Non-Gaussian Based Data Assimilation Lecture 2: Applications

Steven J. Fletcher
Cooperative Institute for Research in the Atmosphere
Colorado State University

Overview of Lecture

1. Do we linearize the Bayesian problem or do we find the Bayesian Problem for the linear increment?
2. Geometric Tangent Linear Approximation
3. Mixed Multiplicative-Additive Incremental 4DVAR
4. Lognormal Detection Algorithm
5. Mixed Lognormal-Gaussian Mixing Ratio – Temperature retrievals
6. Optimal Regions for Descriptive Statistics for Lognormal Based DA
7. Chaotic Signal with Newton-Raphson solver and Lognormal DA

Do we linearize the Bayesian problem or do we find the Bayesian Problem for the linear increment?

The first version of a incremental approach for a lognormal based variational data assimilation method appears in Song *et al.,* (2012). Their starting point was to define an increment as

$$\boldsymbol{x}^{t} = \boldsymbol{x}_{b} \circ \boldsymbol{e}^{\boldsymbol{\Delta x}}$$

Where the circle represents the Hadamard multiplication operator which is a component wise vector multiplication. This increment is then substituted into the background component of the lognormal cost function which results in

$$J_b(\boldsymbol{x}) = \frac{\mathbf{1}}{\mathbf{2}}(\boldsymbol{\Delta x})^{\boldsymbol{T}}\boldsymbol{B}_{\boldsymbol{l}}^{-\mathbf{1}}(\boldsymbol{\Delta x}) + \langle \boldsymbol{\Delta x}, \mathbf{1} \rangle$$

Which has **linearized** the cost function by removing the logarithms but this is not consistent with keeping the probability distributions the same on either side of the equals side as dictated by the uniqueness theorem for moment generating functions. If we assume that the true state is a lognormally distributed random variable (RV), and that the background is also a lognormally distributed RV then we require the exponential term to also be a lognormal RV. This can only occur if $\boldsymbol{\Delta x}$ is a Gaussian RV. Therefore, for us to be able to optimize the increments we need to be solving a Gaussian cost function for it.

In Song *et al.,* (2012) they present results from both the linearized version which did not work well, which is to be expected as the Bayesian problem is not consistent, but when they used what they refer to as the median approach, which is the equation above with out the inner product, they were able to show better results than assuming a Gaussian based approach.

The motivation in Song *et al.,* (2012) for a lognormal approach was to avoid negative values for positive definite variables, that the Gaussian approach kept giving.

Song, H. C., A. Edwards, A. M. Moore and J. Fiechter, 2012: Incremental four-dimensional variational data assimilation of positive-definite oceanic variables using a logarithmic transformation. *Ocean Modell.,* **54-55,** 1—17.

Do we linearize the Bayesian problem or do we find the Bayesian Problem for the linear increment?

In Fletcher and Jones (2014) and alternative incremental formulation was proposed. The alternative formulation focused upon keeping the problem lognormal as much as possible. Therefore, this implied that the formulation linking the true state to the background state needed to be lognormal and not be such that it linearized the cost function. The definition that keeps the problem lognormally consistent is

$$\boldsymbol{x}^{\boldsymbol{t}} = \boldsymbol{x}_{\boldsymbol{b}} \circ \boldsymbol{\Delta x}$$

This now implies that for both sides of the equal sign to be lognormal, if the true state is assumed to be a lognormal RV and so is the background, $\boldsymbol{\Delta x}$ must also be a lognormal RV. Substituting this expression into the full field background component of the cost function results in

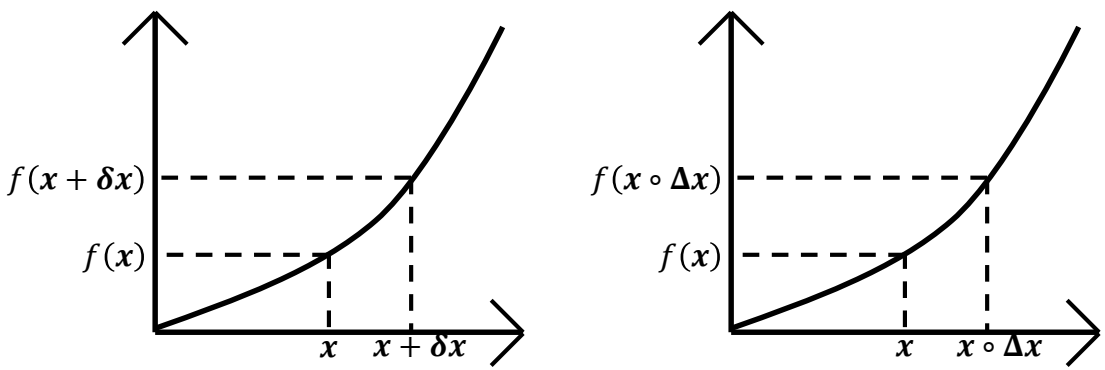
$$J_b(\boldsymbol{x}) = \frac{1}{2}(\ln \boldsymbol{\Delta x})^T \boldsymbol{B}_{\boldsymbol{l}}^{-\mathbf{1}}(\ln \boldsymbol{\Delta x}) + \langle \ln \boldsymbol{\Delta x}\,, \mathbf{1} \rangle$$

Which is a lognormal cost function for the mode of a lognormal distribution.

Fletcher, S.J. and A. S. Jones, 2014: Multiplicative and additive incremental variational data assimilation for mixed lognormal-Gaussian errors. *Mon. Wea. Rev.*, **142,** 2521—2544.

Geometric Tangent Linear Approximation

The question now is how to linearize the observational component with respect to a multiplicative increment? That is to say we need to linearize $\boldsymbol{h}(\boldsymbol{x_b} \circ \mathbf{\Delta x})$. In Fletcher and Jones (2014) it is shown that we can use the same theory behind a tangent linear approximation with an additive increment but now using an multiplicative increment. If we consider the diagrams below



The plot on the left is to illustrate how we obtain the additive tangent linear approximation, whilst the figure on the right is to illustrate how we can obtain an expression involving multiplicative increments.

Geometric Tangent Linear Approximation

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

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

And for 4DVAR by

$$\boldsymbol{h_i\left(M_i\left(x_b(t_0) \circ \Delta \mathrm{x}(t_0)\right)\right) \approx h\left(M_i\left(x_b(t_0)\right)\right) + \frac{\partial h_i}{\partial x}\frac{\partial M_i}{\partial x} x_b(t_0)(\Delta \mathrm{x}(t_0) - 1)}$$

Mixed Multiplicative-Additive Incremental 4DVAR

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

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

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

Mixed Multiplicative-Additive Incremental 4DVAR

$$
\begin{aligned}
J(\Delta_{mx}) &= \frac{1}{2}\Delta x_{mx}^{T} B\, \Delta x_{mx}^{T} + \Delta x_{mx}^{T}\begin{pmatrix} 0_{bp} \\ 1_{bq} \end{pmatrix} \\
&+ \frac{1}{2}\sum_{i=1}^{N_o}\begin{pmatrix} y_{opi} - h_{opi}(M_i(x_b(t_0))) - H_{opi}\overline{M_i}\Delta x_{mx} \\ \ln y_{oqi} - \ln h_{oqi}(M_i(x_b(t_0))) - W_{oi}^{-1}H_{oqi}\overline{M_i}\Delta x_{mx} \end{pmatrix}^{T} R_i^{-1} \\
&\times \begin{pmatrix} y_{opi} - h_{opi}(M_i(x_b(t_0))) - H_{opi}\overline{M_i}\Delta x_{mx} \\ \ln y_{oqi} - \ln h_{oqi}(M_i(x_b(t_0))) - W_{oi}^{-1}H_{oqi}\overline{M_i}\Delta x_{mx} \end{pmatrix} \\
&+ \begin{pmatrix} y_{opi} - h_{opi}(M_i(x_b(t_0))) - H_{opi}\overline{M_i}\Delta x_{mx} \\ \ln y_{oqi} - \ln h_{oqi}(M_i(x_b(t_0))) - W_{oi}^{-1}H_{oqi}\overline{M_i}\Delta x_{mx} \end{pmatrix}^{T}\begin{pmatrix} 0_{opi} \\ 1_{oqi} \end{pmatrix}
\end{aligned}
$$

