4.7

AN ALTERNATIVE TO BIAS CORRECTION IN RETRIEVALS AND DIRECT RADIANCES ASSIMILATION

Steven J. Fletcher* and Milija Zupanski
Cooperative Institute for Research in the Atmosphere,
Colorado State University,
Fort Collins, CO

1. Introduction

It is well known that the foundations of 3-dimensional variational data assimilation (3D VAR DA) is based upon an assumption made about the variables, observations and therefore the errors being Normal, Gaussian, distributed, (Lorenc, 1986). However, we do not live in a *Normal* world and actually there are several variables that are positive definite especially as we enter meso-scale and cloud resolving data assimilation regimes, (Mielke et al., 1977; Miles et al., 2000; Sengupta et al., 2004). This problem of non-Gaussian variables is not solely restricted to the geosciences and a good review of where positive definite variables which are often approximated with a lognormal distribution can be found in (Limpert et al., 2001) and also in the medical sciences (Townsend, 2004).

This problem of non-Normal variables may appear just as a concern for the smaller dimensional modelling and assimilation fields but we have a similar problem with moisture variables in the synoptic scale data assimilation. As pointed out in (Dee and Da Silva, 2003) there are many different options to chose from. The one that we are concerned with is the logarithm of the specific humidity which is used in the Canadian Meteorological Service's mid-atmosphere model and assimilation (Polavarapu et al., 2005). Another reason for an interest in this variable is that is the variable that is retrieved in a 1-D VAR humidity retrieval, (Poli et al., 2002; Deblonde and English, 2003).

Although it may appear easier to take the logarithm of the humidity, i.e. if humidity is lognormal then log of humidity is Normal, this transform between distributions is not a 1 to 1 invertible relationship and hence information is lost.

On the observational side, with the advances in satellites we have cloud variable observations that are showing signs of being lognormally distributed, (Stephens and Coauthors, 2002), along with climatologies from radiosondes also showing the non-Normal structure, (Soden and Lanzante, 1996).

Given these motivations the aim of this paper is to highlight the bias which is introduced with two methods currently used to deal with approximately lognormal variables and observations. The two techniques that we investigate are transforming the lognormal variable to a Normal variable and the second technique is to assume a Normal distribution for the variables and observations. The first technique is used in humidity retrievals, (Poli et al., 2002; Deblonde and English, 2003), whilst the second if used for direct radiance assimilation, (Derber and Wu, 1998). We present these two techniques in the next section.

For us to be able to quantify this bias we introduce the theory from (Fletcher and Zupanski, 2006a) where lognormal observations assimilation is address through using the maximum likelihood estimator for the lognormal distribution. The reason for this choice is that the maximum likelihood estimator for the lognormal distribution is the only bounded and unique of the three estimators for this distribution, (Heyde, 1963; Evans et al., 2000; Fletcher and Zupanski, 2006a).

In Section 3 we briefly summarise the results from (Fletcher and Zupanski, 2006a) which are the associated cost function for lognormal observations and the extension to lognormal backgrounds. We also present the nonlinear solution to the maximum likelihood framework to compare to the results for the two current methods.

In Section 4 we present the difference in the solutions between assuming a transform of the variable and assimilating with the modal model present in Section 3. We also show the difference in the solution when a Normal distribution is used for a lognormal variable and show the difference in the solutions again in Section 5. We finish with some brief remarks.

2. Retrievals and Direct Radiances Assimilation

The importance of the moisture field is well known and there has been a review in (Dee and Da Silva, 2003) which looks at the pros and cons of the different methods used at the operational centres. The method to obtain a humidity field from a brightness temperature field can be derived from considering the following 1-D cost function, i.e defined in the z coordinate,

$$\begin{aligned} J(\mathbf{x}) &= \frac{1}{2}(\mathbf{x}-\mathbf{x_b})^T \mathbf{B}^{-1}(\mathbf{x}-\mathbf{x_b}) \\ &+ (\mathbf{y}-\mathbf{h}(\mathbf{x}))^T \mathbf{R}^{-1}(\mathbf{y}-\mathbf{h}(\mathbf{x})), \end{aligned} \quad (1)$$

where $\mathbf{x}$ is our analysis state vector, $\mathbf{x_b}$ is our background or first guess state, $\mathbf{B}$ is the background covariance matrix comprising of the correlations and standard deviations. Note: Correlations are only representing the linear relationship between the random variables. The vector $\mathbf{y}$

