Mixed Lognormal-Gaussian Incremental Variational Data Assimilation

Steven J. Fletcher, Anton J. Kliewer, Andrew S. Jones and John M. Forsythe

Cooperative Institute for Research in the Atmosphere (CIRA)

Colorado State University (CSU)

Steven.fletcher@colostate.edu

Outline of Talk

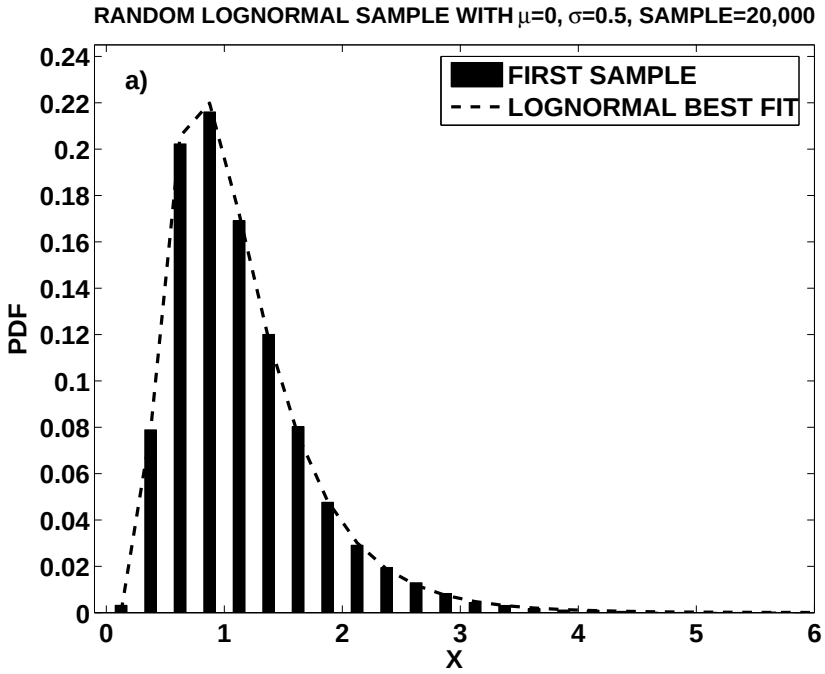
1) Mixed Lognormal-Gaussian Full Field VAR
2) Do we linearise the Bayesian problem or find the Bayesian problem for the increment?
3) Lognormal Incremental VAR - Geometric Tangent Linear Theory
4) Mixed Multiplicative-Additive Lognormal-Gaussian VAR
5) Examples with Lorenz 63 model
6) Comparison to a Gaussian incremental system
7) Conclusions

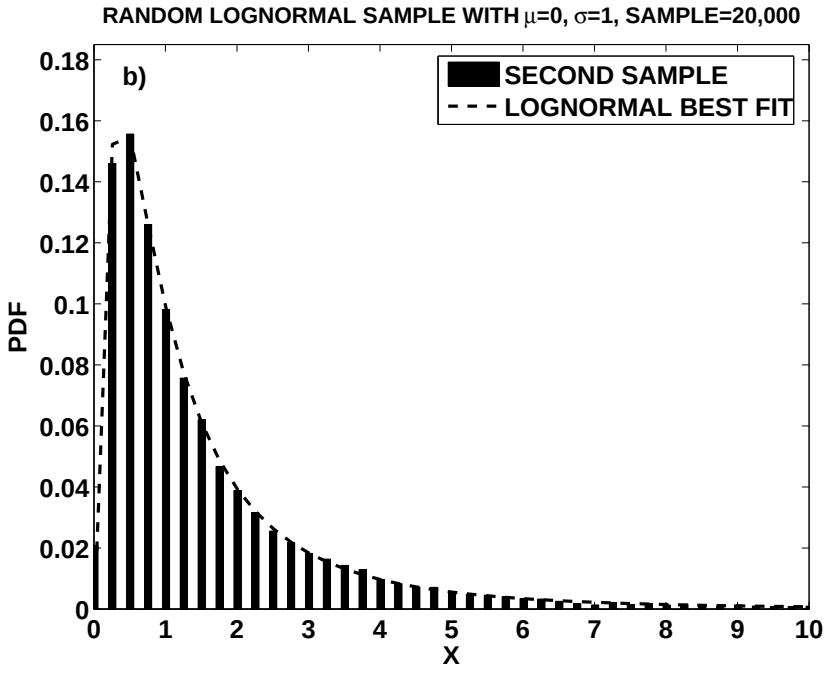
Lognormal Full Field Theory

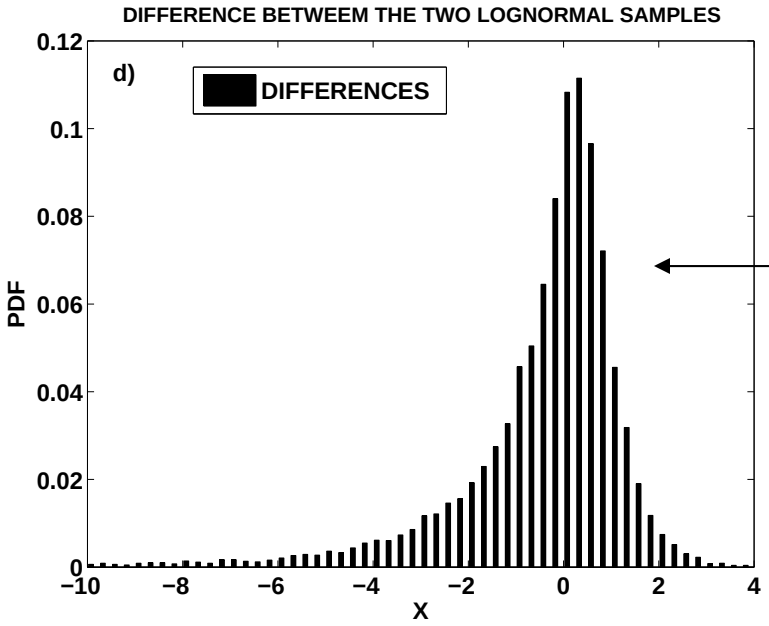
The starting point for lognormal based data assimilation is the definition of the errors. This was first proposed in Cohn (1997) where, due to the geometric behaviour of the lognormal distribution the errors could not be defined as the difference between two variables.