Mixed Multiplicative-Additive Incremental 4DVAR

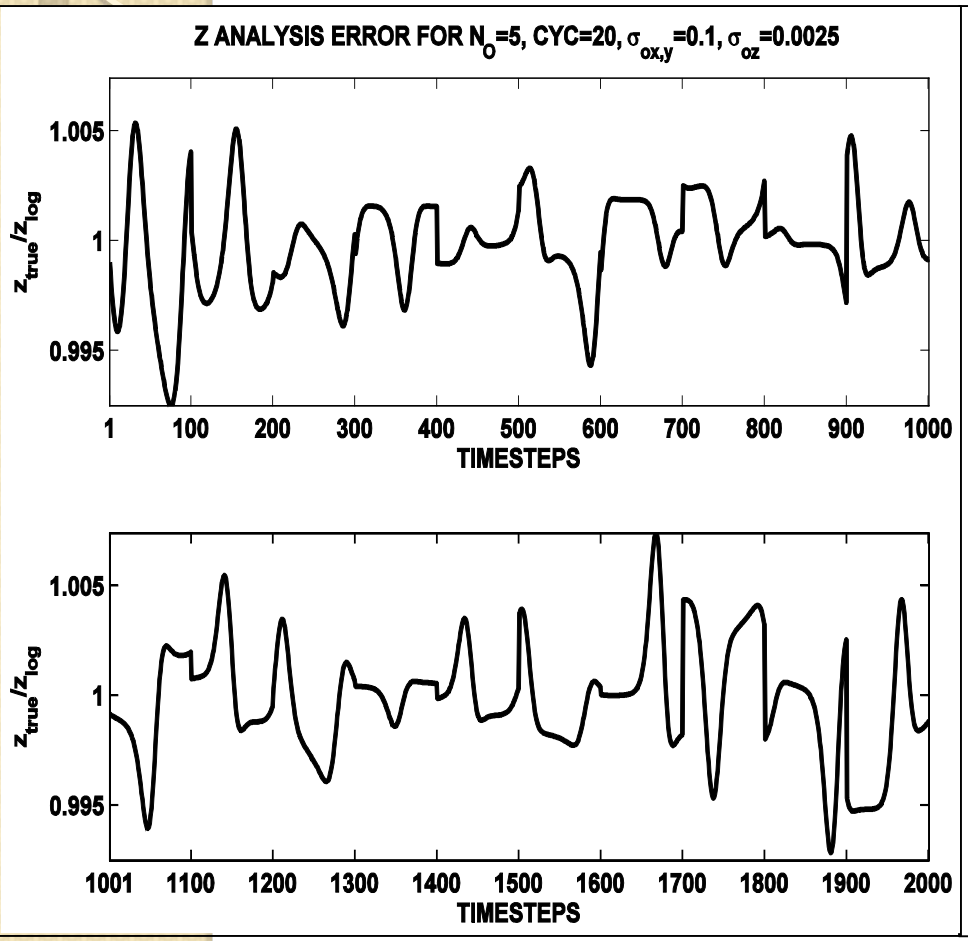
The Lorenz model is given by the following non-linear system of three ordinary differential equation

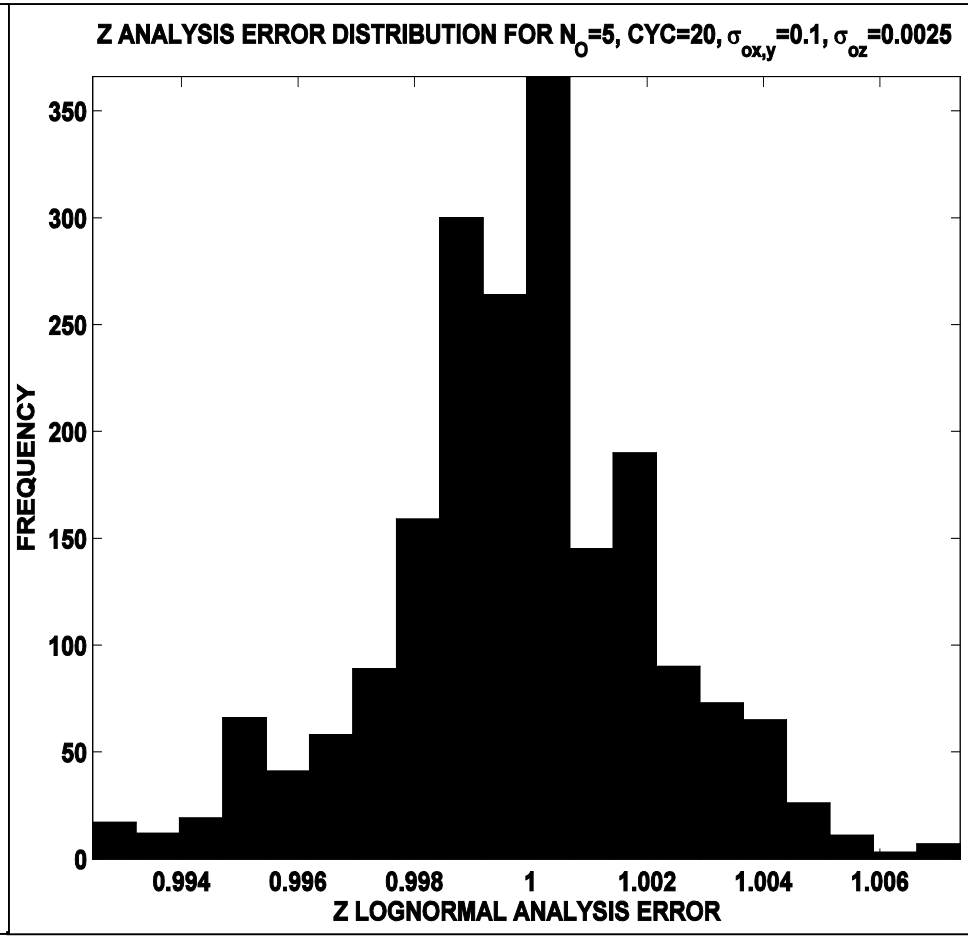
$$\boxed{\begin{aligned} \dot{x} &= \alpha(y - x) \\ \dot{y} &= \rho x - y - xz \\ \dot{z} &= xy - \beta z \end{aligned}}$$

The system is linearized and then discretized using the modified Euler scheme. The adjoint of this scheme is calculated analytically. The minimization of the cost function is achieved through the ***fminsearch*** routine in MATLAB which uses a Nelder-Mead algorithm.

The observations are calculated by adding random perturbations from a Gaussian distribution, (x and y components) and multiplying perturbations from a lognormal distribution (z component) to the true model run.

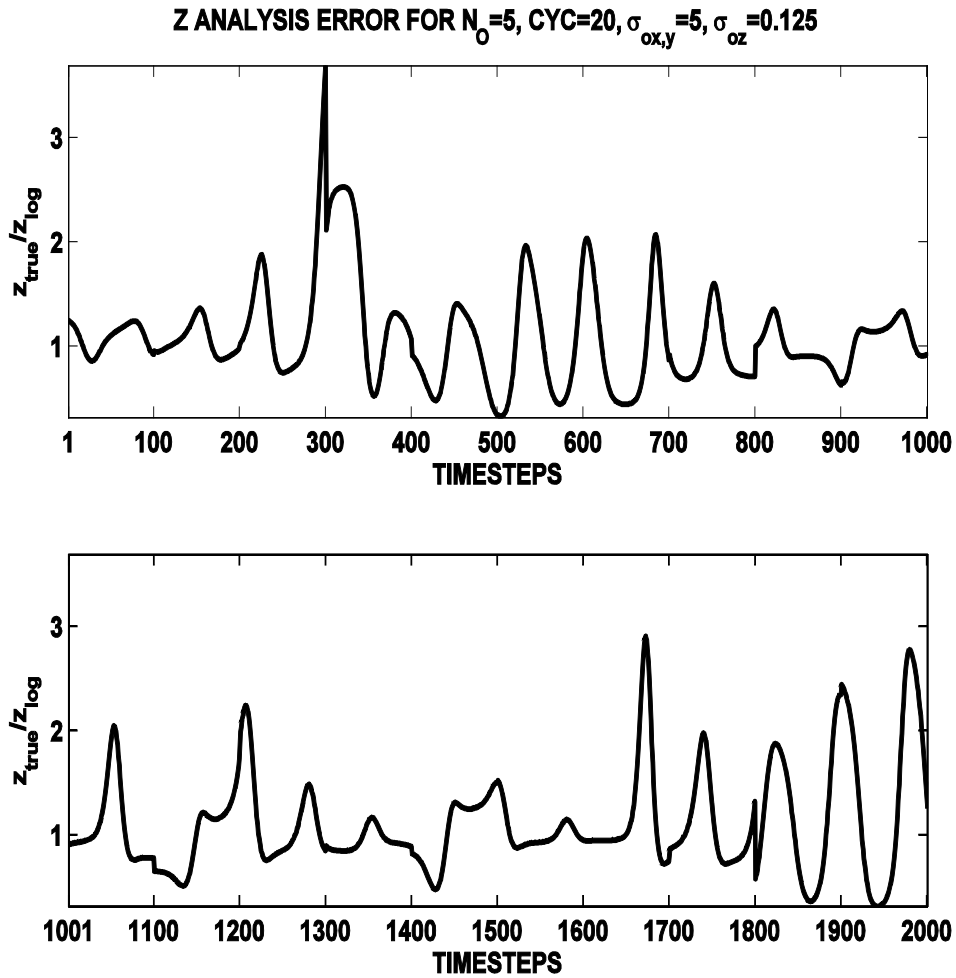
Mixed Multiplicative-Additive Incremental 4DVAR

