equiv \ln \boldsymbol{\Phi}$, subject to

$$\begin{aligned} \mathbf{x}_{k,p} &= \boldsymbol{\mathcal{M}}_k(\mathbf{x}_{k-1}) + \boldsymbol{\Phi}_p, \\ [\mathbf{x}_{k,q}]_s &= [\boldsymbol{\mathcal{M}}_k(\mathbf{x}_{k-1})]_s[\boldsymbol{\Phi}_q]_s, \quad s = p+1, p+2, \ldots, q, \\ \mathbf{y}_{i,k} &= \mathbf{h}_i(\mathbf{x}_k). \end{aligned}$$

The gradients for (70) is derived in Appendix C, whilst in Appendices A and B the derivations for a constant Gaussian and lognormal bias are presented, which given these derivations results in

$$\nabla J = \begin{pmatrix} \nabla_{\mathbf{x}_0} J \\ \nabla_{\boldsymbol{\Phi}} J \end{pmatrix}, \tag{75}$$

$$\begin{aligned} \nabla_{\mathbf{x}_0} J = {} & \widehat{\mathbf{W}}_b^T \left(\mathbf{B}_0^{-1} \begin{pmatrix} \mathbf{x}_{0,p} - \mathbf{x}_{b,0,p} \\ \ln \mathbf{x}_{0,q} - \ln \mathbf{x}_{b,0,q} \end{pmatrix} + \begin{pmatrix} \mathbf{0}_p \\ \mathbf{1}_q \end{pmatrix} \right) \\ & - \sum_{i=1}^{N_o} \widetilde{\mathbf{M}}_{1,k}^T \mathbf{H}_i^T \widehat{W}_{o,i}^T \left(\mathbf{R}_i^{-1} \begin{pmatrix} \mathbf{y}_{p_i,i} - \mathbf{y}_{p_i,i,k} \\ \ln \mathbf{y}_{q_i,i} - \ln \mathbf{y}_{q_i,i,k} \end{pmatrix} \right. \\ & \left. + \begin{pmatrix} \mathbf{0}_{p_i,i} \\ \mathbf{1}_{q_i,i} \end{pmatrix} \right) \end{aligned} \tag{76}$$

$$\begin{aligned} \nabla_{\boldsymbol{\Phi}} J = {} & -\sum_{i=1}^{N_0} \left(\widetilde{\boldsymbol{\mathcal{M}}}_k^T + \sum_{n=0}^{k-2} \widetilde{\boldsymbol{\mathcal{M}}}_{k-n-1}^T \widetilde{\mathbf{M}}_{k-n,k}^T \right) \\ & \times \mathbf{H}_i^T \widehat{\mathbf{W}}_{o,i}^T \left(\mathbf{R}_i^{-1} \begin{pmatrix} \mathbf{y}_{p_i,i} - \mathbf{y}_{p_i,i,k} \\ \ln \mathbf{y}_{q_i,i} - \ln \mathbf{y}_{q_i,i,k} \end{pmatrix} + \begin{pmatrix} \mathbf{0}_{p_i,i} \\ \mathbf{1}_{q_i,i} \end{pmatrix} \right) \\ & + \widehat{\mathbf{W}}_m \left(\mathbf{Q}^{-1} \begin{pmatrix} \boldsymbol{\Phi}_p \\ \ln \boldsymbol{\Phi}_q \end{pmatrix} + \begin{pmatrix} \mathbf{0}_p \\ \mathbf{1}_q \end{pmatrix} \right), \end{aligned} \tag{77}$$

where

$$\widetilde{\mathbf{M}}_k = \begin{pmatrix} \mathbf{M}_{k,p,p} & \mathbf{M}_{k,p,q} \\ \widehat{\mathbf{M}}_{k,q,p} & \widehat{\mathbf{M}}_{k,q,q} \end{pmatrix} \quad \widetilde{\boldsymbol{\mathcal{M}}}_k = \begin{pmatrix} \mathbf{I}_{p,p} & \mathbf{0}_{p,q} \\ \mathbf{0}_{q,p} & \widehat{\boldsymbol{\mathcal{M}}}_{k,q,q} \end{pmatrix},$$

$$[\widehat{\mathbf{M}}_k]_{i,j} = \frac{\partial \boldsymbol{\mathcal{M}}_k(\mathbf{x}_{k-1})_i}{\partial x_j}\Phi_i \qquad i = 1, 2, \ldots, N_s, \quad j = 1, 2, \ldots, N_s$$

$$[\widehat{\boldsymbol{\mathcal{M}}}_k]_{i,j} = \boldsymbol{\mathcal{M}}_k(\mathbf{x}_{k-1})_i \frac{\partial \Phi_i}{\partial \Phi_j}, \qquad i = 1, 2, \ldots, N_s, \quad j = 1, 2 \ldots, N_s,$$

and

$$\hat{\mathbf{W}}_m = \begin{pmatrix} \mathbf{I}_p & \mathbf{0}_{p,q} \\ \mathbf{0}_{q,p} & diag\left(\boldsymbol{\Phi}_j^{-1}\right) \end{pmatrix}, \quad j = p+1, p+2, \ldots, p+q.$$

The gradients for the transform approach is

$$\begin{aligned} \nabla_{\mathbf{x}_0} J = {} & \mathbf{B}_0^{-1} \begin{pmatrix} \mathbf{x}_{0,p} - \mathbf{x}_{b,0,p} \\ \mathbf{X}_{0,q} - \mathbf{X}_{b,0,q} \end{pmatrix} - \begin{pmatrix} \mathbf{I}_{p,p} & \mathbf{0}_{p,q} \\ \mathbf{0}_{q,p} & \left(\frac{\partial \mathbf{x}_0}{\partial \mathbf{X}_0}\right)_{q,q} \end{pmatrix} \\ & \times \sum_{i=1}^{N_o} \widetilde{\mathbf{M}}_{1,k}^T \mathbf{H}_i^T \widehat{W}_{o,i}^T \mathbf{R}_i^{-1} \begin{pmatrix} \mathbf{y}_{p_i,i} - \mathbf{y}_{p_i,i,k} \\ \ln \mathbf{y}_{q_i,i} - \ln \mathbf{y}_{q_i,i,k} \end{pmatrix} \end{aligned} \tag{78}$$