NOTE: There is no known distribution of the difference between two lognormally distributed independent random variables.
However, it is known that the distribution of the differences is NEITHER a Gaussian distribution nor a lognormal distribution.

Therefore we use the property that the ratio (and product) of two lognormal independent random variables is also a lognormally distributed random variable.







ASSUMED GAUSSIAN APPROACH

NOTE: DIFFERNCE IS NOT A GAUSSIAN

Fletcher (2010)

The lognormal distributed observational errors from Cohn (1997) are defined as

$$\varepsilon_{o,i} = \frac{y_i}{h_i(x)} \qquad (1)$$

where $i = 1, \ldots, N_o, y_i$ are the observations, and $h_i(x)$ is the non-linear observation operator.

The definition for lognormally distributed background errors comes from Fletcher and Zupanski (2007) and they are defined by

$$\varepsilon_{b,j} = \frac{x_j^t}{x_{b,j}} \qquad (2)$$

These are combined with the standard Gaussian error definition to form the mixed lognormal-Gaussian error definition (Fletcher and Zupanski, 2006b, 2007)

FULL FIELD MIXED GAUSSIAN-LOGNORMAL 3D AND 4D VAR

The resulting 3DVAR and 4DVAR cost functions are

$$J(x) = \frac{1}{2}\boldsymbol{\varepsilon}_b^T \mathbf{B}^{-1}\boldsymbol{\varepsilon}_b + \boldsymbol{\varepsilon}_b^T \begin{pmatrix} \mathbf{0}_{bp} \\ \mathbf{1}_{bq} \end{pmatrix} + \frac{1}{2}\boldsymbol{\varepsilon}_o^T \mathbf{R}^{-1}\boldsymbol{\varepsilon}_o + \boldsymbol{\varepsilon}_o^T \begin{pmatrix} \mathbf{0}_{op} \\ \mathbf{1}_{oq} \end{pmatrix} \qquad (3)$$

where

$$\boldsymbol{\varepsilon}_b = \begin{pmatrix} \boldsymbol{x}_p^t - \boldsymbol{x}_{bp} \\ \ln \boldsymbol{x}_q^t - \ln \boldsymbol{x}_{bq} \end{pmatrix} \textit{ and } \boldsymbol{\varepsilon}_o = \begin{pmatrix} \boldsymbol{y}_{op} - \boldsymbol{h}_{op}(\boldsymbol{x}) \\ \ln \boldsymbol{y}_{oq} - \ln \boldsymbol{h}_{oq}(\boldsymbol{x}) \end{pmatrix}$$

Fletcher and Zupanski (2007)

and

$$J(\boldsymbol{x}_0) = \frac{1}{2}\boldsymbol{\varepsilon}_{0b}^T \mathbf{B}^{-1}\boldsymbol{\varepsilon}_{0b} + \boldsymbol{\varepsilon}_{0b}^T \begin{pmatrix} \mathbf{0}_{bp} \\ \mathbf{1}_{bp} \end{pmatrix} + \frac{1}{2}\sum_{i=1}^{N_o} \boldsymbol{\varepsilon}_{0i}^T \mathbf{R}_i^{-1}\boldsymbol{\varepsilon}_{oi} + \sum_{i=1}^{N_o} \boldsymbol{\varepsilon}_{oi}^T \begin{pmatrix} \mathbf{0}_{opi} \\ \mathbf{1}_{oqi} \end{pmatrix} \qquad (4)$$

where

$$\boldsymbol{\varepsilon}_{0b} = \begin{pmatrix} \boldsymbol{x}_p^t(\boldsymbol{t}_0) - \boldsymbol{x}_{bp}(\boldsymbol{t}_0) \\ \ln \boldsymbol{x}_q^t(\boldsymbol{t}_0) - \ln \boldsymbol{x}_{bq}(\boldsymbol{t}_0) \end{pmatrix} \textit{ and } \boldsymbol{\varepsilon}_{oi} = \begin{pmatrix} \boldsymbol{y}_{pi} - \boldsymbol{h}_{pi}\big(\boldsymbol{M}_i(\boldsymbol{x}(\boldsymbol{t}_0))\big) \\ \ln \boldsymbol{y}_{qi} - \ln \boldsymbol{h}_{qi}\big(\boldsymbol{M}_i(\boldsymbol{x}(\boldsymbol{t}_0))\big) \end{pmatrix}$$

Fletcher (2010)

Do we linearise the Bayesian problem or find the Bayesian problem for the increment?

In Song *et al.* (2012) the first version of a lognormal distribution incremental VAR is presented. The starting point in Song *et al.* (2012) is to defined the relationship between the true state and the background as

$$x^t = x_b \circ e^{\Delta x} \qquad (5)$$

NOTE: Eqn (5) is equivalent to a log-linearisation of the full field cost function. However, this raises the point that for both side to be lognormal, the increment in (5) has to be Gaussian distributed.

Therefore, the associated cost function for the increment should be from a Gaussian distribution. Which is the median approach in Song *et al.* (2012) and shows positive results against an assumed Gaussian approach.

The problem here is that this increment is used to linearize the background component, which means that the Bayesian problem is not consistent for the Gaussian increment to make (5) lognormal.