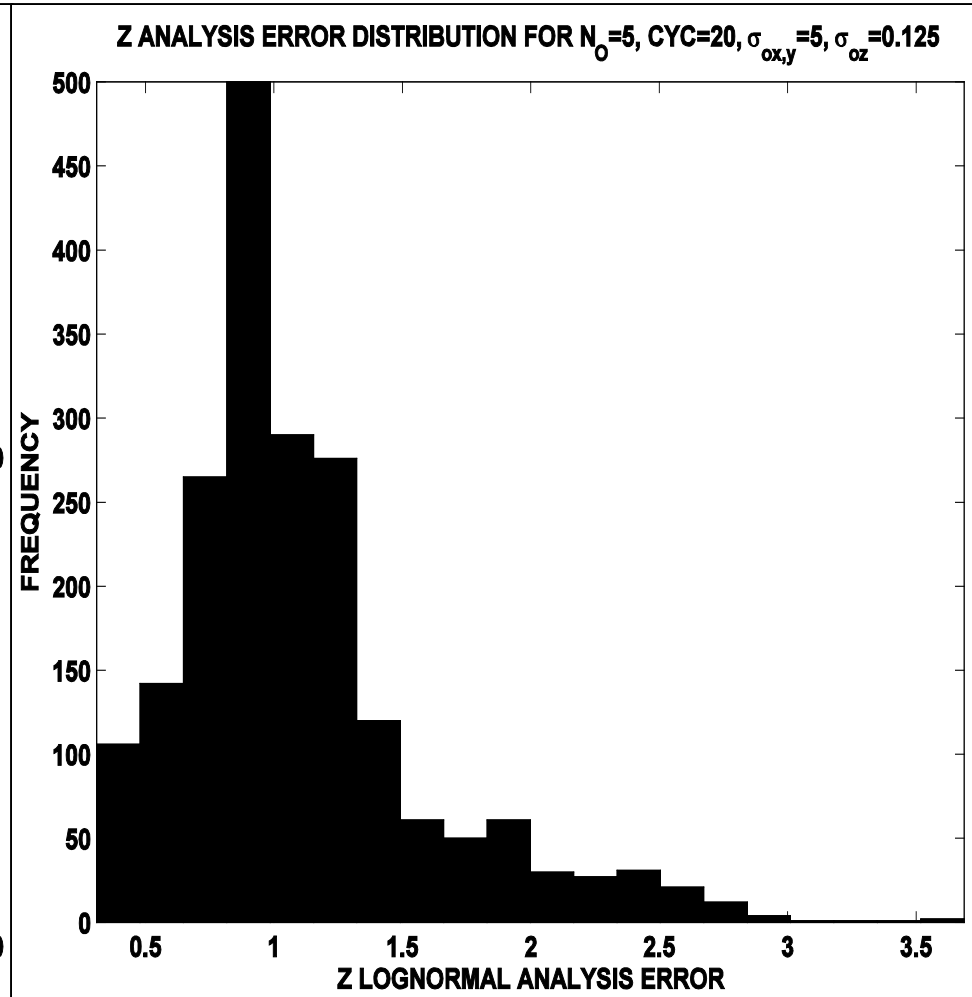




Results for 20 cycles of 100ts with few accurate obs

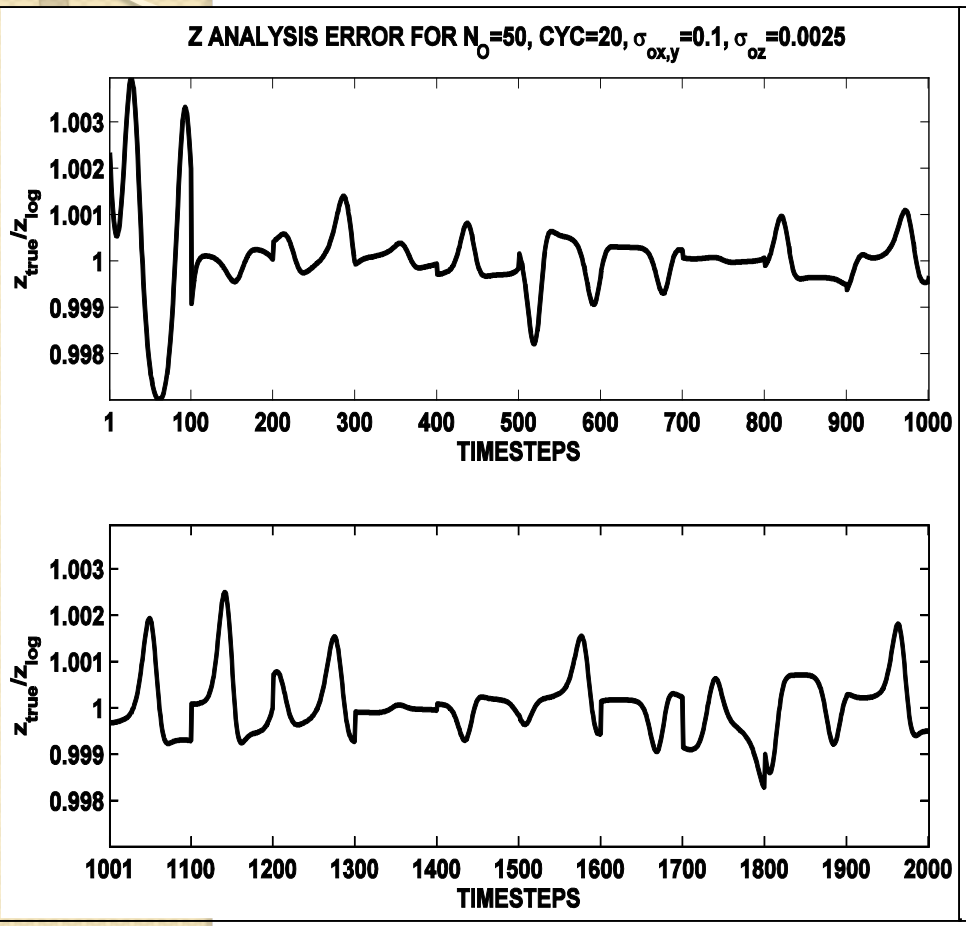
Mixed Multiplicative-Additive Incremental 4DVAR