*Corresponding author address: Dr. Steven J. Fletcher, Cooperative Institute for Research in the Atmosphere, Colorado State University, 1375 Campus Delivery, Fort Collins, CO 80523-1375; Email: fletcher@cira.colostate.edu

contains the brightness temperature observations, $\mathbf{h}(\mathbf{x})$ is the **non-linear** forward operator which transforms the state vectors to the observations and $\mathbf{R}$ is the observation covariance matrix contain defined as $\mathbf{R} = \mathbf{E} + \mathbf{F}$ where $\mathbf{E}$ is the instrument error covariance matrix and $\mathbf{F}$ is the representativeness error covariance matrix, (Deblonde and English, 2003).

The basis of a specific humidity retrieval is to transform the state variable, specific humidity, $\mathbf{q}$, to the new transformed variable

$$\mathbf{Q} = \ln \mathbf{q}. \tag{2}$$

If we assume that $\mathbf{Q}$ is now Normally distributed then the original variable, specific humidity is lognormally distributed. The main advantage of the lognormal distribution is that is defined in terms of the statistical parameters of $\ln \mathbf{q}$, i.e.

$$\mathbf{q} \sim LN\left(\boldsymbol{\mu}, \sigma^2\right), \tag{3}$$

where

$$\boldsymbol{\mu} = \mathrm{E}\left(\ln \mathbf{q}\right) \quad \text{and} \quad \sigma^2 = \mathrm{E}\left(\left(\ln \mathbf{q}\right)^2\right) - \mathrm{E}\left(\ln \mathbf{q}\right)^2. \tag{4}$$

The importance of (4) is that it highlights the fact that the statistics that are calculated for the Normal variable, $\mathbf{Q}$, are consistent with the distribution of the lognormal variable, $\mathbf{q}$. This fact is important when we consider the analysis due to transforming from Normal to lognormal variables in Section 4.

The advantage of transforming is that the variable $\mathbf{Q}$ fits into the current Normal frameworks and therefore no changes are needed to the minimization algorithms to minimize the cost function. After the analysis is performed we simple invert (2) which then gives us a value for the model state of specific humidity. However, this introduces a bias in the analysis and we quantify this in Section 4.

Let us now consider direct radiance assimilation using the Normal variational cost function. This process is non-linear causing and hence raises doubts about the assumption that $\mathbf{y}$ and $\mathbf{h}(\mathbf{x})$ are Normally distributed. A more detailed explanation of direct radiance assimilation algorithm can be found in (Derber and Wu, 1998). The reason to look at the direct radiance assimilation method is that a form of bias correction has to be applied to the analysis state from the minimization of the cost function defined in (1). Therefore, by assuming a Normal framework for non-Normal variables we have introduced an error which we try to correct through some form of bias analysis. We investigate this bias in Section 5.

3. Maximum likelihood approach for Lognormal data assimilation

In this section we summarise the maximum likelihood framework derived in Fletcher and Zupanski (2006a). As we have mentioned before the mode of a multivariate lognormal distribution is the only unique and bounded estimator of this distribution, see Fletcher and Zupanski (2006a) for a full detailed discussion of the three estimators. Another advantage of this method is that we do not assume any linearity of the observation operator therefore allow for more realistic cases.

We start with the definition of the relative observation error. This come from (Cohn, 1997) as the ratio of two lognormal variables is itself lognormal, (Townsend, 2004). The definition of the errors is given by

$$\varepsilon_i^o = \frac{y_i}{h_i(\mathbf{x})}, \quad i = 1, 2, \ldots, N_o. \tag{5}$$

To derive the maximum likelihood approach we start from the Bayesian model in (Lorenc, 1986) and use the definition of the multivariate lognormal distribution given by

$$\begin{aligned} f(\mathbf{x}) &= \frac{1}{(2\pi)^{\frac{n}{2}} |\Sigma|^{\frac{1}{2}}} \left(\prod_{i=1}^{n} \frac{1}{x_i}\right) \\ &\times \exp\left\{-\frac{1}{2}\left(\ln \mathbf{x} - \boldsymbol{\mu}\right)^T \Sigma^{-1} \left(\mathbf{x} - \boldsymbol{\mu}\right)\right\}, \end{aligned} \tag{6}$$

where f is the proability density function, pdf, n is the dimensional of $\mathbf{x}$, $\boldsymbol{\mu}_i = \mathrm{E}(\ln \mathbf{x}_i)$, Σ is the covariance matrix of $\ln \mathbf{x}$ and $|\Sigma|$ is the determinant of Σ. Note: That if a bias is known then μ can be different from zero to compensate in the minimization for this.