Lognormal Incremental VAR

The starting point is to realize that unlike in the Gaussian framework we can not apply an additive increment if we want to maintain a lognormal framework, we have to define the relationship between the true state and the background as

$$x^t = x_b \circ \Delta \mathbf{x} \qquad (6)$$

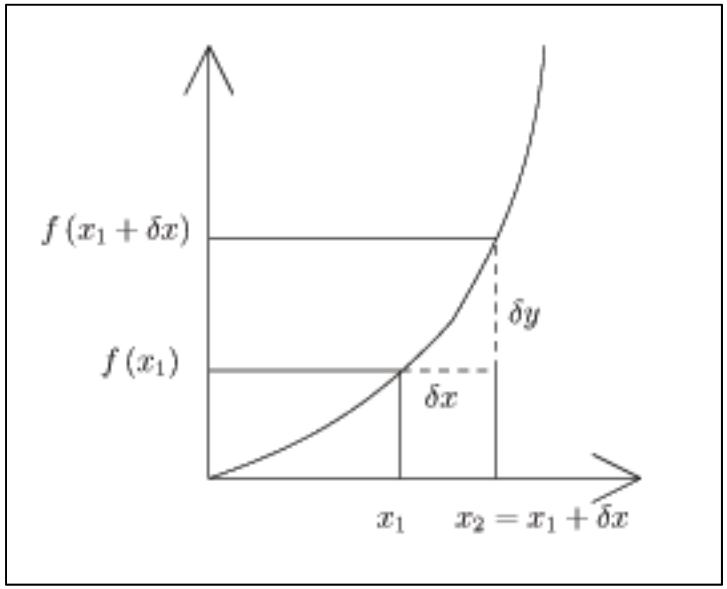
This then allows the background component of the cost function to be

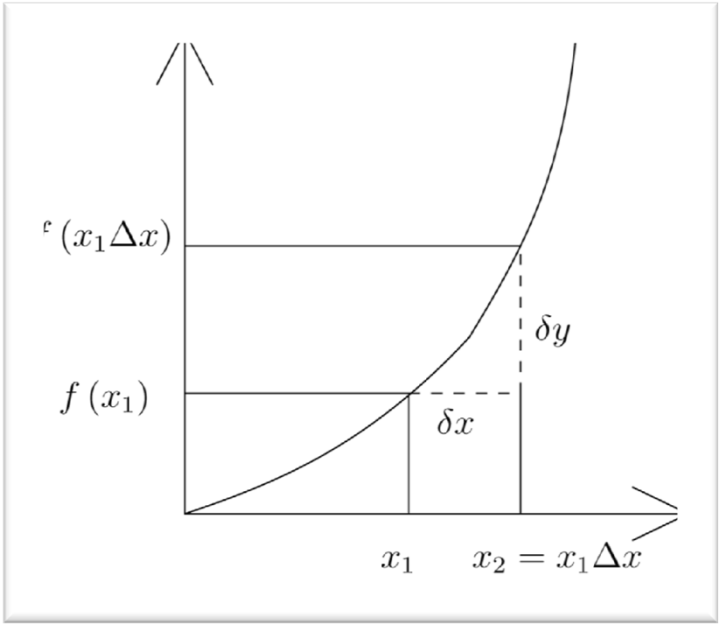
$$J_b(\Delta \mathbf{x}) = \frac{1}{2} \ln \Delta \mathbf{x}^T \boldsymbol{B}^{-1} \ln \Delta \mathbf{x} + \ln \Delta \mathbf{x}^{\mathrm{T}} \mathbf{1} \qquad (7)$$

Therefore, (7) is consistent with a Bayesian problem for a lognormally distributed increment, but is not a linearisation.

Geometric Tangent Linear Modeling

We have to address how we are to linearise $h(x_b \circ \Delta \mathbf{x})$. Again as with the background we require the observational component to be as consistent with a lognormal distribution as possible. Therefore, we can not simply use the standard tangent linear approximation. To overcome this problem, in Fletcher and Jones (2014), we define the Geometric Tangent Linear approximation as





Geometric Tangent Linear Modeling

This then allows us to be able to use standard derivative results for a multiplicative increment. Which means that for lognormal 3DVAR the observation operator can be approximated by

$$h(x_b \circ \Delta\mathbf{x}) \approx h(x_b) + \frac{\partial h}{\partial x} x_b(\Delta\mathbf{x} - 1) \qquad (7)$$

And for 4DVAR by

$$h_i(M_i(x_b(t_0) \circ \Delta\mathbf{x}(t_0))) \approx h(M_i(x_b(t_0))) + \frac{\partial h_i}{\partial x}\frac{\partial M_i}{\partial x} x_b(t_0)(\Delta\mathbf{x}(t_0) - 1) \qquad (8)$$

Fletcher and Jones (2014).

Mixed Multiplicative-Additive Lognormal-Gaussian Incremental VAR

As we do not live in a just Gaussian or lognormal world, we have to combine the two approaches, as we have done for the full field (Fletcher, 2010). We therefore define our incremental vector as

$$\mathbf{\Delta x_{mx}} = \begin{pmatrix} \delta x_{bp}(t_0) \\ \mathbf{\Delta x_{bq}}(t_0) \end{pmatrix} \qquad (9)$$

This then gives the following 4DVAR cost function (Fletcher and Jones, 2014).

$$J(\Delta_{mx}) = \frac{1}{2}\Delta x_{mx}^{T} B\, \Delta x_{mx}^{T} + \Delta x_{mx}^{T} \begin{pmatrix} 0_{bp} \\ 1_{bq} \end{pmatrix}$$

$$+\frac{1}{2}\sum_{i=1}^{N_o} \begin{pmatrix} y_{opi} - h_{opi}(M_i(x_b(t_0))) - H_{opi}\overline{M_i}\Delta x_{mx} \\ \ln y_{oqi} - \ln h_{oqi}(M_i(x_b(t_0))) - W_{oi}^{-1} H_{oqi}\overline{M_i}\Delta x_{mx} \end{pmatrix}^{T} R_i^{-1}$$

$$\times \begin{pmatrix} y_{opi} - h_{opi}(M_i(x_b(t_0))) - H_{opi}\overline{M_i}\Delta x_{mx} \\ \ln y_{oqi} - \ln h_{oqi}(M_i(x_b(t_0))) - W_{oi}^{-1} H_{oqi}\overline{M_i}\Delta x_{mx} \end{pmatrix}$$

$$+ \begin{pmatrix} y_{opi} - h_{opi}(M_i(x_b(t_0))) - H_{opi}\overline{M_i}\Delta x_{mx} \\ \ln y_{oqi} - \ln h_{oqi}(M_i(x_b(t_0))) - W_{oi}^{-1} H_{oqi}\overline{M_i}\Delta x_{mx} \end{pmatrix}^{T} \begin{pmatrix} 0_{opi} \\ 1_{oqi} \end{pmatrix}$$