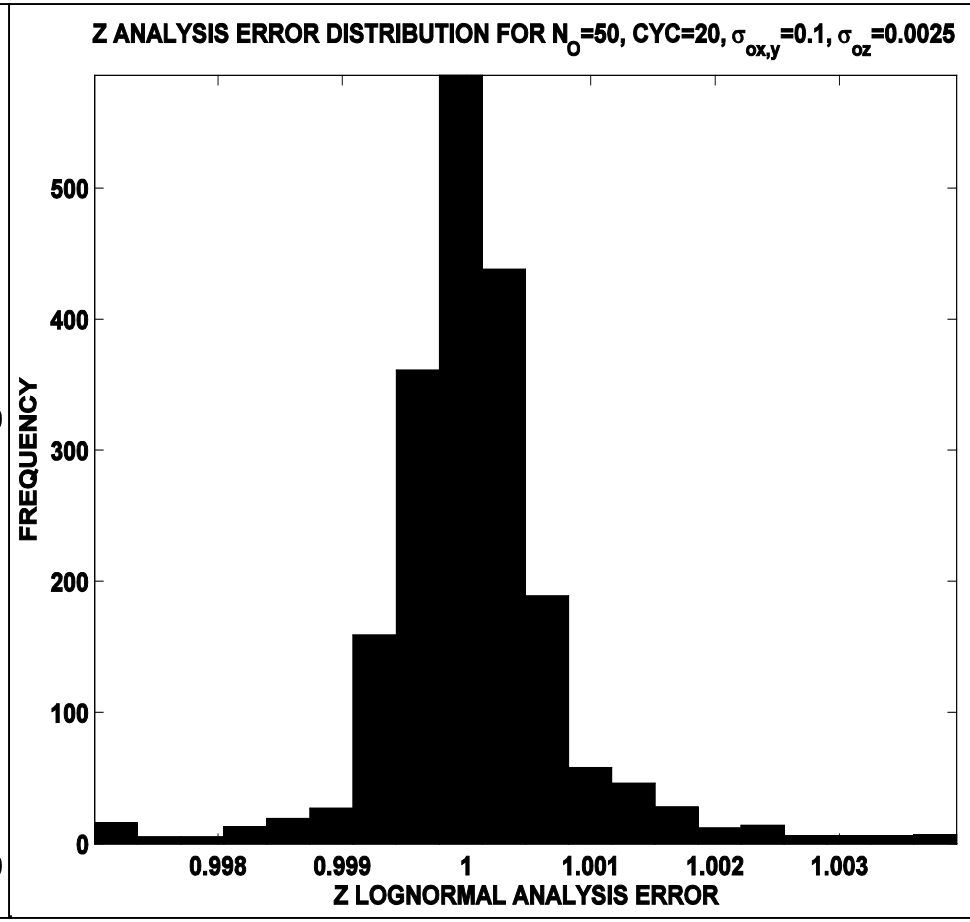




Results with same window lengths and same number of obs but less accurate

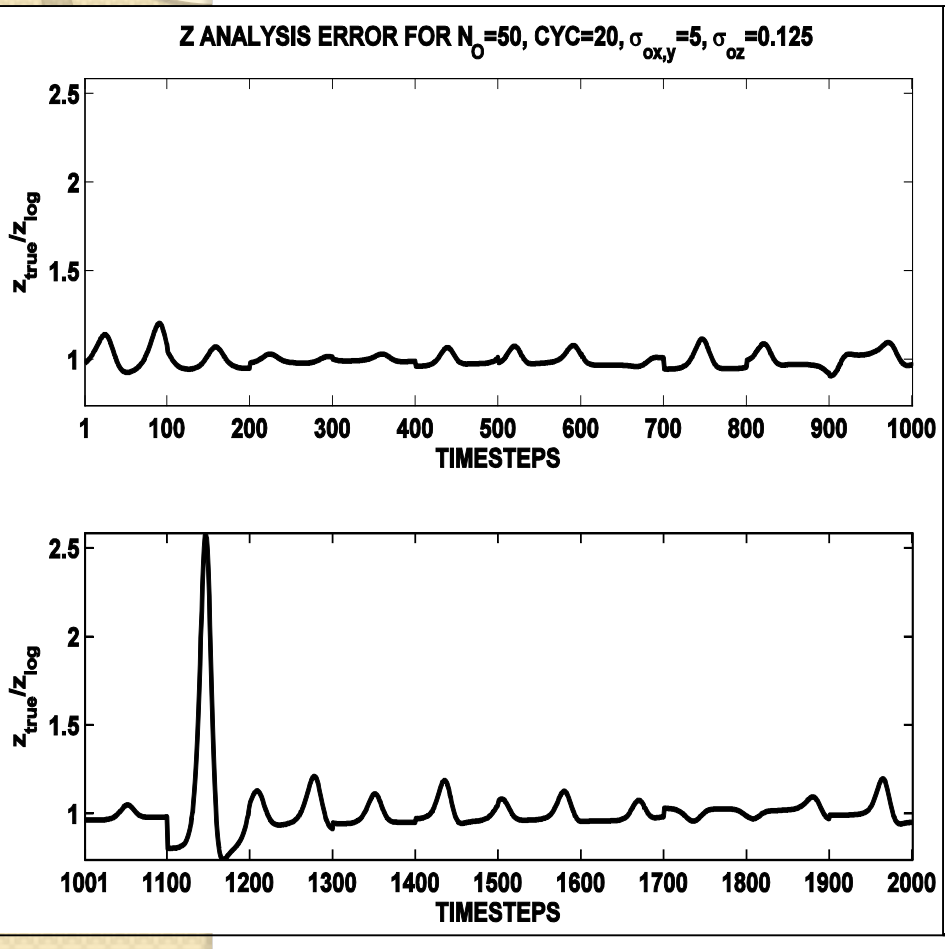
Mixed Multiplicative-Additive Incremental 4DVAR

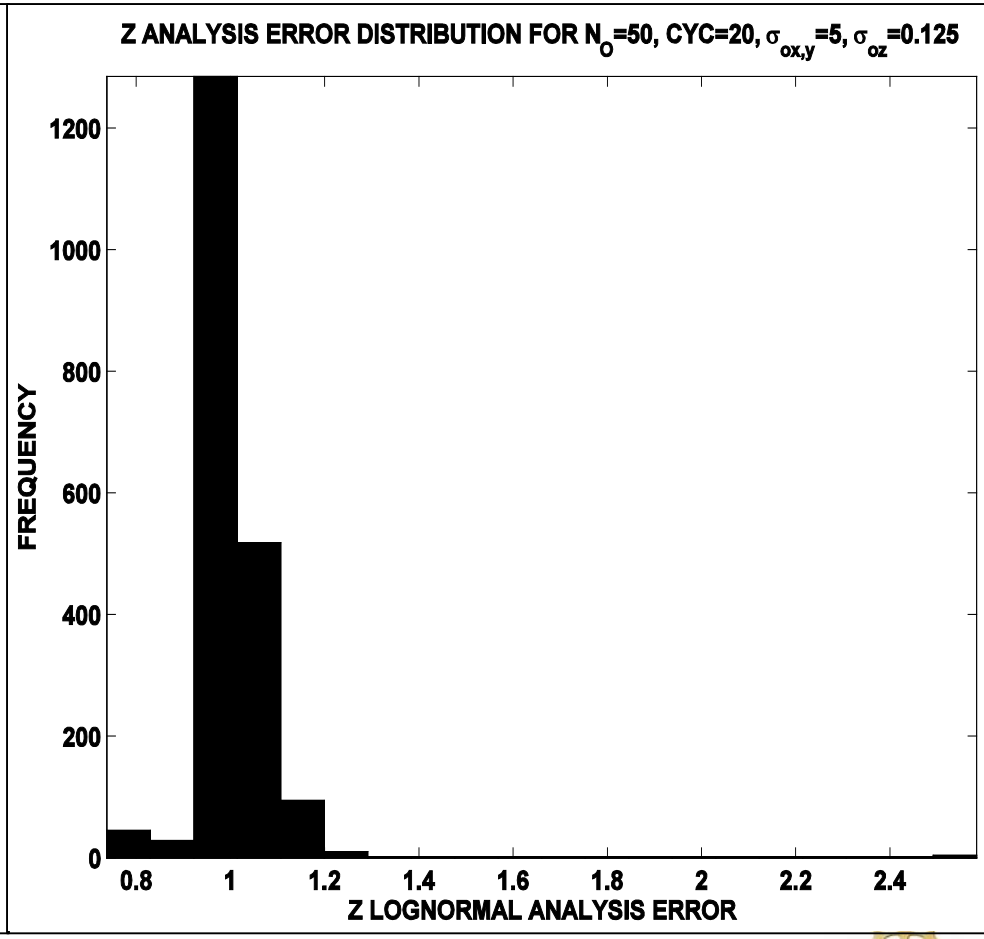




Results for same number of assimilation windows but with accurate observations every other time step

Mixed Multiplicative-Additive Incremental 4DVAR

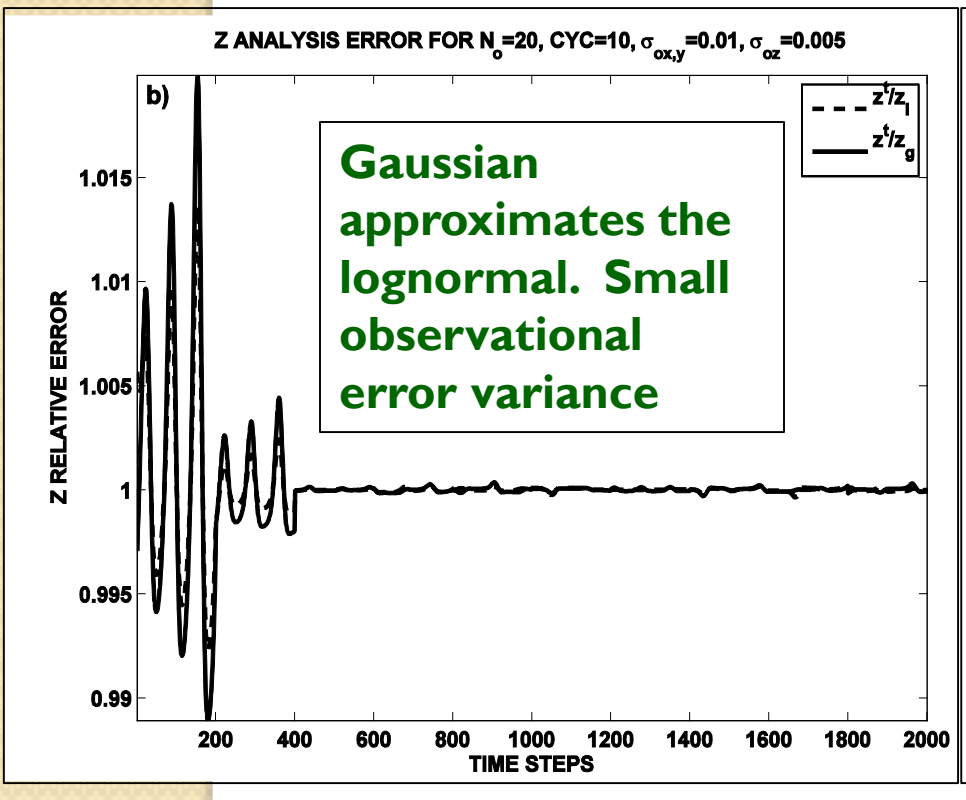


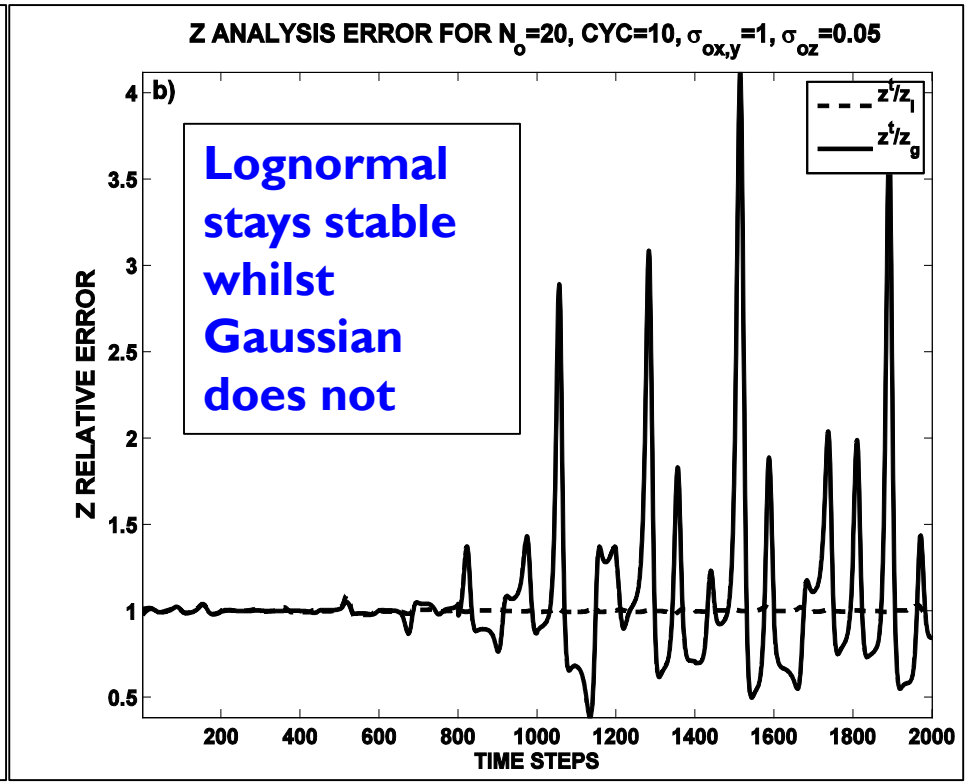


Results for same number of assimilation windows but with accurate observations every other time step

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.