The important term in (6) is the product of the x_i's. This is the term that enables us to transform between the lognormal and the Normal distribution but also, as shown in (Fletcher and Zupanski, 2006a), this terms enables us to obtain the mode of the analysis distribution.

From the definition of the observational errors which we are going to assume are lognormally distributed, (5), and the definition of the lognormal pdf (6) we obtain the following conditional pdf as set out in (Lorenc, 1986) and (Fletcher and Zupanski, 2006a) which is

$$\begin{aligned} f(\boldsymbol{\varepsilon}^o) = \frac{1}{(2\pi)^{\frac{N_o}{2}} |\Sigma|^{\frac{1}{2}}} \left(\prod_{i=1}^{N_o} \frac{\mathbf{h}_i(\mathbf{x})}{\mathbf{y}_i}\right) \times \\ \exp\left\{-\frac{1}{2}\left(\ln \mathbf{y} - \ln \mathbf{h}(\mathbf{x})\right)^T \mathbf{R}^{-1} \left(\ln \mathbf{y} - \ln \mathbf{h}(\mathbf{x})\right)\right\}, \end{aligned} \tag{7}$$

where we have assumed that $\boldsymbol{\mu} = \mathbf{0}$ so that the observations are unbiased, $\mathbf{R}$ is the standard Normal observational covariance matrix and N_o is the total number of observations.

To form the maximum likelihood approach we take th negative logarithm of (7). This then results in the cost function of the form

$$\begin{aligned} J_o(\mathbf{x}) &= \frac{1}{2}\left(\ln \mathbf{y} - \ln \mathbf{h}(\mathbf{x})\right)^T \mathbf{R}^{-1} \left(\ln \mathbf{y} - \ln \mathbf{h}(\mathbf{x})\right) \\ &+ \left(\ln \mathbf{y} - \ln \mathbf{h}(\mathbf{x})\right)^T \mathbf{1}_{N_o}, \end{aligned} \tag{8}$$

where $J_o(\mathbf{x})$ represents the observational component of the full 3D VAR type cost function and $\mathbf{1}_n$ is a vector of 1's of dimension $n \times 1$. The reason for writing the summation that appears in the second line of (8) as the product of vectors is because this helps in highlighting the difference between transforming and not transforming in Section 5.

We also have to allow for lognormal background variables as this is the situation in the humidity retrievals

case. We start with the definition of the lognormal background error given by

$$\boldsymbol{\varepsilon}_{b,i} = \frac{\mathbf{x}_i}{\mathbf{x_b}_i}. \tag{9}$$

Using the definition of the lognormal distribution, (6), and taking the negative logarithm of the pdf gives us the lognormal background cost function as

$$\begin{aligned} J_b &= \frac{1}{2}\left(\ln \mathbf{x} - \ln \mathbf{x_b}\right)^T \mathbf{B}^{-1}\left(\ln \mathbf{x} - \ln \mathbf{x_b}\right) \\ &+ \left(\ln \mathbf{x} - \ln \mathbf{x_b}\right)^T \mathbf{1}_N, \end{aligned} \tag{10}$$

where N is the total number of state variables.

Combining (10) and (8) we obtain the full 3D lognormal VAR cost function, $J(\mathbf{x})$, defined as

$$\begin{aligned} J(\mathbf{x}) &= \frac{1}{2}\left(\ln \mathbf{x} - \ln \mathbf{x_b}\right)^T \mathbf{B}^{-1}\left(\ln \mathbf{x} - \ln \mathbf{x_b}\right) \\ &+ \frac{1}{2}\left(\ln \mathbf{y} - \ln \mathbf{h}(\mathbf{x})\right)^T \mathbf{R}^{-1}\left(\ln \mathbf{y} - \ln \mathbf{h}(\mathbf{x})\right) \\ &+ \left(\ln \mathbf{x} - \ln \mathbf{x_b}\right)^T \mathbf{1}_N + \left(\ln \mathbf{y} - \ln \mathbf{h}(\mathbf{x})\right)^T \mathbf{1}_{N_o}. \end{aligned} \tag{11}$$