Example with the Lorenz 63 model

The Lorenz model is given by the following non-linear system of three ordinary differential equation

$$\dot{x} = \alpha(y - x)$$
$$\dot{y} = \rho x - y - xz$$
$$\dot{z} = xy - \beta z$$

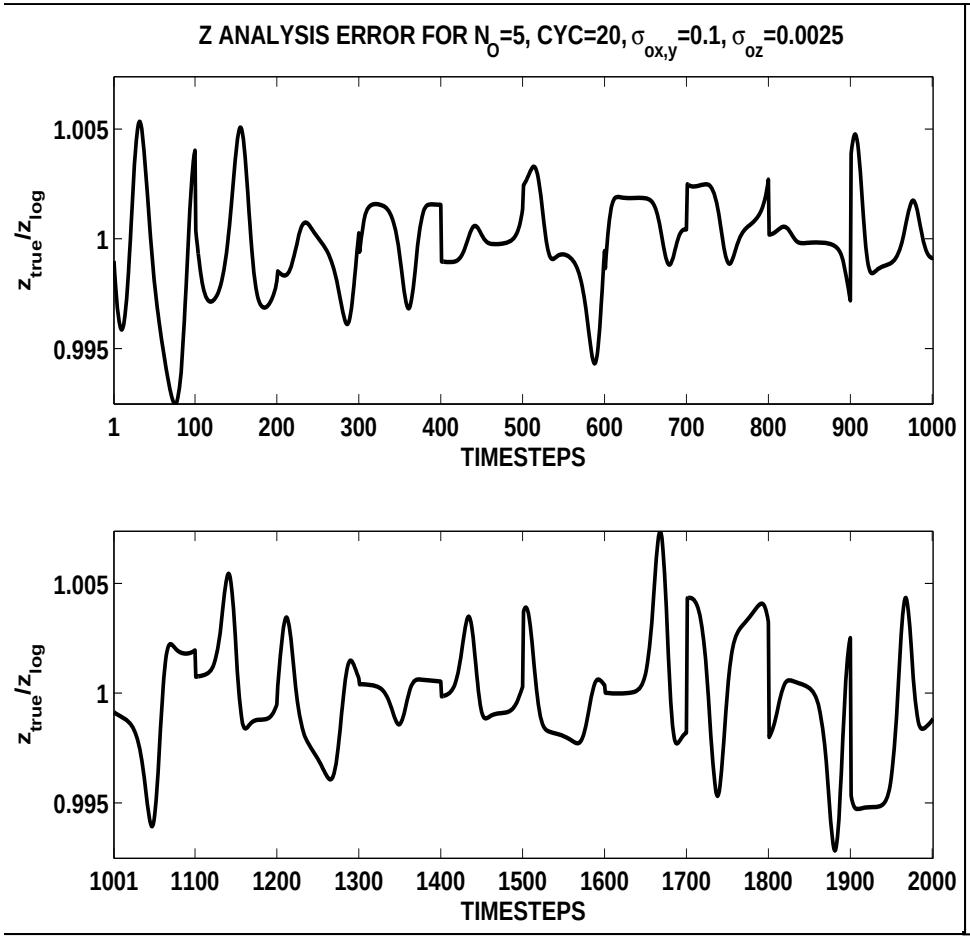
The system is linearized and then descretized using the modified Euler scheme. The adjoint of this scheme is calculated analytically.

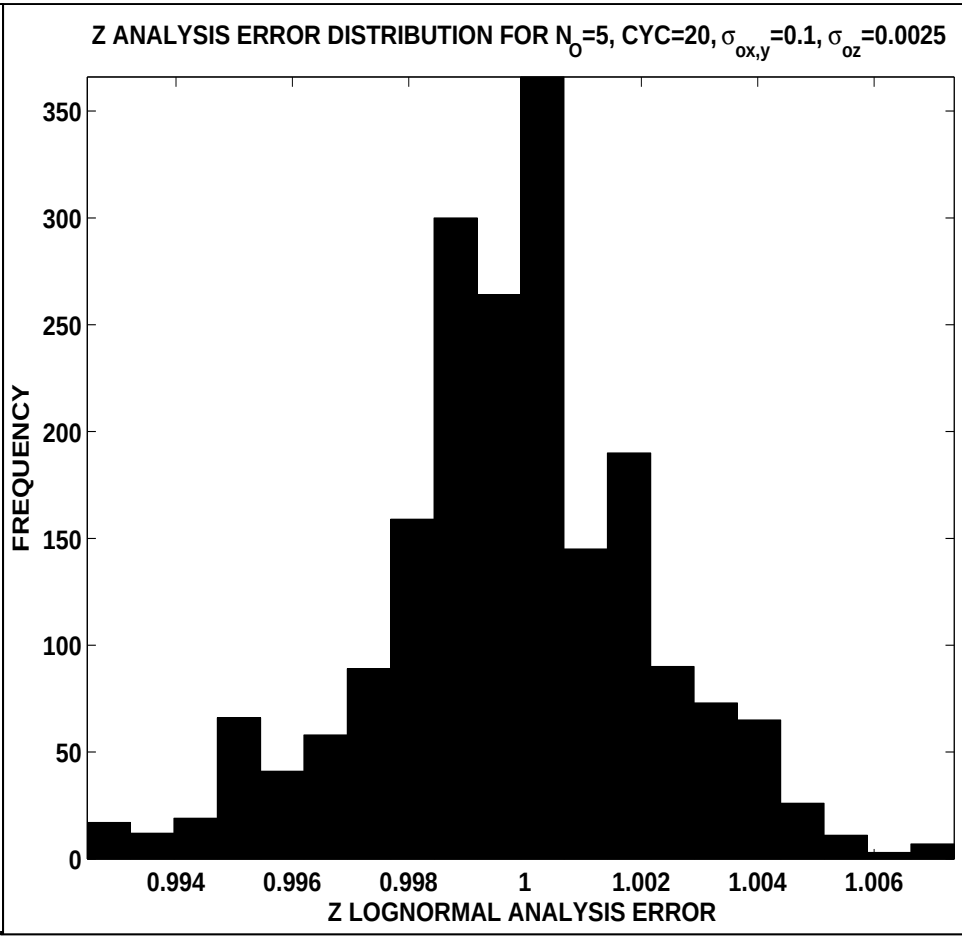
The observations are calculated by adding, or multiplying random perturbations from a Gaussian distribution, for the lognormal the background state is multiplied by the exponential of the scaled Gaussian increment