$$\begin{aligned} \nabla_{\hat{\boldsymbol{\Phi}}} J = {} & -\sum_{i=1}^{N_0} \begin{pmatrix} \mathbf{I}_{p,p} & \mathbf{0}_{p,q} \\ \mathbf{0}_{q,p} & \left(\frac{\partial \boldsymbol{\Phi}}{\partial \hat{\boldsymbol{\Phi}}}\right)_{q,q} \end{pmatrix} \\ & \times \left(\widetilde{\boldsymbol{\mathcal{M}}}_k^T + \sum_{n=0}^{k-2} \widetilde{\boldsymbol{\mathcal{M}}}_{k-n-1}^T \widetilde{\mathbf{M}}_{k-n,k}^T \right) \\ & \times \mathbf{H}_i^T \widehat{\mathbf{W}}_{o,i}^T \mathbf{R}_i^{-1} \begin{pmatrix} \mathbf{y}_{p_i,i} - \mathbf{y}_{p_i,i,k} \\ \ln \mathbf{y}_{q_i,i} - \ln \mathbf{y}_{q_i,i,k} \end{pmatrix} \\ & + \mathbf{Q}^{-1} \begin{pmatrix} \boldsymbol{\Phi}_p \\ \hat{\boldsymbol{\Phi}}_q \end{pmatrix}. \end{aligned} \tag{79}$$

5.7. Weak constraint experimental design

The weak constraint experiments follow a similar set up as for the strong constraint experiments, in that the observations are calculated in the same way, the model error covariance matrix is calculated similar to the background error covariance. The main difference between the **Q** and **B** calculations is that the **Q** matrix is calculated from the averaged differences between the true model run with no model bias and the bias model run with the true initial conditions.

The background trajectory is the same as the one from the strong constraint experiment. The model errors are generated randomly from a Gaussian distribution with a specific defined variance, and mean zero. These are then added to the two Gaussian variable model equations at each time step and the exponential of the number generated for the lognormal component

multiplies the model equation for the z component. The model bias for the z component appears in the adjoint equations as shown in Appendix B and so are included in the adjoint program along with Φ_z which we are trying to find to counter the bias. The first guess for the model error of the control vector is always (0 0 1) for the modal approach and (0 0 0) for the transform approach.

5.8. Weak constraint experiments

The first problem that occurs when designing a model error experiment with the Lorenz 1963 system is sensitivity of the model. It does not take much pushing in any direction for the model to either become a stable decaying solution staying on one arm of the attractor, or completely unstable and blowing up to infinity. Therefore, the model biases introduce could result in either these solutions, which is different to the changing true solution as seen in Fig. 6. We present 4 small experiments to demonstrate first that it is possible to account for a lognormal bias in a system given standard and highly accurate observations. The first difference from the strong constraint experiments is that both DA systems are not stable for as large a window as shown for the strong constraint experiments. This could possible be due to the trajectory of the first guess being to far away from the true solution by the later times, making the tangent linear assumption invalid.

The four results that are shown are different combinations of model and observation error values as well as four different assimilation window lengths. The first experiment is where the model error variance is σ^2_{me}, is 0.25 units, $\sigma^2_o = 0.25$ units, with 10 observations in the window. The results for this experiment are shown in Fig. 12 where we have a similar result as for the strong constraint with the two analysis statistics able to approximate the true solution quite well for the x component whilst for the z component there is a slight difference but the two are quite similar.

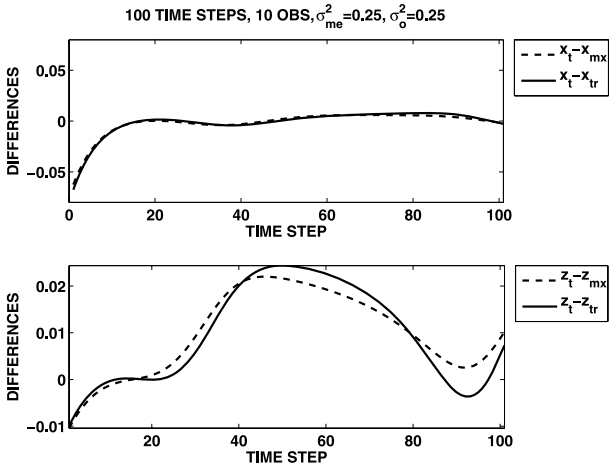

Fig. 12. Plots of the departures from the true solution by both the mixed approach (dashed line) and the transform approach (solid line) for x and z component of the model for $N_o = 10$, $\sigma^2_o = 0.25$, $\sigma^2_{me} = 0.25$ with a window of 100 time steps.

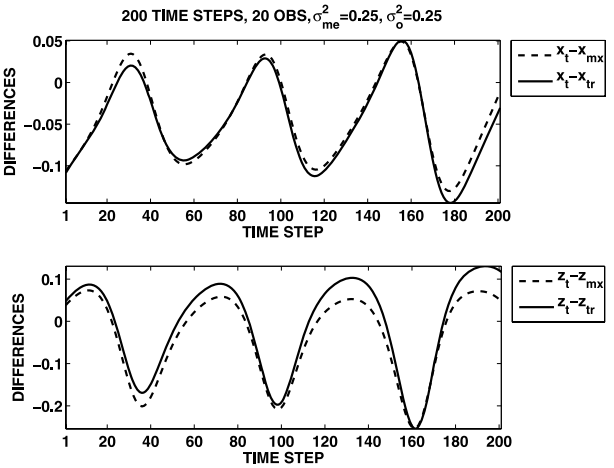

Fig. 13. Plots of the departures from the true solution by both the mixed approach (dashed line) and the transform approach (solid line) for x and z component of the model for $N_o = 20$, $\sigma^2_o = 0.25$, $\sigma^2_{me} = 0.25$ with a window of 200 time steps.

The second experiment is with a window of 200 time steps, $\sigma^2_{me} = 0.25$, $\sigma^2_o = 0.25$ and 20 observations which is the equivalent to the first experiment with observations every ten time steps. As shown in Fig. 13 the two solutions are starting to separate from each other with the modal approach slightly performing better than the transform approach for the x component but clearly performing slightly better than the transform approach for the z component. Therefore, the effects of the window length are starting to appear as in the strong constraint situation.