For us to be able to identify the bias introduced through transforming or using the wrong distribution we require the solution to (11). It may be obvious at this point that the solution is non-linear but it is the structure of the solution that we are interested in. To obtain the state which minimizes (11) we need to differentiate (11), set to zero and rearrange which gives

$$\begin{aligned} &\mathbf{x} = \mathbf{x_b} \exp\left\{-\mathbf{B}\mathbf{1}_N\right\} \times \\ &\exp\left\{\mathbf{B}\mathbf{W}_b^{-T}\mathbf{W}_o^T\mathbf{H}^T\mathbf{R}^{-1}\left(\ln \mathbf{y} - \ln \mathbf{h}(\mathbf{x}) + \mathbf{R}\mathbf{1}_{N_o}\right)\right\}, \end{aligned} \tag{12}$$

where

$$\begin{aligned} \mathbf{W}_{b(i,i)} &= \frac{1}{x_i}, \qquad i = 1, 2, \ldots, N, \\ \mathbf{W}_{o(j,j)} &= \frac{1}{h_j(\mathbf{x})}, \quad j = 1, 2, \ldots, N_o, \\ \mathbf{H} &= \frac{\partial \mathbf{h}}{\partial \mathbf{x}} \end{aligned}$$

where these two matrices arise from the differentiation of the logarithms in (11).

NOTE 1: We can not find the solution to (12) explicitly as the right hand side of the equation is a function of $\mathbf{x}$. We can, however, use an iterative solver like the quasi-Newton or a conjugate gradient method to iterate to find a solution.

NOTE 2: In the formulation used above we **do not** use the logarithm of the state variable in the observation operator. The reason for this is that the lognormal framework does not allow us to interchange the logarithm and the non-linear observation operator. Therefore it is the variable $\mathbf{h}(\mathbf{x})$ that is lognormal as this is the component that is compared to the observations.

4. Differences between current and new approach for humidity retrievals

In the last section we derived the non-linear solution to (11) given by (12). It is with this solution we can compare how the other two techniques differ from the true solution given by (12).

In this section we are consider the technique where we transform the observations and the state variable to be defined for $\hat{\mathbf{x}} = \ln \mathbf{x}$ which has the following cost function

$$\begin{aligned} J(\hat{\mathbf{x}}) &= \frac{1}{2}\left(\hat{\mathbf{x}} - \hat{\mathbf{x}}_b\right)^T \mathbf{B}^{-1}\left(\hat{\mathbf{x}} - \hat{\mathbf{x}}_b\right) \\ &+ \left(\ln \mathbf{y} - \ln \hat{\mathbf{h}}(\hat{\mathbf{x}})\right)^T \mathbf{R}^{-1}\left(\ln \mathbf{y} - \ln \hat{\mathbf{h}}(\hat{\mathbf{x}})\right). \end{aligned} \tag{13}$$

This then makes (13) a Normal cost function and hence the solution is given by

$$\mathbf{x} = \mathbf{x_b} \exp\left\{\mathbf{B}\hat{\mathbf{W}}_o^T\hat{\mathbf{H}}^T\mathbf{R}^{-1}\left(\ln \mathbf{y} - \ln \hat{\mathbf{h}}(\ln \mathbf{x})\right)\right\}, \tag{14}$$

where

$$\hat{\mathbf{H}} = \frac{\partial \hat{\mathbf{h}}}{\partial \hat{\mathbf{x}}} \text{ and } \hat{\mathbf{W}}_{o(j,j)} = \frac{1}{\hat{h}_j}, \quad j = 1, 2, \ldots, N_o.$$

We can see from (12) that this solution has the appearance of being similar to a lognormal mode as it contains the covariance terms in the solution in both the background and the observational component of the solution. However, if we compare this with (14) we see that this solution has no structure of the mode present. It is actually representing the median of the analysis space which over estimates the most likely state for a multivariate lognormal. The other fact is that the median which (14) represents is the one associated with the transform between the Normal and the lognormal distributions but is not unique.

5. Differences between current and new approach for direct radiance assimilation

We now consider the case where the wrong distribution is used to represent the variables and the observations. The assumption made in direct radiance assimilation is that the observations are Normal. Therefore the cost function that is used is (1) in the full 3D form. Therefore the standard non-linear solution is

$$\mathbf{x} = \mathbf{x_b} + \tilde{\mathbf{B}}\mathbf{H}^T\tilde{\mathbf{R}}^{-1}\left(\mathbf{y} - \mathbf{h}(\mathbf{x})\right), \tag{15}$$

where here we have the tildes above the covariance matrix as these are the best Normal approximation to the data and not the correct lognormal or any other distribution statistics.