The initial conditions for the true solution, the initial background and the true increment at the initial time are

STATE	TRUE	BACKGROUND	INCREMENT
x	-5.4458	-5.9	0.4542
y	-5.4841	-5.0	-.4841
z	22.5606	24.0	0.94

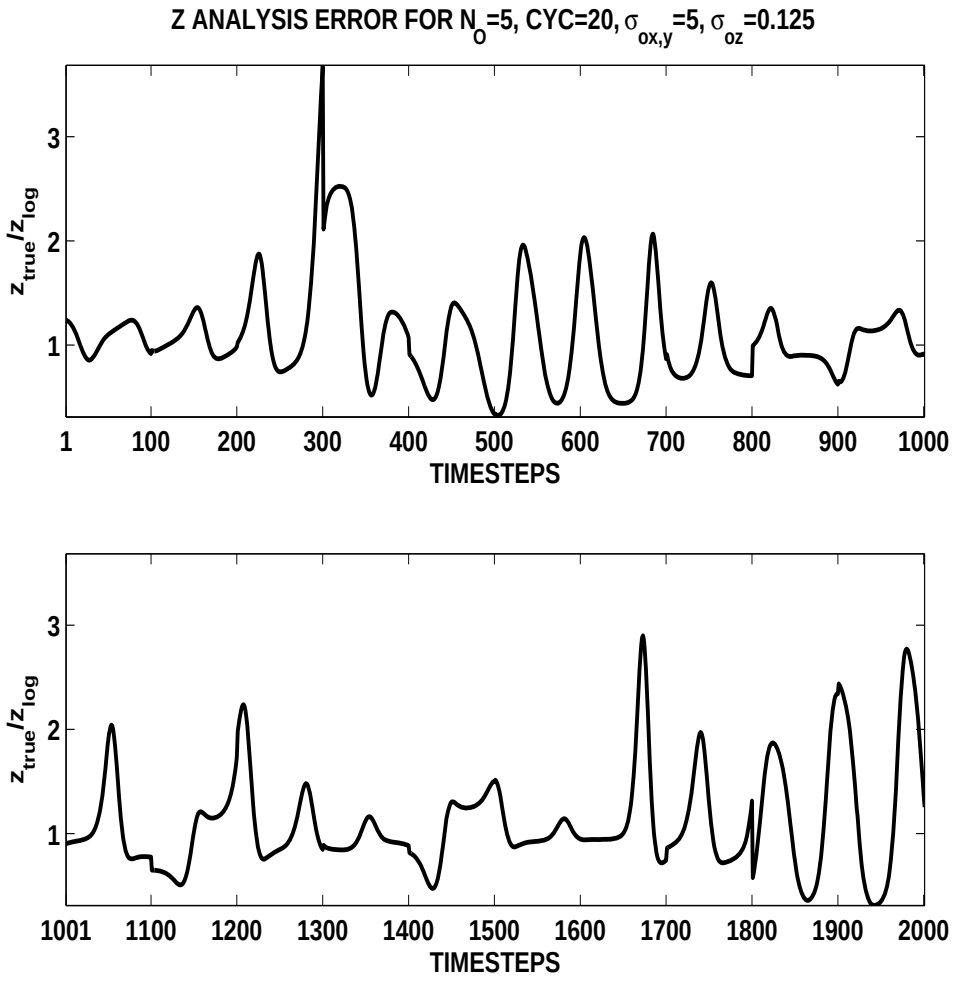
Results for 20 cycles of 100ts with few accurate obs