For the third experiment the assimilation window is extended by another 100 times steps and an extra 10 observations are included with the same observation error as the last two experiments, but now the model error is doubled to 0.5 units, Fig. 14. The clear result from this experiment is that the transform approach fails to capture the bias as well as the correct initial conditions, whilst the modal approach is able to capture most of the bias and is more accurate than the transform approach for both the x and z component. The significance of this experiment is that it is demonstrating the limits on the transform approach. Here for these combination, remembering that the model error biases are not 0.5 but are randomly selected from a Gaussian distribution with $\sigma^2 = 0.5$. Whilst it can be seen that the mixed modal approach does not capture the exact true solution it is nearer than the transform method.

The last experiment, Fig. 15, has an assimilation window of 400 time steps, this is the largest window that the modal

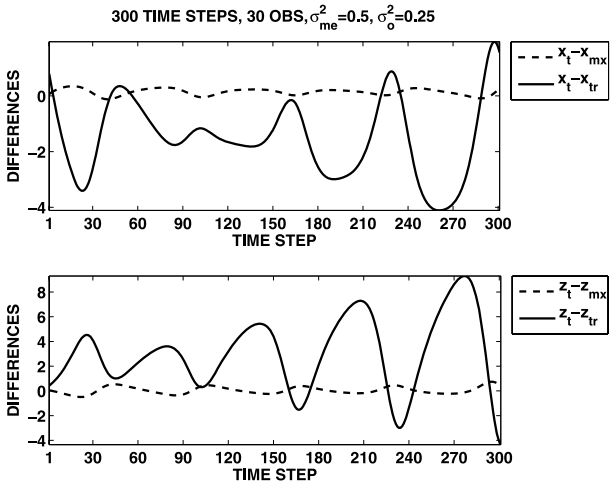

Fig. 14. Plots of the departures from the true solution by both the mixed approach (dashed line) and the transform approach (solid line) for x and z component of the model for $N_o = 30$, $\sigma_o^2 = 0.2$, $\sigma_{me}^2 = 0.5$ with a window of 300 time steps.

approach was stable to for the all the different set ups that were tried. In this experiment there are fewer observations, only 20 for the whole window, but these are quite accurate observations with $\sigma_o^2 = 0.01$ units and with $\sigma_{me}^2 = 0.2$ units. The significance of this experiment is to demonstrate that for larger assimilation windows, with fewer but highly accurate observations, recall the observations are generated from the true solution, and a relatively small model error variance the transform approach does not approximate the true solution. It is also clear that the modal approach does not capture the true solution but it is more accurate than the transform approach.

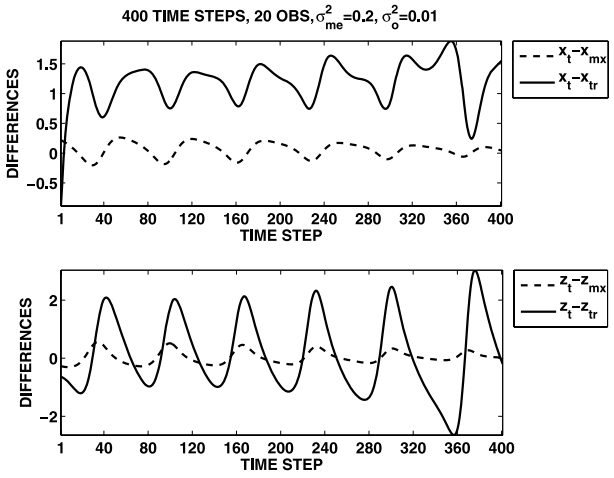

Fig. 15. Plots of the departures from the true solution by both the mixed approach (dashed line) and the transform approach (solid line) for x and z component of the model for $N_o = 20$, $\sigma_o^2 = 0.01$, $\sigma_{me}^2 = 0.2$ with a window of 400 time steps.

6. Conclusions and further work

In this paper, a brief overview of current techniques to derive 4D VAR have been presented. The two approaches were a weighted least-squares approach and a functional minimization set up. These two approaches have been extended to both lognormal errors and mixed lognormal-Gaussian errors. It has been shown that the weighted least-squares approach, by a transform of the lognormal random variable into a Gaussian random variable, obtains a median of the multivariate lognormal distribution and not the mode. It has also been shown that it is possible to define a functional so that the minimum of the functional is the mode of the multivariate lognormal distribution. One of the significance of these derivations is that transformed weighted least-squares approach is used in some operational centres in data assimilation systems.

Whilst it is possible to derive a functional approach for the mode when there are lognormal and mixed Gaussian-lognormal errors, a more general maximum likelihood approach was needed. In Section 4, two different probability models were derived through Bayesian network and conditional independence theory. The two approaches were for the strong and weak constraint versions of 4D VAR. During the derivation no correlation assumptions were made, making (46) and (51) applicable for any distributions. It was shown that the current Gaussian strong constraint 4D VAR is obtained from this maximum likelihood approach. The consequence of this is that although the weighted least-squares method finds the median, for the Gaussian situation this is know to be equivalent to the maximum likelihood approach, Clarke and Cooke (2004). It is also shown in Section 4 that the functional approach for the mode, presented in Section 4.3, for lognormal errors was equivalent to (46) evaluated with the lognormal distribution.

In Section 5 results from experiments comparing the transform approach with the probability approach are presented for the strong constraint formulation of 4D VAR. In these experiments the size of the assimilation windows are varied, as well as the variance of the observational error and the number of observations in each window. The results showed that the transform (or unbiased approach) performed as well as the probability approach for short windows with multiple observations, even when $\sigma^2 = 1, 5$ units. However, when the windows were lengthened the unbiased approach could not capture the true solution whilst the probability approach could. This was true for both small and large observational error variance values. In the final experiment there is only one window but with the same number of observations as the second experiment with two windows. The results showed that the unbiased approach did not capture the true solution and converged quicker than the maximum likelihood approach, which did capture the correct solution. While these results are in a simpler ordinary differential equations model they have demonstrated that there are consequences for

which approach to use when dealing with non-Gaussian random variables in a 4D VAR framework.