Lognormal Detection Algorithm

There is a lot of growing evidence that the distribution of certain atmospheric, and oceanic fields, do not have a Gaussian distribution continuously throughout the year. Therefore the question becomes, how can we know when to use a lognormal or a Gaussian based data assimilation system?

Dr. Anton Kliewer, a NSF funded postdoctoral fellow at CIRA/CSU has been developing an algorithm that is based upon a composite of three statistical tests which determine if the data that you are testing are presenting a lognormal signal, a Gaussian signal or a non-Gaussian signal that is not a lognormal but does not determine this other distribution.

Similar work to detect non-Gaussian signals is being undertaken at Météo-France.

Lognormal Detection Algorithm

Hypothesis Tests:

Shapiro-Wilk – combines the information that is contained in the normal probability plot with the information obtained from the estimator of the standard deviation of the sample.

Jarque-Bera – a goodness-of-fit test of whether sample data have the skewness and kurtosis matching a normal distribution.

Chi-Squared Goodness-of-Fit – determines whether there is a significant difference between the expected frequencies and the observed frequencies from a probability distribution.

Composite – combination of the previous three tests.

(Kliewer *et al*, 2015a)

Kliewer, A. J, S. J. Fletcher, J. M. Forsythe and A. S. Jones 2015a: Identifying Non-normal and lognormal characteristics of temperature, mixing ratio, surface pressure, and winds for data assimilation systems. Submitted to *Nonlinear Processes in the Geosciences*

Lognormal Detection Algorithm

For Shapiro-Wilk and Jarque-Bera, the null hypothesis is that the data comes from a Gaussian distribution. The alternative hypothesis is the data does not come from a Gaussian distribution.

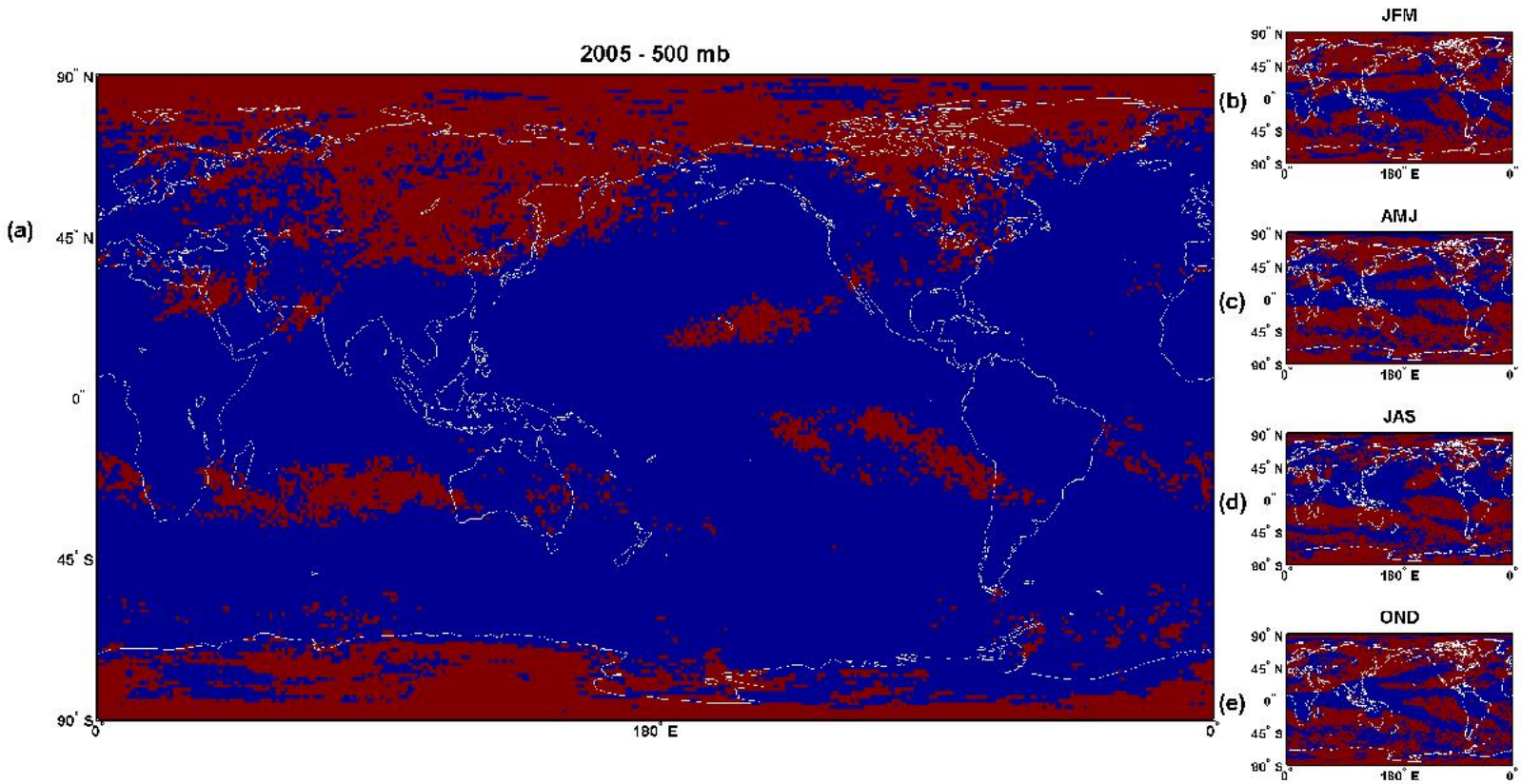
For the Chi-squared the null hypothesis is that the data comes from a lognormal distribution. The alternative hypothesis is the data does not come from a lognormal distribution.

The composite test combines these results – a positive result (red on images) indicates that the data does not come from a Gaussian distribution (as determined by Shapiro-Wilk and Jarque-Bera) and it does come from a lognormal distribution (as determined by Chi-squared). A negative result (blue on image) indicates that either Shapiro-Wilk or Jarque-Bera determined the data to come from a Gaussian distribution, or the Chi-squared determined the data does not come from a lognormal distribution.

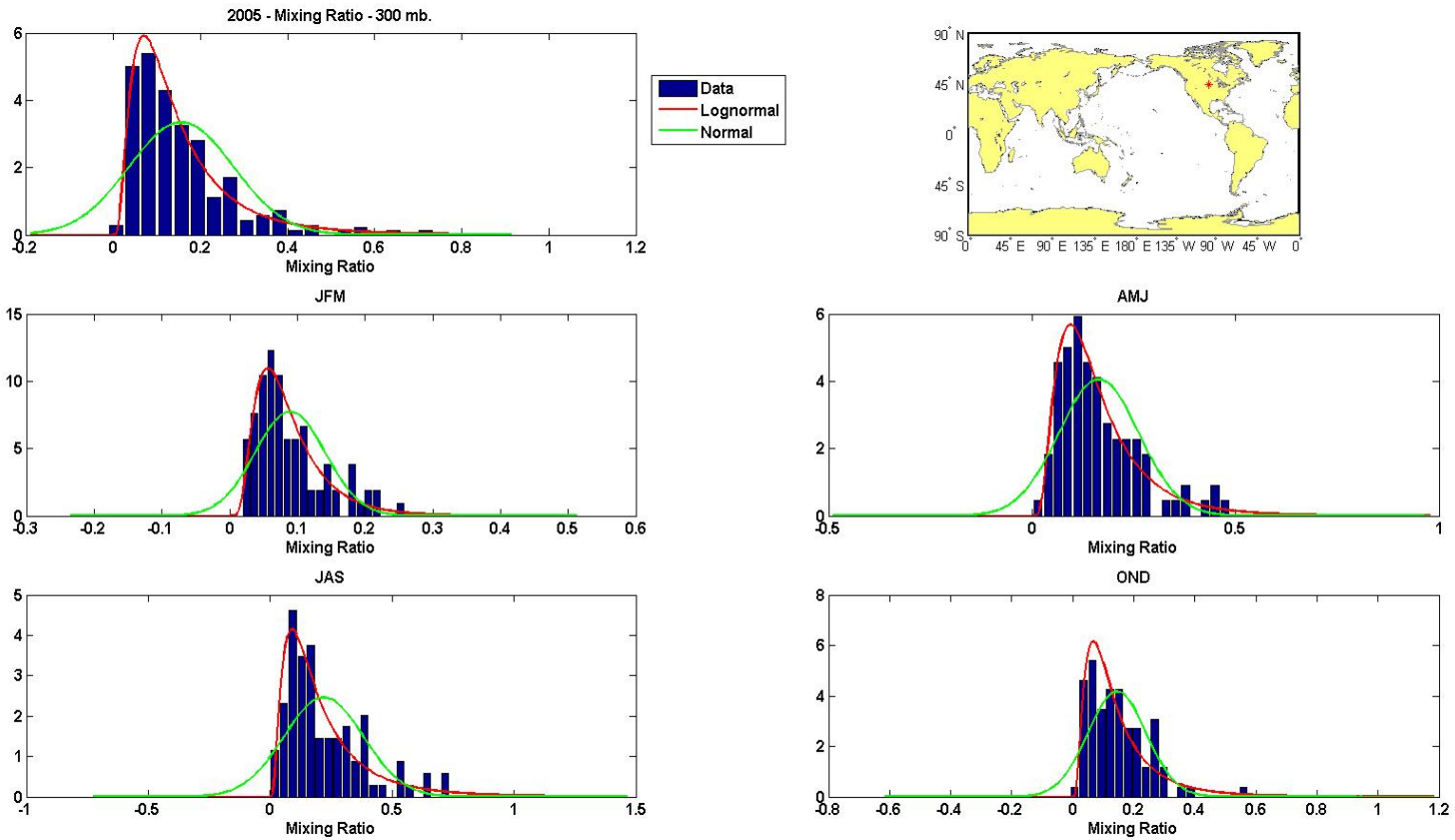
All tests conducted at the α=0.01 significance level (99% confidence).