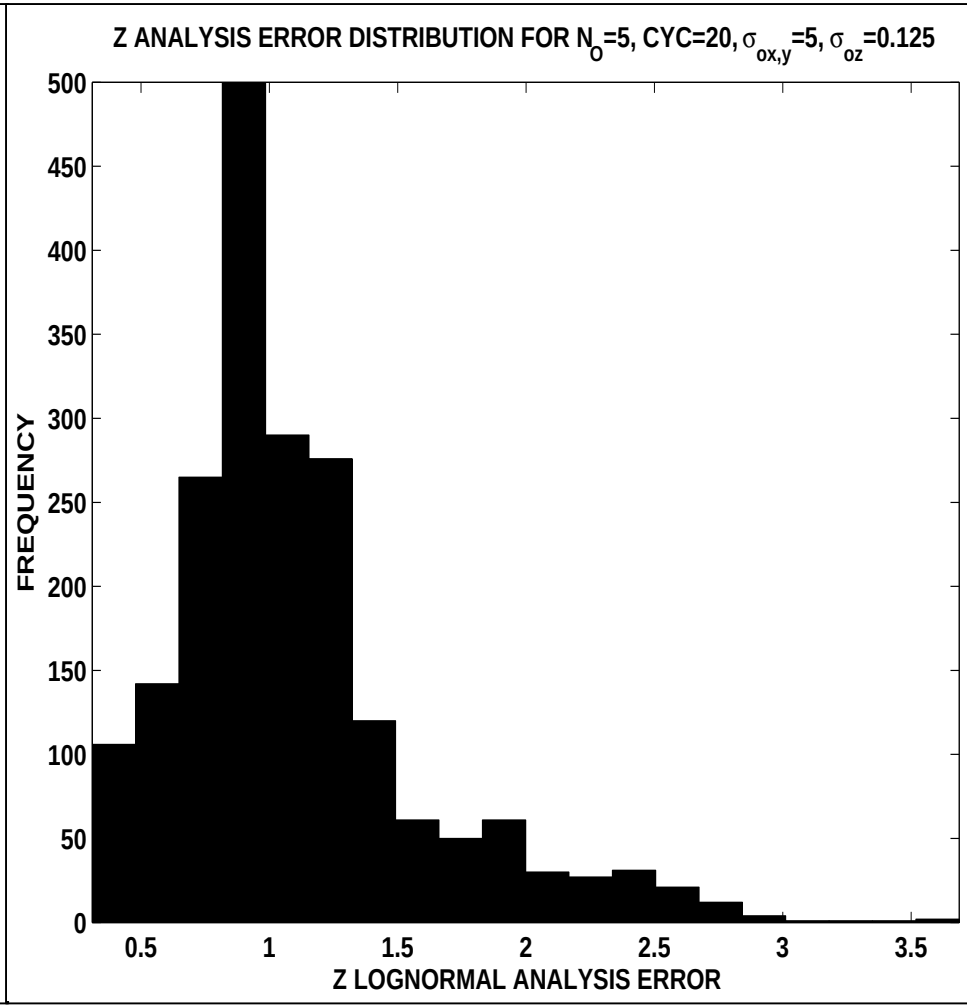




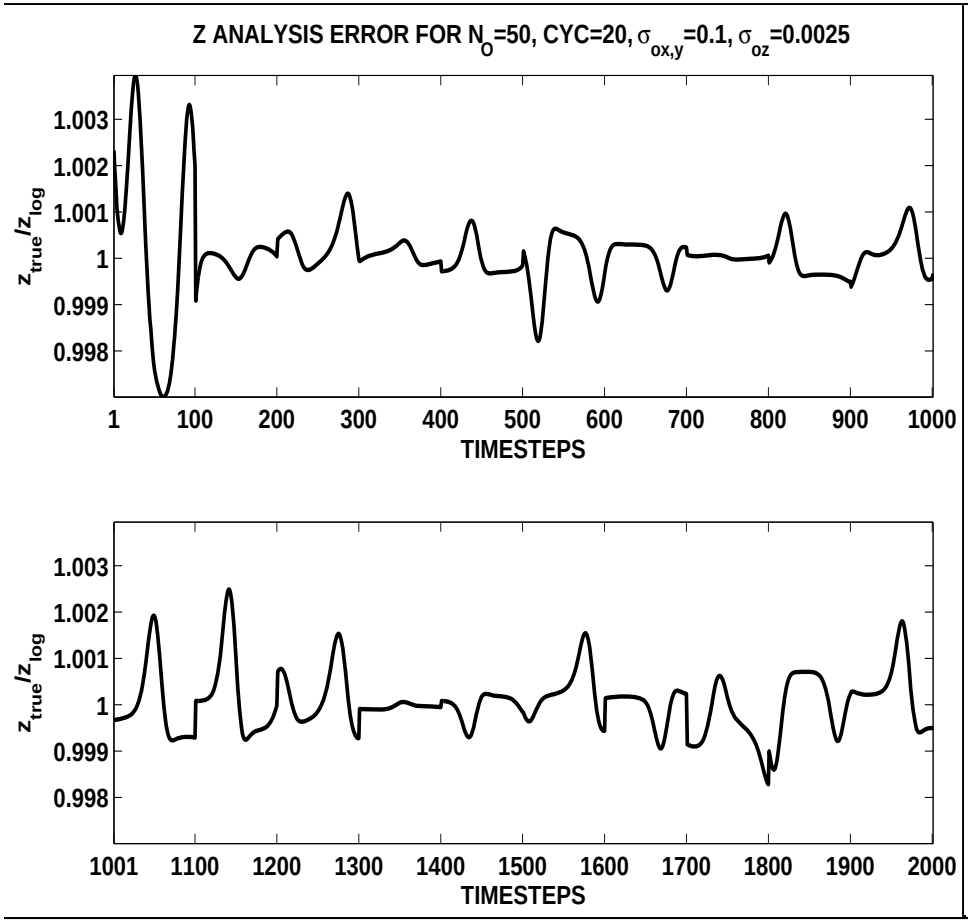
Fletcher and Jones (2014)

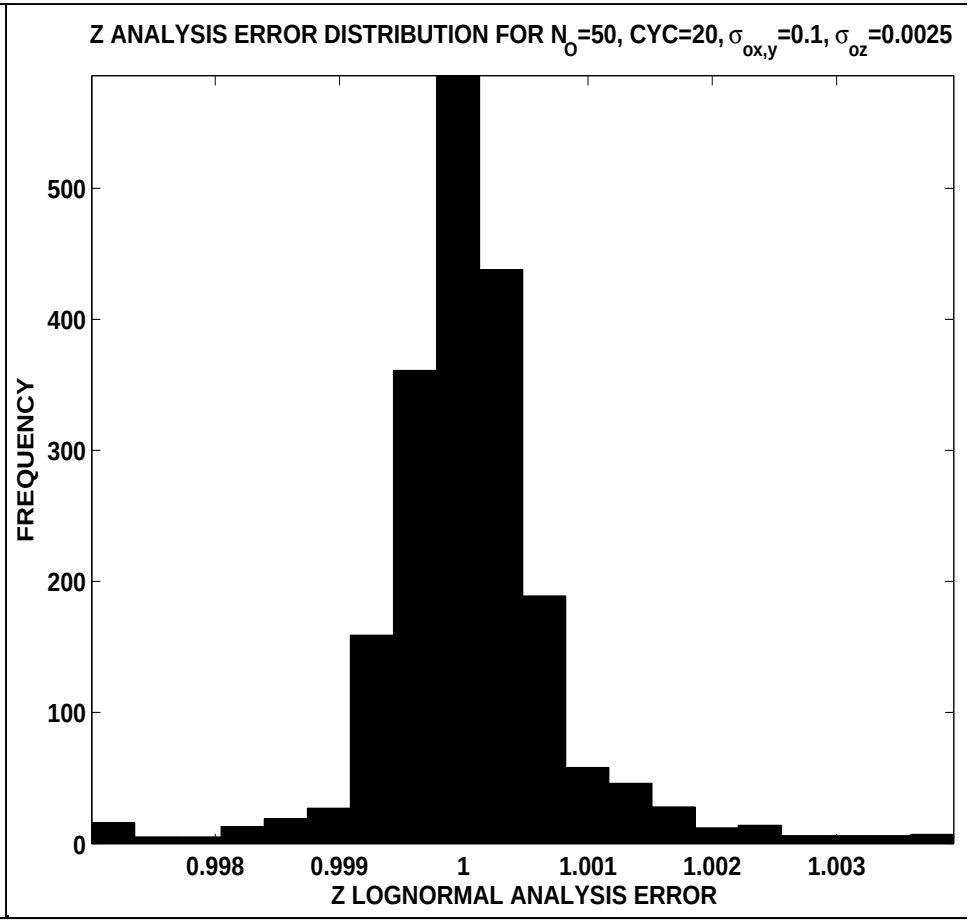
Results with same window lengths and same number of obs but less accurate