In Section 5.7, a new cost function and gradient are presented for the weak constraint version of 4D VAR for mixed Gaussian-lognormal model errors and the associated cost function and gradient for the transform approach, the proof of the derivations are shown in Appendices A, B and C. The results from a series of experiments for the model error formulation are presented in Section 5.8, where it was assumed that there was a constant bias that is randomly drawn from a Gaussian distribution for the x and y component of the Lorenz 1963 model, and from a lognormal distribution for the z component. For the x and y components the biases was added to the discrete equations while for the z component the discrete equation was multiplied by the random number. It was shown that although it was not possible to go out to as large a window as for the strong constraint where the bias was not present, the modal approach was more reliable than the transform approach when the window length was larger than 200 time steps, given a large variance for the model error distributions or very accurate observations, which were generated from the true solution.

In conclusion, the modal approach has been shown to be a possible viable alternative to transforming lognormal variables into their Gaussian counterpart for both the strong and weak constraint, albeit a constant random bias and in a simpler model. As part of this work a general probability model was derived from Bayesian network and conditional independence theory, which allows for a 4D VAR system for errors of any distribution based upon a maximum likelihood approach.

The plans for further work are to develop this new approach in both the 3-D and 4-D approaches for cloud prediction systems. These systems are highly reliant on the correct prediction of humidity and moisture variables in general. In Section 2, is was shown that a linear function of humidity was not a Gaussian random variable during certain times of the year, and that it is better represented by a lognormal distribution. Therefore, assimilation systems need to be developed that can handle these variables more consistently.

7. Acknowledgments

I am grateful to Drs. Andy Jones, Manajit Sengupta, Dusanka and Milija Zupanski for useful discussion for the motivation for this work. I am also grateful to Dr. Sengupta for providing the data from the ARM SGP site to demonstrate that humidity can be a non-Gaussian variable. ARM is a program in the United States Department of Energy. I am also extremely grateful to Dr. Dusanka Zupanski for helping us understand how to solve the model error problem presented here. I am very grateful for the constructive and useful reviews from Prof. Carlos Pires and the other anonymous reviewer. This work was funded by the Department of Defence (DoD) Center for Geosciences/Atmospheric Research at Colorado State University under Cooperative Agreement (W911NF-060200015) with the Army Research Laboratory.

8. Appendix A: Derivation of the model error experiment gradient

The model error, or weak constraint formulation, as set out in Zupanski (1997), for the Gaussian errors, constant bias case has the associated functional

$$J(\mathbf{x}_0, \mathbf{\Phi}) = \frac{1}{2}(\mathbf{x}_0 - \mathbf{x}_{b,0})^T \mathbf{B}^{-1}(\mathbf{x}_0 - \mathbf{x}_{b,0}) + \frac{1}{2}\sum_{i=1}^{N_o}(\mathbf{y}_i - \mathbf{y}_m)^T \mathbf{R}_i^{-1}(\mathbf{y}_i - \mathbf{y}_m) + \frac{1}{2}\mathbf{\Phi}^T \mathbf{Q}^{-1}\mathbf{\Phi}, \tag{A1}$$

where $\boldsymbol{\mathcal{M}}$ and $\mathbf{h}$ are postprocessing operators imposed as weak constrains defined by

$$\mathbf{x}_m = \boldsymbol{\mathcal{M}}_{m,m-1}(\mathbf{x}_{m-1}) + \mathbf{\Phi}, \tag{A2}$$

$$\mathbf{y}_i = \mathbf{h}_i(\mathbf{x}_m) + \boldsymbol{\varepsilon}_i^o, \tag{A3}$$

where the subscript m represents the number of model evaluations needed from the start of the window to the specific observation time.

An approximation to the conditions for a minimum to (A1) is to consider the case when the first variation of (A1) combined with (A2), δJ is zero, Zupanski (1997). The first variation can be written in terms of the gradient with respect to the two control variables $\mathbf{x}_0$ and $\mathbf{\Phi}$, Zupanski (1997). Therefore the first variation of (A1) is defined as

$$\delta J = \delta \mathbf{x}_0^T \nabla_{\mathbf{x}_0} J + \delta \mathbf{\Phi}^T \nabla_{\mathbf{\Phi}} J \equiv \begin{pmatrix} \nabla_{\mathbf{x}_0} J & \mathbf{0} \\ \mathbf{0} & \nabla_{\mathbf{\Phi}} J \end{pmatrix} \begin{pmatrix} \delta \mathbf{x}_0 \\ \delta \mathbf{\Phi} \end{pmatrix} = \mathbf{0}. \tag{A4}$$

We require (A4) to be true $\forall \delta \mathbf{x}_0, \delta \mathbf{\Phi} \in \mathbb{R}^{N_s}$ which therefore implies that the two gradients must both be zero simultaneously for us to solve the problem.

We start by considering the first variation of (A1) with respect to $\delta \mathbf{x}_0$, $\delta \mathbf{y}_m$ and $\delta \mathbf{\Phi}$, combined with the first variations of (A2) and (A3) which results in

$$\delta J = \delta \mathbf{x}_0^T \mathbf{B}^{-1}(\mathbf{x}_0 - \mathbf{x}_{b,0}) - \sum_{i=1}^{N_o} \delta \mathbf{y}_m^T \mathbf{R}_i^{-1}(\mathbf{y}_i - \mathbf{y}_m) + \delta \mathbf{\Phi}^T \mathbf{Q}^{-1}\mathbf{\Phi}, \tag{A5}$$

$$\delta \mathbf{x}_m = \mathbf{M}_m \delta \mathbf{x}_{m-1} + \delta \mathbf{\Phi}, \tag{A6}$$

$$\delta \mathbf{y}_m = \mathbf{H}_i \delta \mathbf{x}_m. \tag{A7}$$

However, we require (A5) only in terms of $\delta \mathbf{x}_0$ and $\delta \mathbf{\Phi}$, therefore substituting (A6) into (A7) it is possible obtain an expression

for $\delta \mathbf{y}_m$ in terms of $\delta \mathbf{x}_0$ and $\delta \mathbf{\Phi}$. By substituting (A6) for $\delta \mathbf{x}_m$ we arrive at

$$\delta \mathbf{y}_m = \mathbf{H}_i(\mathbf{M}_m \delta \mathbf{x}_{m-1} + \delta \mathbf{\Phi}). \tag{A8}$$

The step just proformed can be repeated for $\delta \mathbf{x}_{m-1}$ which results in

$$\begin{aligned}\delta \mathbf{y}_m &= \mathbf{H}_i \left(\mathbf{M}_m \left(\mathbf{M}_{m-1} \delta \mathbf{x}_{m-2} + \delta \mathbf{\Phi}\right) + \delta \mathbf{\Phi}\right) \\ &\equiv \mathbf{H}_i \left(\mathbf{M}_m \mathbf{M}_{m-1} \delta \mathbf{x}_{m-2} + \mathbf{M}_m \delta \mathbf{\Phi} + \delta \mathbf{\Phi}\right).\end{aligned} \tag{A9}$$