For this study one year (2005) of Global Forecast System 0 hr forecasts from the 0z runs of mixing ratio were treated as observations of the ``true state'' and analyzed using statistical hypothesis testing procedures.

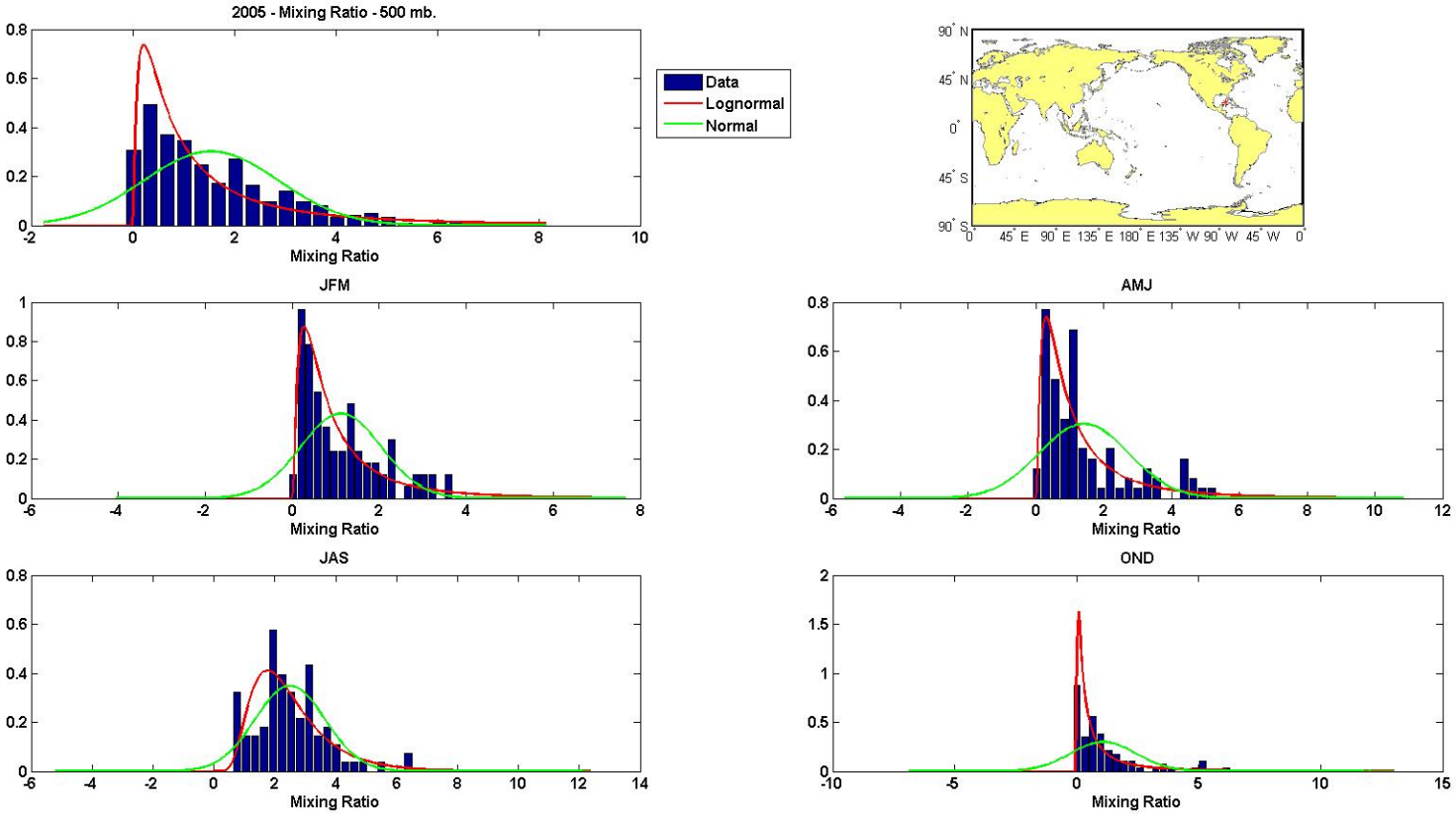
Global 500hPa mixing ratio



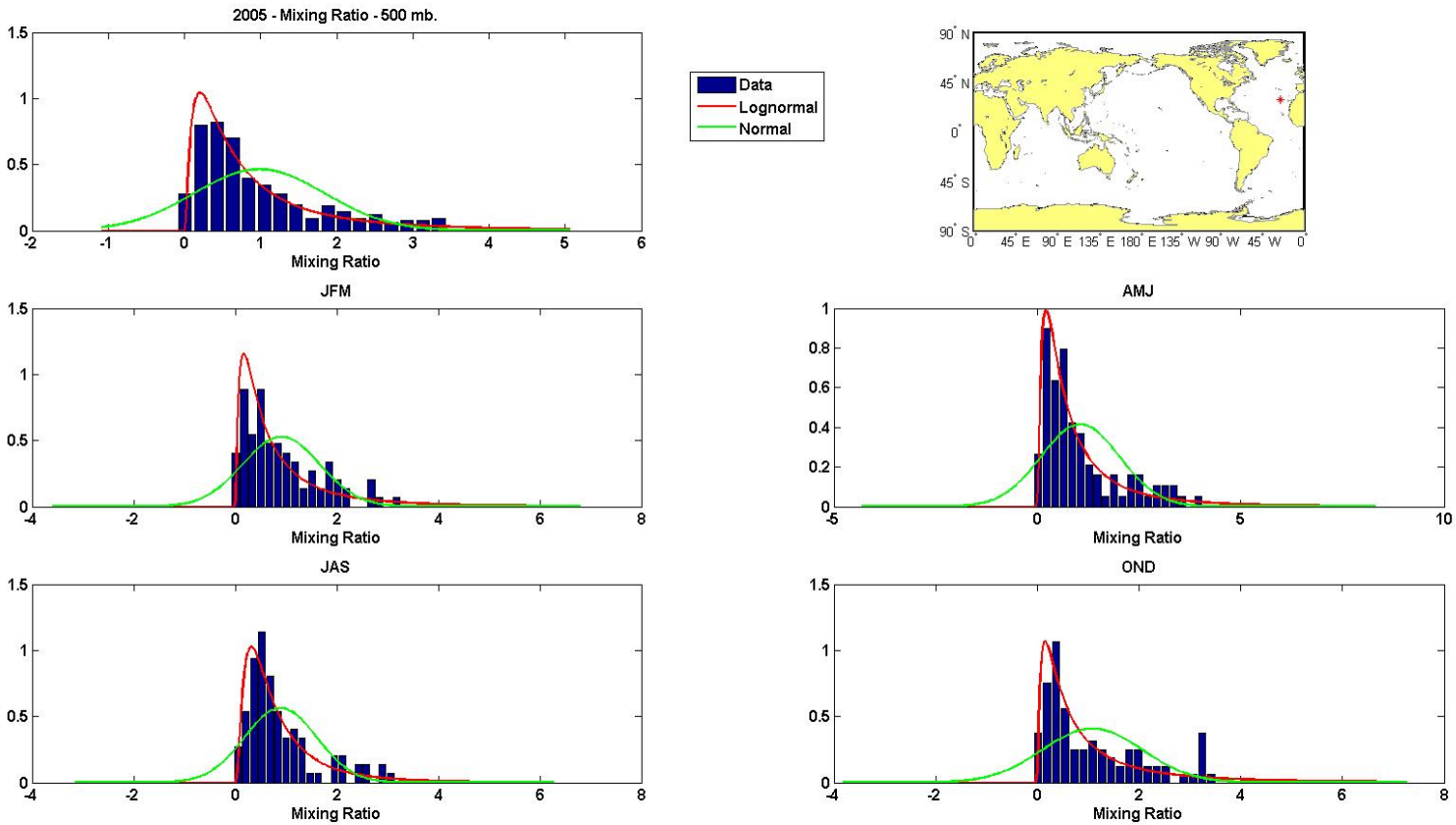
Continental United States



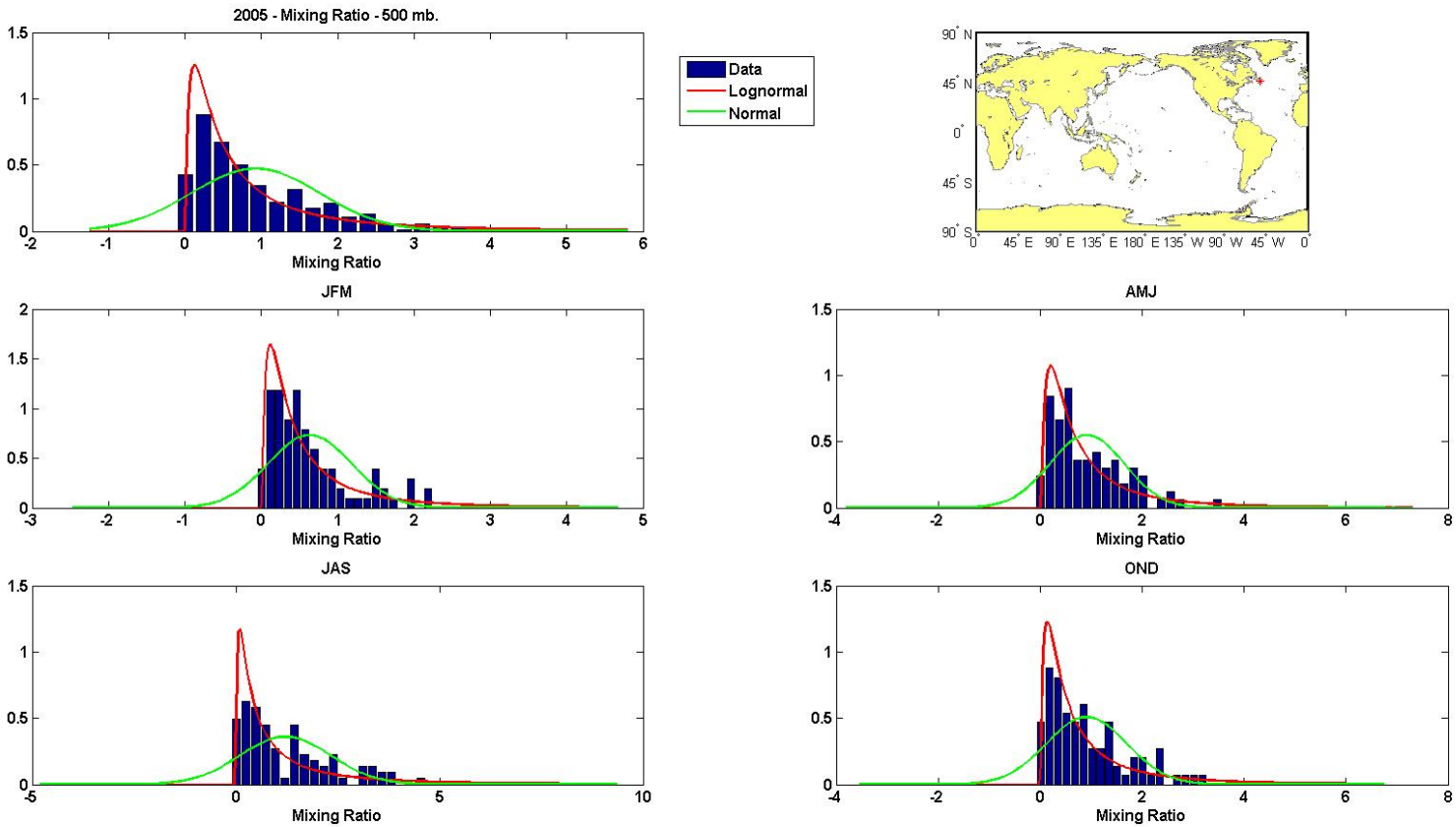
Gulf of Mexico



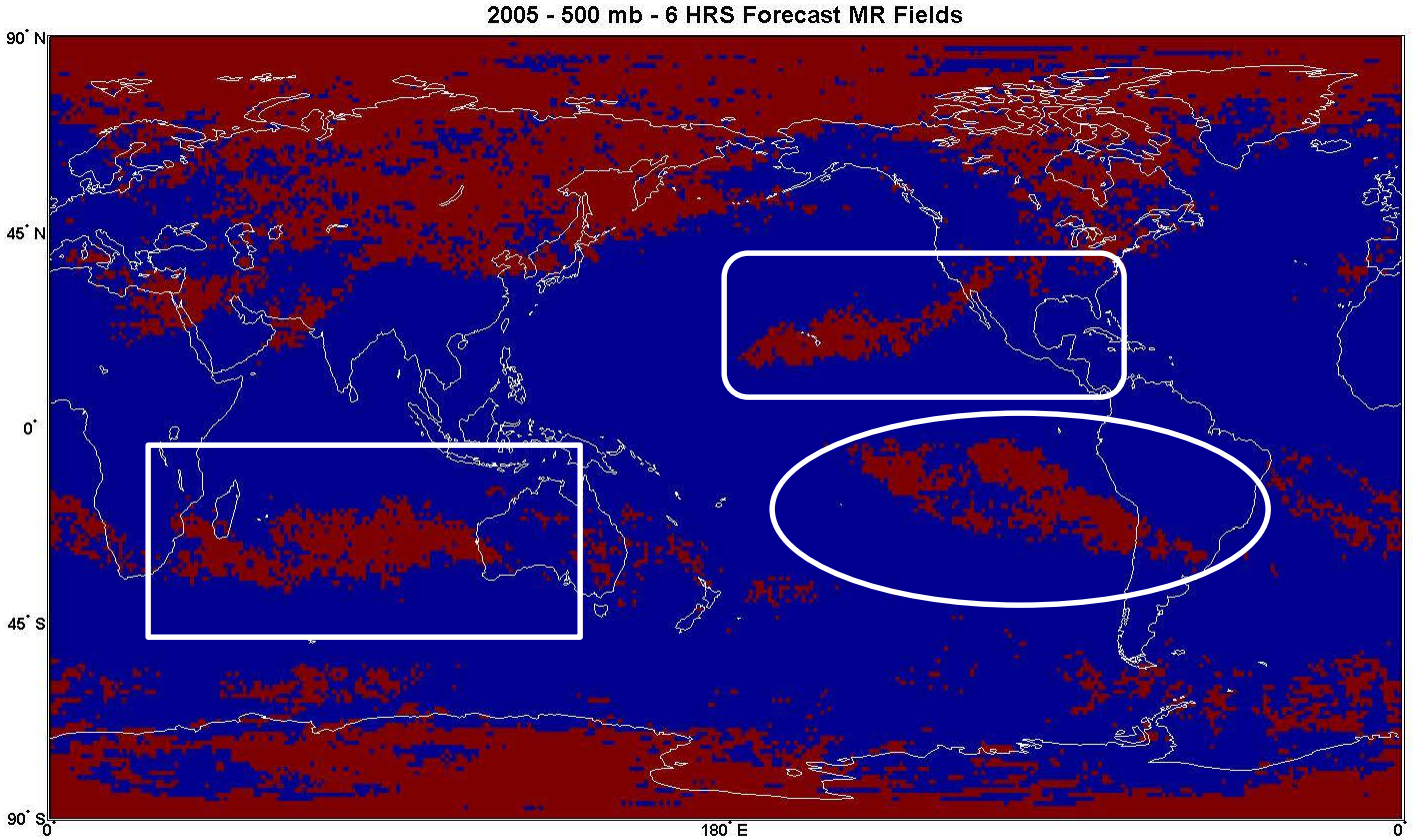
Off the coast of Africa



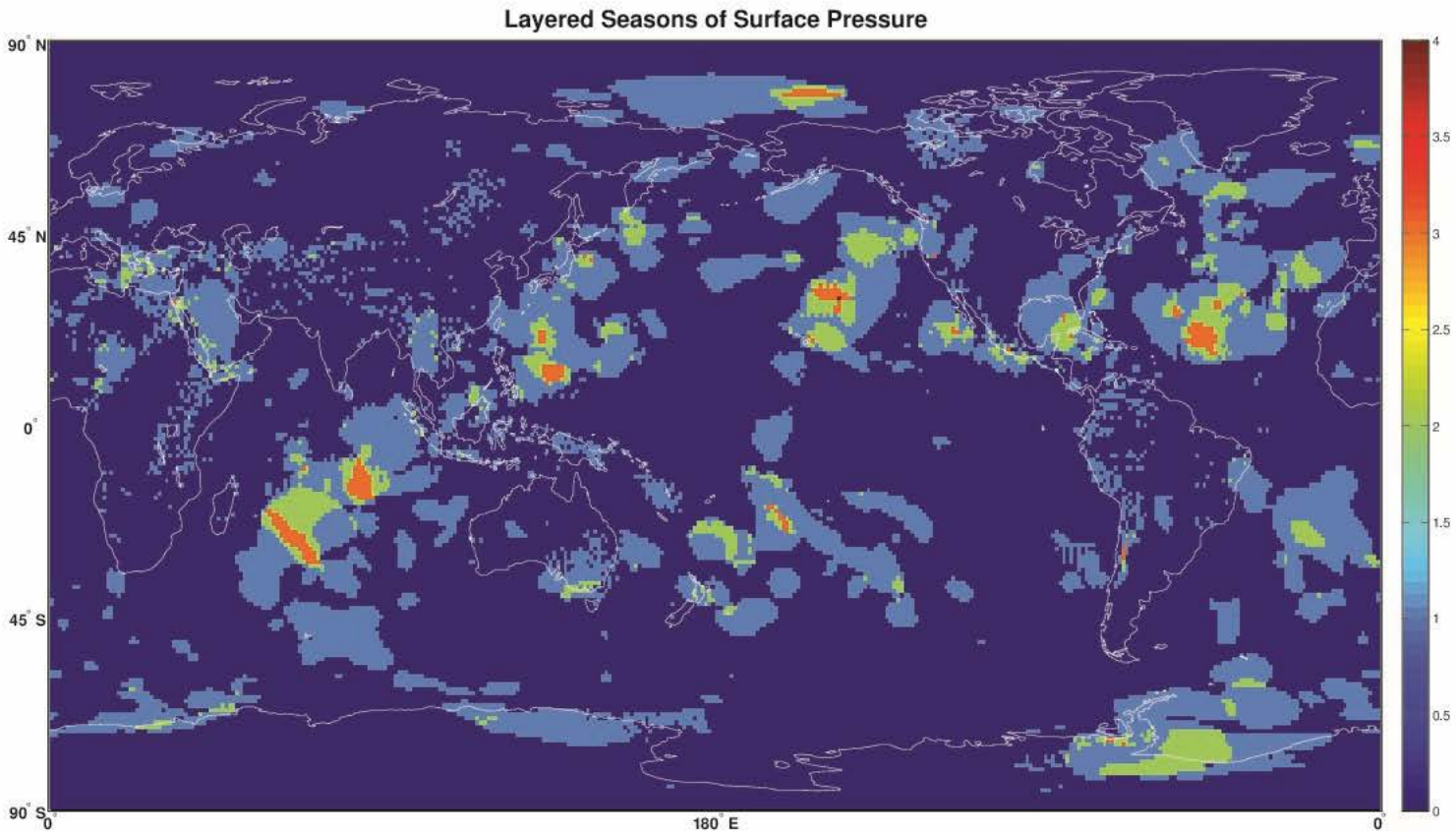
North Atlantic/Mid-Latitudes



Same analysis but now with the 6 hour forecast which is used as the first guess for the next assimilation cycle.

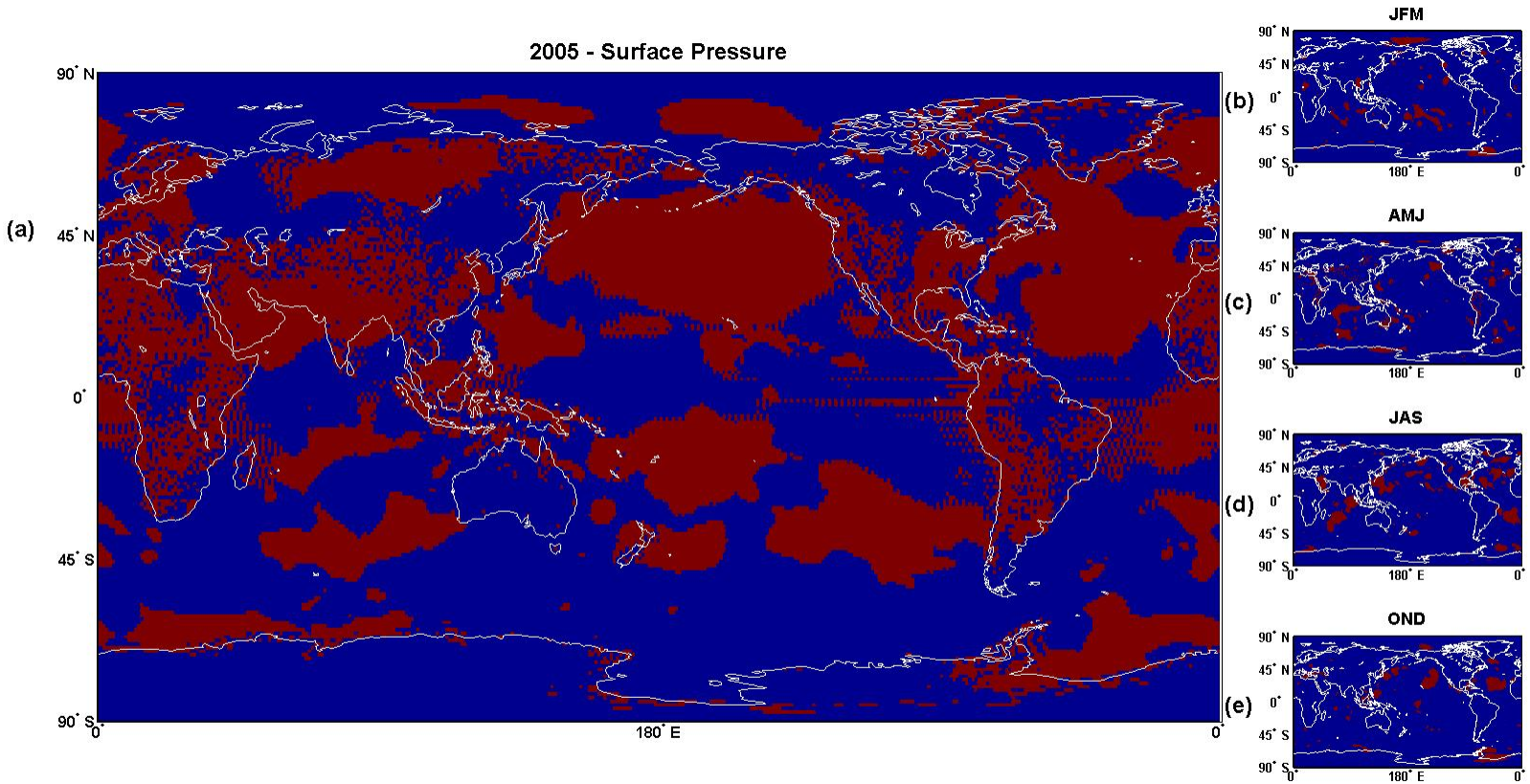


Surface Pressure Non-Gaussian Detection



Indicators of when surface pressure by three month groupings had a non-Gaussian signal, but not a lognormal signal.