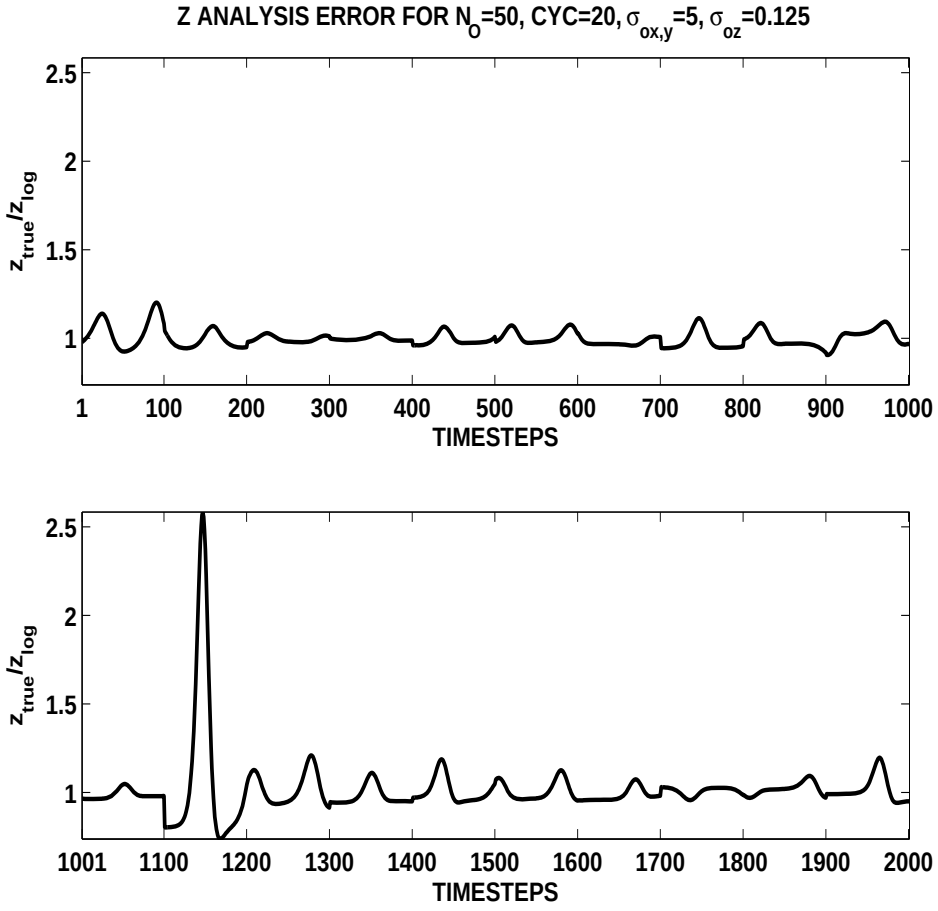


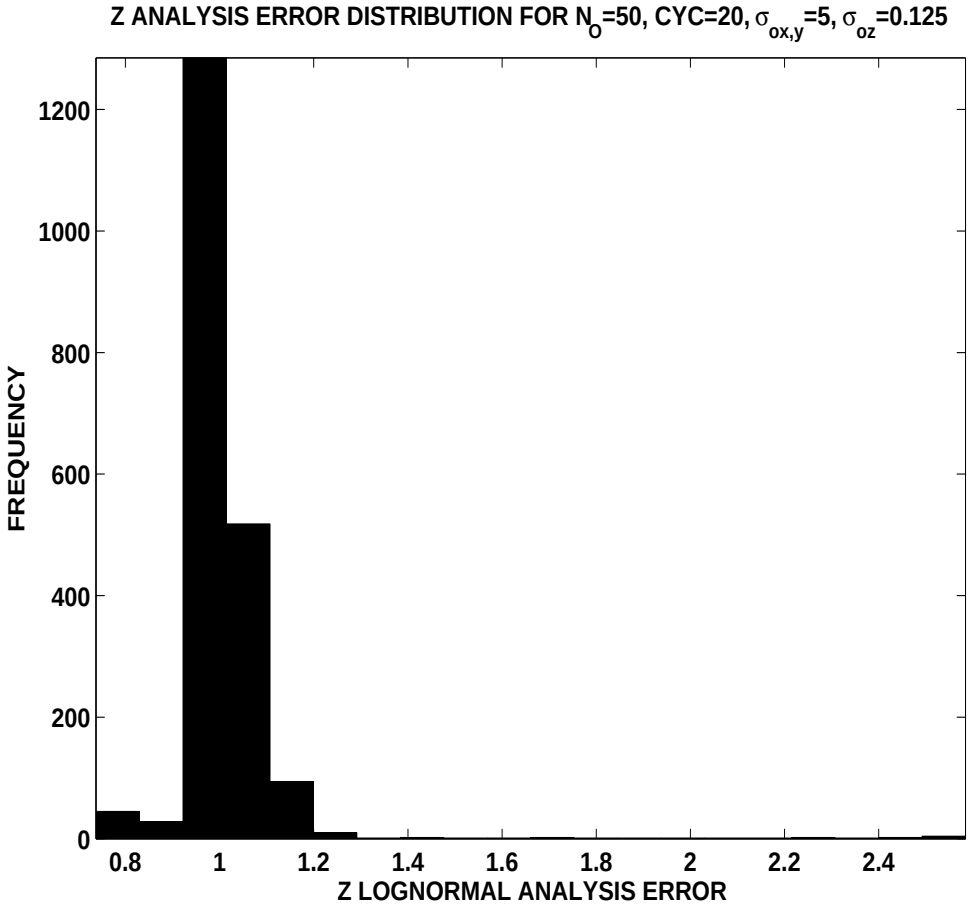
Results for same number of assimilation windows but with accurate observations every other time step