Continuing the substitution as detailed above to $t = 0$ results in

$$\begin{aligned}\delta \mathbf{y}_m = {} & \mathbf{H}_i \mathbf{M}_m \mathbf{M}_{m-1} \ldots \mathbf{M}_1 \delta_{\mathbf{x}_0} + \mathbf{H}_i (\mathbf{M}_m \mathbf{M}_{m-1} \ldots \mathbf{M}_2 \\ & + \mathbf{M}_m \mathbf{M}_{m-1} \ldots \mathbf{M}_3 + \ldots \mathbf{M}_m \mathbf{M}_{m-1} + \mathbf{M}_m + \mathbf{I}) \delta \mathbf{\Phi}.\end{aligned} \tag{A10}$$

Substituting (A10) into (A5) results in

$$\begin{aligned}\delta J = {} & \delta \mathbf{x}_o^T \left(\mathbf{B}^{-1}(\mathbf{x}_0 - \mathbf{x}_{b,0}) - \sum_{i=1}^{N_o} \mathbf{M}_{m,0}^T \mathbf{H}_i^T \mathbf{R}_i^{-1} (\mathbf{y}_i - \mathbf{y}_m)\right) \\ & - \delta \mathbf{\Phi}^T \left(\sum_{i=1}^{N_o} \left(\mathbf{I} + \sum_{l=1}^{m-2} \mathbf{M}_{m,m-l}^T\right)\right. \\ & \left. \times \mathbf{H}_i^T \mathbf{R}_i^{-1} (\mathbf{y}_i - \mathbf{y}_m) + \mathbf{Q}^{-1} \mathbf{\Phi}\right).\end{aligned} \tag{A11}$$

It can clearly be seen in (A11) that the first line is $\nabla_{\mathbf{x}_0} J$ and the second line is $\nabla_{\mathbf{\Phi}} J$. Therefore, the problem is to find $\mathbf{x}_0$ and $\mathbf{\Phi}$ such that (A11) is approximately zero.

9. Appendix B: Derivation of the Lognormal model error gradient

In Appendix A, the derivation of the Gaussian model bias gradient was derived, however, for the lognormal case the starting point is quite different. The cost function is defined as

$$\begin{aligned}J(\mathbf{x}_0, \mathbf{\Phi}) = {} & \frac{1}{2} (\ln \mathbf{x}_0 - \mathbf{x}_{b,0})^T \mathbf{B}_0^{-1} (\ln \mathbf{x}_0 - \mathbf{x}_{b,0}) \\ & + (\ln \mathbf{x}_0 - \mathbf{x}_{b,0})^T \mathbf{1}_{N_s} \\ & + \frac{1}{2} \sum_{i=1}^{N_o} (\ln \mathbf{y}_i - \ln \mathbf{y}_k)^T \mathbf{R}_i^{-1} (\ln \mathbf{y}_i - \ln \mathbf{y}_k) \\ & + (\ln \mathbf{y}_i - \ln \mathbf{y}_k)^T \mathbf{1}_{N_{o,i}} \\ & + (\ln \mathbf{\Phi})^T \mathbf{Q}^{-1} (\ln \mathbf{\Phi}) + (\ln \mathbf{\Phi})^T \mathbf{1}_{N_s},\end{aligned} \tag{B1}$$

subject to

$$[\mathbf{x}_k]_i = [\boldsymbol{\mathcal{M}}_k(\mathbf{x}_{k-1})]_i \, \mathbf{\Phi}_i, \tag{B2}$$

$$\mathbf{y}_k = \mathbf{h}_i(\mathbf{x}_k). \tag{B3}$$

Taking the first variation of the problem described in (B1) we have

$$\begin{aligned}\delta J = {} & \delta \mathbf{x}_0^T \mathbf{W}_B^T \mathbf{B}_0^{-1} ((\ln \mathbf{x} - \ln \mathbf{x}_{b,0}) + \mathbf{1}_{N_s}) \\ & - \sum_{i=1}^{N_o} \delta \mathbf{y}_k^T \mathbf{W}_{o,i}^T \left(\mathbf{R}_i^{-1} (\ln \mathbf{y}_i - \ln \mathbf{y}_k) + \mathbf{1}_{N_{o,i}}\right) \\ & + \delta \mathbf{\Phi}^T \mathbf{W}_m^T \left(\mathbf{Q}^{-1} (\ln \mathbf{\Phi}) + \mathbf{1}_{N_s}\right),\end{aligned} \tag{B4}$$

subject to

$$\delta \mathbf{x}_k = \widehat{\mathbf{M}}_k \delta \mathbf{x}_{k-1} + \widehat{\boldsymbol{\mathcal{M}}}_k \delta \mathbf{\Phi}, \tag{B5}$$

$$\delta \mathbf{y}_k = \mathbf{H}_i \delta \mathbf{x}_k, \tag{B6}$$

where

$$\begin{aligned}[\widehat{\mathbf{M}}_k]_{i,j} &= \frac{\partial \boldsymbol{\mathcal{M}}_k (\mathbf{x}_{k-1})_i}{\partial x_j} \Phi_i \qquad i = 1, 2, \ldots, N_s, \\ &\qquad\qquad\qquad\qquad\quad j = 1, 2, \ldots, N_s \\ [\widehat{\boldsymbol{\mathcal{M}}}_k]_{i,j} &= \boldsymbol{\mathcal{M}}_k (\mathbf{x}_{k-1})_i \frac{\partial \Phi_i}{\partial \Phi_j}, \quad i = 1, 2, \ldots, N_s, \\ &\qquad\qquad\qquad\qquad\quad j = 1, 2 \ldots, N_s, \\ \mathbf{W}_m &= diag\left(\frac{1}{\Phi_n}\right), \qquad n = 1, 2, \ldots, N_s.\end{aligned}$$