If we now compare the solutions (15) and (12) we see that there is a major discrepency between the two and as such there is bias being introduce through both the solution only representing a form of a Taylor series expansion of the exponential but also the wrong statistics are being used for the covariance matrices.

6. Conclusions

In this paper we have try to warn about using transforms for lognormal variables so that they are compatible with the current Normal frameworks in 3D VAR. We have also shown that the assumption that a variable is Normal when it is not can lead to a major bias away from the true solution. The advantage of a full lognormal data assimilation scheme derived from a maximum likelihood approach is that the solution to the cost function is the only unique and bounded estimator of the lognormal distribution, (Fletcher and Zupanski, 2006a). Another advantage of the framework presented in Section 3 is that we have made no assumption about the observation operator being linear.

The plans for this work is to extend the applicability of this method for the operational weather and ocean prediction centers. Another plan is to see the impact on both humidity retrievals and humidity assimilation in general through using a more skewed distribution than the Normal.

This work has been partially extended in (Fletcher and Zupanski, 2006b) where we have derived a hybrid assimilation scheme where both Normal and lognormal observations can be assimilated simultaneously. This would be the more practical approach for the humidity retrievals as the state vector contains the temperature, which is assumed Normal, and the specific humidity which is possible better approximated with a lognormal.

Acknowledgments

We are grateful to the Super Computing Division at the National Center for Atmospheric Research for the use of Bluesky for the numerical implementation of the MLEF. This work funded in part by the National Science Foundation of America's Collaboration in Mathematical Geoscience Grant 0327651 and by the DoD Center for Geosciences/Atmospheric Research at Colorado State University under Cooperative Agreement (W911NF-060200015) with the Army Research Laboratory

REFERENCES

Cohn, S. E., 1997: An introduction to estimation theory. *J. Met. Soc. Japan*, **75**, 257–288.

Deblonde, G. and S. English, 2003: One-dimensional variational retrievals from SSMISS-simulated observations. *J. Appl. Meteor.*, **42**, 1406–1420.

Dee, D. P. and A. M. Da Silva, 2003: The choice of variable for atmospheric moisture analysis. *Mon. Wea. Rev.*, **131**, 155–171.

Derber, J. C. and W.-S. Wu, 1998: The use of TVOS cloud-cleared radiances in the NCEP SSI analysis system. *Mon. Wea. Rev*, **126**, 2287–2299.

Evans, M., N. Hastings, and B. Peacock, 2000: *Statistical distributions*. John Wiley and Sons, Inc., pp 221.

Fletcher, S. J. and M. Zupanski, 2006a: A data assimilation method for lognormally distributed observational errors. *In Print Quart. J. Roy. Meteor. Soc.*

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

Heyde, C. C., 1963: On a property of the lognormal distribution. *J. Roy. Statist. Soc. Ser. B*, **25**, 392–393.

Limpert, E., W. A. Stahel, and M. Abbt, 2001: Log-normal distributions across the sciences: Keys and clues. *BioSciences*, **51**, 341–352.

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

Mielke, P. W., J. S. Williams, and S.-C. Wu, 1977: Covariance analysis technique based on bivariate lognormal distribution with weather modification applications. *J. App. Met.*, **16**, 83–187.

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

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

Poli, P., J. Joiner, and E. R. Kurinski, 2002: 1DVAR analysis of temperature and humidity using GPS radio occulation refractivity data. *J. Geophys. Res*, **107(D20)**, 444–4468.

Sengupta, M., E. E. Clothiaux, and T. P. Ackerman, 2004: Climatology of warm boundary layer clouds at the ARM SGP site and their comparison to models. *J. Clim.*, **17**, 4760–4782.

Soden, B. J. and J. R. Lanzante, 1996: An assessment of satellite and radiosonde climatologies of upper-tropospheric water vapor. *J. Clim.*, **9**, 1235–1250.

Stephens, G. L. and Coauthors, 2002: The CLOUDSAT mission and the A-Train. *Bull. Amer. Meteor. Soc.*, **83**, 1771–1190.

Townsend, J. P., 2004: Resolution of large and small differences in gene expression using models for the Bayesian analysis of gene expression levels and spotted DNA microarrays. *BMC Bioinformatics*, **5:54**.