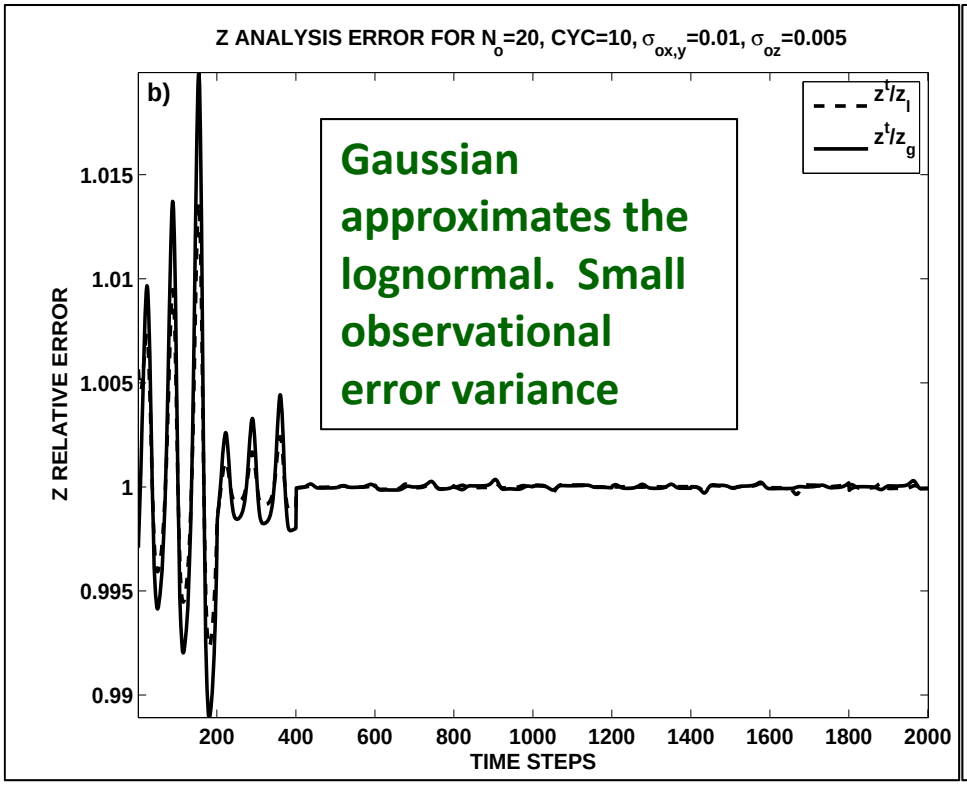
Results for same number of assimilation windows but with accurate observations every other time step

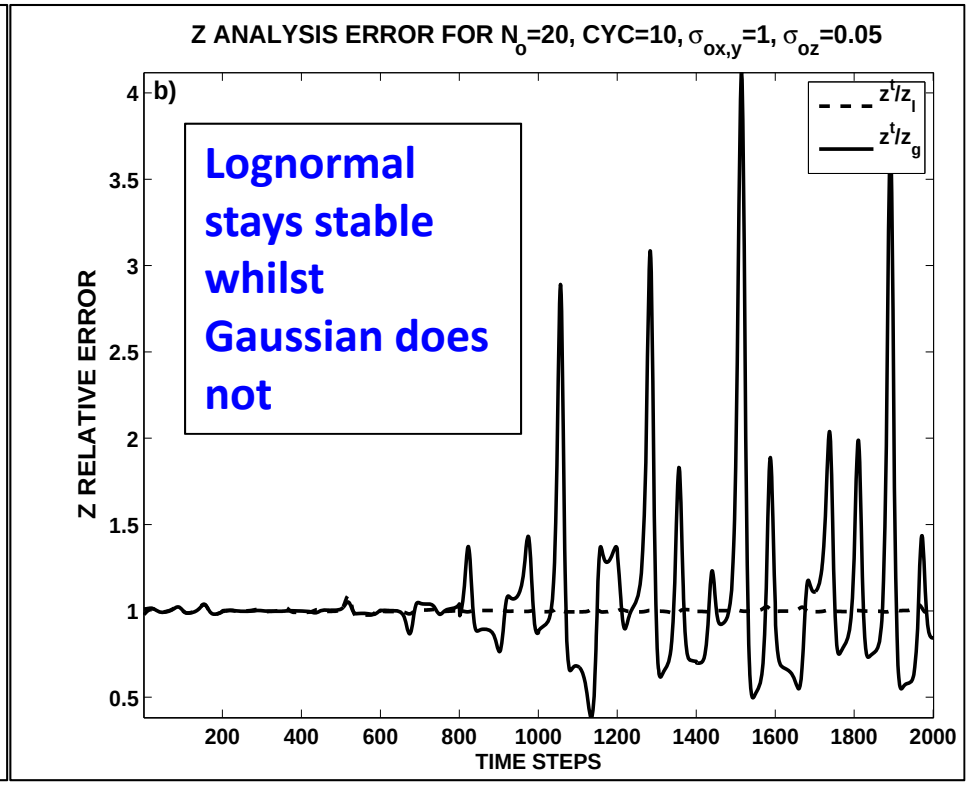




Comparison to a full Gaussian incremental system

These results are from Fletcher and Jones (2014) paper where here we are presenting results from two of the experiments, the first to highlight where the two systems are similar and the case where the lognormal converges but the Gaussian does not.





Conclusions

- It is possible to define an incremental version of the lognormal full field 3D and 4DVAR.
- This is possible through relating a multiplicative increment to an additive one from the gradient definition.
- Have been able to test in a Lorenz 63 model with just inner loops and no QC of observations and monitoring of the tangent linear assumption
- Tested with different number of observations with different variances and with different assimilation window lengths
- Next is to apply this theory in the WRF-GSI system which does not need the adjoint of the WRF model
- Have applied the full field mixed distribution formulation in a temperature-mixing ratio microwave retrieval system and have shown positive results compared to a Gaussian only version (Kliewer *et al.,* 2015).