The first new feature that makes the lognormal approach different to the Gaussian is the inclusion of the model bias term in (B5). However, following the same argument as for the Gaussian case we need to keep expanding (B5) down to $\delta \mathbf{x}_0$. The first expansion of (B5) is

$$\delta \mathbf{x}_k = \widehat{\mathbf{M}}_k (\widehat{\mathbf{M}}_{k-1} \delta \mathbf{x}_{k-2}) + \widehat{\mathbf{M}}_k \widehat{\boldsymbol{\mathcal{M}}}_{k-1} \delta \mathbf{\Phi} + \widehat{\boldsymbol{\mathcal{M}}}_k \delta \mathbf{\Phi}. \tag{B7}$$

Continuing this expansion to $\delta \mathbf{x}_{k-2}$ results in

$$\begin{aligned}\delta \mathbf{x}_k = {} & \widehat{\mathbf{M}}_k \widehat{\mathbf{M}}_{k-1} \widehat{\mathbf{M}}_{k-2} \delta \mathbf{x}_{k-3} + \widehat{\mathbf{M}}_k \widehat{\mathbf{M}}_{k-1} \widehat{\boldsymbol{\mathcal{M}}}_{k-2} \delta \mathbf{\Phi} \\ & + \widehat{\mathbf{M}}_k \widehat{\boldsymbol{\mathcal{M}}}_{k-1} \delta \mathbf{\Phi} + \widehat{\boldsymbol{\mathcal{M}}}_k \delta \mathbf{\Phi}.\end{aligned}$$

Following the derivation to $\delta \mathbf{x}_0$ results in

$$\begin{aligned}\delta \mathbf{x}_k = {} & \widehat{\mathbf{M}}_k \widehat{\mathbf{M}}_{k-1} \ldots \widehat{\mathbf{M}}_1 \delta \mathbf{x}_0 + \widehat{\mathbf{M}}_k \widehat{\mathbf{M}}_{k-1} \ldots \widehat{\mathbf{M}}_2 \widehat{\boldsymbol{\mathcal{M}}}_1 \delta \mathbf{\Phi} \\ & + \widehat{\mathbf{M}}_k \widehat{\mathbf{M}}_{k-1} \ldots \widehat{\mathbf{M}}_3 \widehat{\boldsymbol{\mathcal{M}}}_2 \delta \mathbf{\Phi} + \ldots + \widehat{\mathbf{M}}_k \widehat{\boldsymbol{\mathcal{M}}}_{k-1} \delta \mathbf{\Phi} \\ & + \widehat{\boldsymbol{\mathcal{M}}}_k \delta \mathbf{\Phi},\end{aligned} \tag{B8}$$

which can be simplified to

$$\delta \mathbf{x}_k = \widehat{\mathbf{M}}_{k,1} \delta \mathbf{x}_0 + \left(\widehat{\boldsymbol{\mathcal{M}}}_k + \sum_{n=0}^{k-2} \widehat{\mathbf{M}}_{k,k-n} \widehat{\boldsymbol{\mathcal{M}}}_{k-n-1}\right) \delta \mathbf{\Phi}, \tag{B9}$$

where

$$\widehat{\mathbf{M}}_{k,k-n} = \widehat{\mathbf{M}}_k \widehat{\mathbf{M}}_{k-1} \ldots \widehat{\mathbf{M}}_{k-n}.$$

However, it is the adjoint of (B9) that is required, which can easily be verified as

$$\delta \mathbf{x}_k^T = \delta \mathbf{x}_0^T \widehat{\mathbf{M}}_{1,k}^T + \delta \mathbf{\Phi}^T \left(\widehat{\boldsymbol{\mathcal{M}}}_{k,k-1}^T + \sum_{n=0}^{k-2} \widehat{\boldsymbol{\mathcal{M}}}_{k-n-1}^T \widehat{\mathbf{M}}_{k-n,k}^T\right), \tag{B10}$$

where

$$\widehat{\mathbf{M}}_{k-n,k}^T = \widehat{\mathbf{M}}_{k-n}^T \widehat{\mathbf{M}}_{k-n+1}^T \ldots \widehat{\mathbf{M}}_k^T.$$

Therefore, substituting (B10) into (B6) and then into (B4) results in the first variation of (B1) as

$$\begin{aligned}\delta J = {} & \delta \mathbf{x}_0^T \mathbf{W}_B^T \mathbf{B}_0^{-1} \left((\ln \mathbf{x} - \ln \mathbf{x}_{b,0}) + \mathbf{1}_{N_s} \right) \\ & - \sum_{i=1}^{N_o} \delta \mathbf{x}_0^T \widehat{\mathbf{M}}_{1,k}^T \mathbf{H}_i^T \mathbf{W}_{o,i}^T \left(\mathbf{R}_i^{-1} (\ln \mathbf{y}_i - \ln \mathbf{y}_k) + \mathbf{1}_{N_{o,i}} \right) \\ & - \sum_{i=1}^{N_o} \delta \boldsymbol{\Phi}^T \left(\widehat{\boldsymbol{\mathcal{M}}}_k^T + \sum_{n=0}^{k-2} \widehat{\boldsymbol{\mathcal{M}}}_{k-n-1}^T \widehat{\mathbf{M}}_{k-n,k}^T \right) \\ & \times \mathbf{H}_i^T \mathbf{W}_{o,i}^T \left(\mathbf{R}_i^{-1} (\ln \mathbf{y}_i - \ln \mathbf{y}_k) + \mathbf{1}_{N_{o,i}} \right) \qquad \text{(B11)} \\ & + \delta \boldsymbol{\Phi}^T \mathbf{W}_m^T \left(\mathbf{Q}^{-1} \ln \boldsymbol{\Phi} + \mathbf{1}_{N_s} \right). \qquad \text{(B12)}\end{aligned}$$

10. Appendix C: Mixed type model error gradient

In this appendix the information that has been presented in the last two appendices are combined to give the case where the model error is distributed a mixed Gaussian-lognormal distribution. The formulation of the cost function is

$$\begin{aligned}J(\mathbf{x}_0, \boldsymbol{\Phi}) = {} & \frac{1}{2} \begin{pmatrix} \mathbf{x}_{0,p} - \mathbf{x}_{b,0,p} \\ \ln \mathbf{x}_{0,q} - \ln \mathbf{x}_{b,0,q} \end{pmatrix}^T \mathbf{B}_0^{-1} \begin{pmatrix} \mathbf{x}_{0,p} - \mathbf{x}_{b,0,p} \\ \ln \mathbf{x}_{0,q} - \ln \mathbf{x}_{b,0,q} \end{pmatrix} \\ & + \begin{pmatrix} \mathbf{x}_{0,p} - \mathbf{x}_{b,0,p} \\ \ln \mathbf{x}_{0,q} - \ln \mathbf{x}_{b,0,q} \end{pmatrix}^T \begin{pmatrix} \mathbf{0}_p \\ \mathbf{1}_q \end{pmatrix} \\ & + \frac{1}{2} \sum_{i=1}^{N_o} \begin{pmatrix} \mathbf{y}_{p_i,i} - \mathbf{y}_{p_i,i,k} \\ \ln \mathbf{y}_{q_i,i} - \ln \mathbf{y}_{q_i,i,k} \end{pmatrix}^T \mathbf{R}_i^{-1} \begin{pmatrix} \mathbf{y}_{p_i,i} - \mathbf{y}_{p_i,i,k} \\ \ln \mathbf{y}_{q_i,i} - \ln \mathbf{y}_{q_i,i,k} \end{pmatrix} \\ & + \begin{pmatrix} \mathbf{y}_{p_i,i} - \mathbf{y}_{p_i,i,k} \\ \ln \mathbf{y}_{q_i,i} - \ln \mathbf{y}_{q_i,i,k} \end{pmatrix}^T \begin{pmatrix} \mathbf{0}_{p_i,i} \\ \mathbf{1}_{q_i,i} \end{pmatrix} \\ & + \frac{1}{2} \begin{pmatrix} \boldsymbol{\Phi}_p \\ \ln \boldsymbol{\Phi}_q \end{pmatrix}^T \mathbf{Q}^{-1} \begin{pmatrix} \boldsymbol{\Phi}_p \\ \ln \boldsymbol{\Phi}_q \end{pmatrix} + \begin{pmatrix} \boldsymbol{\Phi}_p \\ \ln \boldsymbol{\Phi}_q \end{pmatrix}^T \begin{pmatrix} \mathbf{0}_p \\ \mathbf{1}_q \end{pmatrix},\end{aligned} \qquad \text{(C1)}$$

subject to

$$\mathbf{x}_{k,p} = \boldsymbol{\mathcal{M}}_k(\mathbf{x}_{k-1}) + \boldsymbol{\Phi}_p, \qquad \text{(C2)}$$

$$[\mathbf{x}_{k,q}]_s = [\boldsymbol{\mathcal{M}}_k(\mathbf{x}_{k-1})]_s [\boldsymbol{\Phi}_q]_s, \quad s = p+1, p+2, \ldots, q, \qquad \text{(C3)}$$

$$\mathbf{y}_{i,k} = \mathbf{h}_i(\mathbf{x}_k). \qquad \text{(C4)}$$

As with the Gaussian and lognormal constant bias case we require the first variation of (C1) with respect to $\delta \mathbf{x}_0$ and $\delta \boldsymbol{\Phi}$. Therefore following the steps presented in Appendices A and B we have

$$\begin{aligned}\delta J = {} & \delta \mathbf{x}_0^T \widehat{\mathbf{W}}_b^T \left(\mathbf{B}_0^{-1} \begin{pmatrix} \mathbf{x}_{0,p} - \mathbf{x}_{b,0,p} \\ \ln \mathbf{x}_{0,q} - \ln \mathbf{x}_{b,0,q} \end{pmatrix} + \begin{pmatrix} \mathbf{0}_p \\ \mathbf{1}_q \end{pmatrix} \right) \\ & - \sum_{i=1}^{N_o} \delta \mathbf{y}_k^T \widehat{W}_{o,i}^T \left(\mathbf{R}_i^{-1} \begin{pmatrix} \mathbf{y}_{p_i,i} - \mathbf{y}_{p_i,i,k} \\ \ln \mathbf{y}_{q_i,i} - \ln \mathbf{y}_{q_i,i,k} \end{pmatrix} + \begin{pmatrix} \mathbf{0}_{p_i,i} \\ \mathbf{1}_{q_i,i} \end{pmatrix} \right) \\ & + \delta \boldsymbol{\Phi}^T \widehat{W}_m^T \left(\mathbf{Q}^{-1} \begin{pmatrix} \boldsymbol{\Phi}_p \\ \ln \boldsymbol{\Phi}_q \end{pmatrix} + \begin{pmatrix} \mathbf{0}_p \\ \mathbf{1}_q \end{pmatrix} \right), \qquad \text{(C5)}\end{aligned}$$

subject to

$$\delta \mathbf{x}_k = \widetilde{\mathbf{M}}_k \delta \mathbf{x}_{k-1} + \widetilde{\mathbf{M}}_k \delta \boldsymbol{\Phi}, \qquad \text{(C6)}$$

$$\delta \mathbf{y}_k = \mathbf{H}_i \delta \mathbf{x}_k, \qquad \text{(C7)}$$

where

$$\widetilde{\mathbf{M}}_k = \begin{pmatrix} \mathbf{M}_{k,p,p} & \mathbf{M}_{k,p,q} \\ \widehat{\mathbf{M}}_{k,q,p} & \widehat{\mathbf{M}}_{k,q,q} \end{pmatrix} \quad \text{and} \quad \widetilde{\boldsymbol{\mathcal{M}}}_k = \begin{pmatrix} \mathbf{I}_{p,p} & \mathbf{0}_{p,q} \\ \mathbf{0}_{q,p} & \widehat{\boldsymbol{\mathcal{M}}}_{k,q,q} \end{pmatrix}.$$

Following the arguments used in Appendices A and B then $\delta \mathbf{x}_k$ can be written as

$$\delta \mathbf{x}_k = \widetilde{\mathbf{M}}_{k,1} \delta \mathbf{x}_0 + \left(\widetilde{\boldsymbol{\mathcal{M}}}_k + \sum_{n=0}^{k-2} \widetilde{\mathbf{M}}_{k,k-n} \widetilde{\boldsymbol{\mathcal{M}}}_{k-n-1} \right) \delta \boldsymbol{\Phi}. \qquad \text{(C8)}$$

Substituting the transpose of (C8) into (C5) results in

$$\begin{aligned}\delta J = {} & \delta \mathbf{x}_0^T \widehat{\mathbf{W}}_b^T \left(\mathbf{B}_0^{-1} \begin{pmatrix} \mathbf{x}_{0,p} - \mathbf{x}_{b,0,p} \\ \ln \mathbf{x}_{0,q} - \ln \mathbf{x}_{b,0,q} \end{pmatrix} + \begin{pmatrix} \mathbf{0}_p \\ \mathbf{1}_q \end{pmatrix} \right) \\ & - \sum_{i=1}^{N_o} \delta \mathbf{x}_0^T \widetilde{\mathbf{M}}_{1,k}^T \mathbf{H}_i^T \widehat{W}_{o,i}^T \left(\mathbf{R}_i^{-1} \begin{pmatrix} \mathbf{y}_{p_i,i} - \mathbf{y}_{p_i,i,k} \\ \ln \mathbf{y}_{q_i,i} - \ln \mathbf{y}_{q_i,i,k} \end{pmatrix} \right. \\ & \left. + \begin{pmatrix} \mathbf{0}_{p_i,i} \\ \mathbf{1}_{q_i,i} \end{pmatrix} \right) - \sum_{i=1}^{N_0} \delta \boldsymbol{\Phi}^T \left(\widetilde{\boldsymbol{\mathcal{M}}}_k^T + \sum_{n=0}^{k-2} \widetilde{\boldsymbol{\mathcal{M}}}_{k-n-1}^T \widetilde{\mathbf{M}}_{k-n,k}^T \right) \\ & \times \mathbf{H}_i^T \widehat{\mathbf{W}}_{o,i}^T \left(\mathbf{R}_i^{-1} \begin{pmatrix} \mathbf{y}_{p_i,i} - \mathbf{y}_{p_i,i,k} \\ \ln \mathbf{y}_{q_i,i} - \ln \mathbf{y}_{q_i,i,k} \end{pmatrix} + \begin{pmatrix} \mathbf{0}_{p_i,i} \\ \mathbf{1}_{q_i,i} \end{pmatrix} \right) \\ & + \delta \boldsymbol{\Phi} \widehat{\mathbf{W}}_m \left(\mathbf{Q}^{-1} \begin{pmatrix} \boldsymbol{\Phi}_p \\ \ln \boldsymbol{\Phi}_q \end{pmatrix} + \begin{pmatrix} \mathbf{0}_p \\ \mathbf{1}_q \end{pmatrix} \right). \qquad \text{(C9)}\end{aligned}$$

References

Beaulieu, N. C. and Rajwani, N. 2004. Highly accurate simple closed form approximations to lognormal sum distributions and densities. *IEEE Comm. Lett.* **9**, 709–711.

Clarke, G. M. and Cooke, D. 2004. *A Basic Course in Statistics*. Oxford University Press, New York, USA, 734 pp.

Cohn, S. E. 1997. An introduction to estimation theory *J. Meteor. Soc. Jpn.* **75**, 257–288.

Courtier, P. and Talagrand, O. 1990. Variational assimilation of meteorological observations with the direct and adjoint shallow-water equations. *Tellus* **42A**, 531–549.

Daley, R. 1992. The effects of serially correlated observation and model errors on atmospheric data assimilation. *Mon. Wea. Rev.* **164**, 120–177.

Fletcher, S. J. and Zupanski, M 2006a. A data assimilation method for log-normally distributed observational errors. *Quart. J. Roy. Meteor. Soc.* **132**, 2505–2519.

Fletcher, S. J. and Zupanski, M 2006b. A hybrid normal and lognormal distribution for data assimilation. *Atmos. Sci. Lett.* **7**, 43–46.

Fletcher, S. J. and Zupanski, M 2007. Implications and impacts of transforming lognormal variables into normal variables in VAR. *Meteorologische Zeitschrift* **16**, 755–765.

Le Dimet, F.-X. and Talagrand, O. 1986. Variational algorithm for analysis and assimilation adjustment problem with advective constraints. *Tellus* **38A**, 97–110.

Lewis, J. M. and Derber, J. C. 1985. The use of adjoints equations to solve a variational adjustment problem with advective constraints. *Tellus* **37A**, 309–322.

Lorenc, A. C. 1986. Analysis methods for numerical weather prediction. *Q. J. Roy. Meteor. Soc* **112**, 1177–1194.

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

Mielke, P. W., Williams, J. S. and Wu, S.-C. 1977. Covariance analysis techniques based upon bivariate lognormal distribution with modification application. *J. Appl. Met.* **16**, 183–187.

Miles, M. L., Verlinde, J. and Clothiaux, E. E. 2000. Cloud droplet size distribution in low-level stratiform clouds. *J. Atmos. Sci.* **57**, 295–311.

Pearl, J. 2007. *Causality, Models, Reasoning and Inference* Cambridge University Press., New York, USA., 384 pp.

Pedlosky, J. 1987. *Geophysical Fluid Dynamics* Springer, New York, USA., 710 pp.

Polavarapu, S., Ren, S., Rochen, Y., Sankey, D., Ek, N. and co-authors. 2005. Data assimilation with the Canadian middle atmosphere model. *Atmos.-Ocean* **43**(1), 77–100.

Sasaki, Y. 1970 Some basic formalisms in numerical variational analysis. *Mon. Wea. Rev.*, **98**, 875–883.

Sengupta, M., Clothiaux, E. E. and Ackerman, T. P. 2002 Climatology of warm boundary layers clouds at the ARM SGP site and their comparisons to models. *J. Clim.*, **17**, 4760–4782.

Talagrand, O. 1988 Four-dimensional variational data assimilation. *ECMWF Seminar Proceedings on Data Assimilation and the use of Satellite Data.*, ECMWF, Reading, U.K. 1–30.

Thépaut, J.-N. and Courtier, P., 1991. Four-dimensional variational data assimilation using the adjoint of a multilevel primitive equation model. *Quart. J. Roy. Meteor. Soc.*, **117**, 1225–1254.

Thépaut, J.-N., Hoffman, R. N. and Courtier, P., 1993. Interactions of dynamics and observations in four-dimensional variational assimilation. *Mon Wea. Rev.*, **121**, 3393–3414.

Van Leeuwen, P. J. and Evensen, G. 1996. Data assimilation and inverse methods in terms of a probabilistic formulation. *Mon. Wea. Rev.*, **124**, 2898–2913.

Zupanksi, D. 1997. A general weak constraint applicable to operational 4DVAR data assimilation systems. *Mon. Wea. Rev*, **125**, 2274–2292.