Surface Pressure Non-Gaussian Detection



A composite for the whole of 2005 and broken down by 3 month groupings to detect non-Gaussian signals for surface pressure.

Mixed Lognormal-Gaussian Mixing Ratio – Temperature Retrievals

At CIRA we have implemented the full-field mixed lognormal-Gaussian theory into the CIRA1-Dimensional Optimal Estimator (C1DOE). C1DOE is a microwave retrieval system that retrieves emissivities, mixing-ratio and temperature.

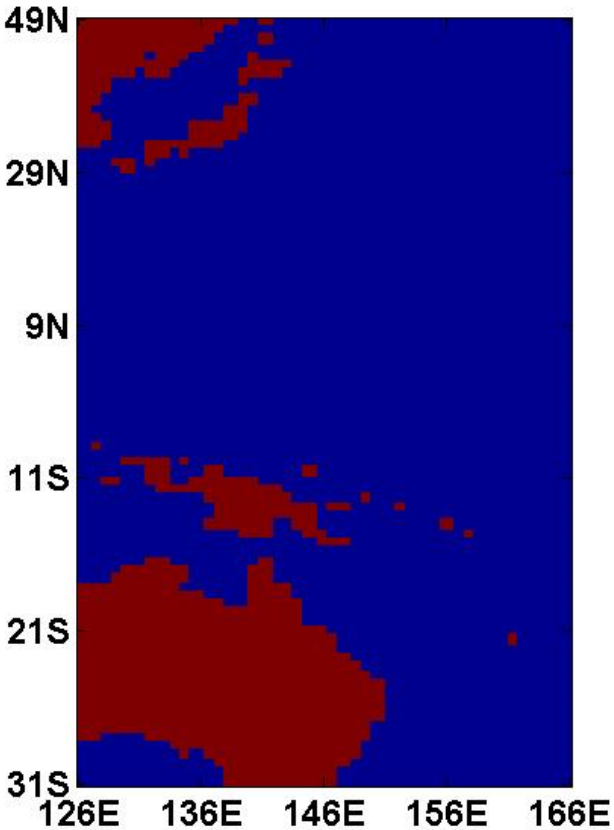
For the experiments that we performed we updated the background error covariance matrices for both the Gaussian and mixed formulations. These updates were calculated from GDAS data for the month of September 2005. We have adapted the retrieval system to also perform the transform approach as well as the mixed formulation.

C1DOE is a microwave radiance/brightness temperature based retrieval system using the radiative transfer theory set out in Liebe (1989) and the subsequent revisions in 1992. A good description of C1DOE formulations can be found in Engelen and Stephens (1999). The retrievals are performed on 7 vertical levels at 1000hPa, 850hPa, 700hPa, 500hPa, 200hPa and 100hPa.

The experiment were ran for 10 days from 09/01/2005 – 09/10/2005. The cost function is solved using a Newton-Raphson Solver where we are using the correct Jacobians and are inverting the analytical Hessian matrices.

Engelen, R. J. and G.L. Stephens, 1999: Characterization of water-vapour retrievals from TOVS/HIRS, and SSM/T-2 measurement. *Q. J. Roy. Meteorol. Soc.,* **125,** 331—351.
Liebe, H.J., 1989: MPM – An atmospheric millimeter-wave propagation model. *Int. J. Infrared and Millimeter Waves,* **10,** 631—650.

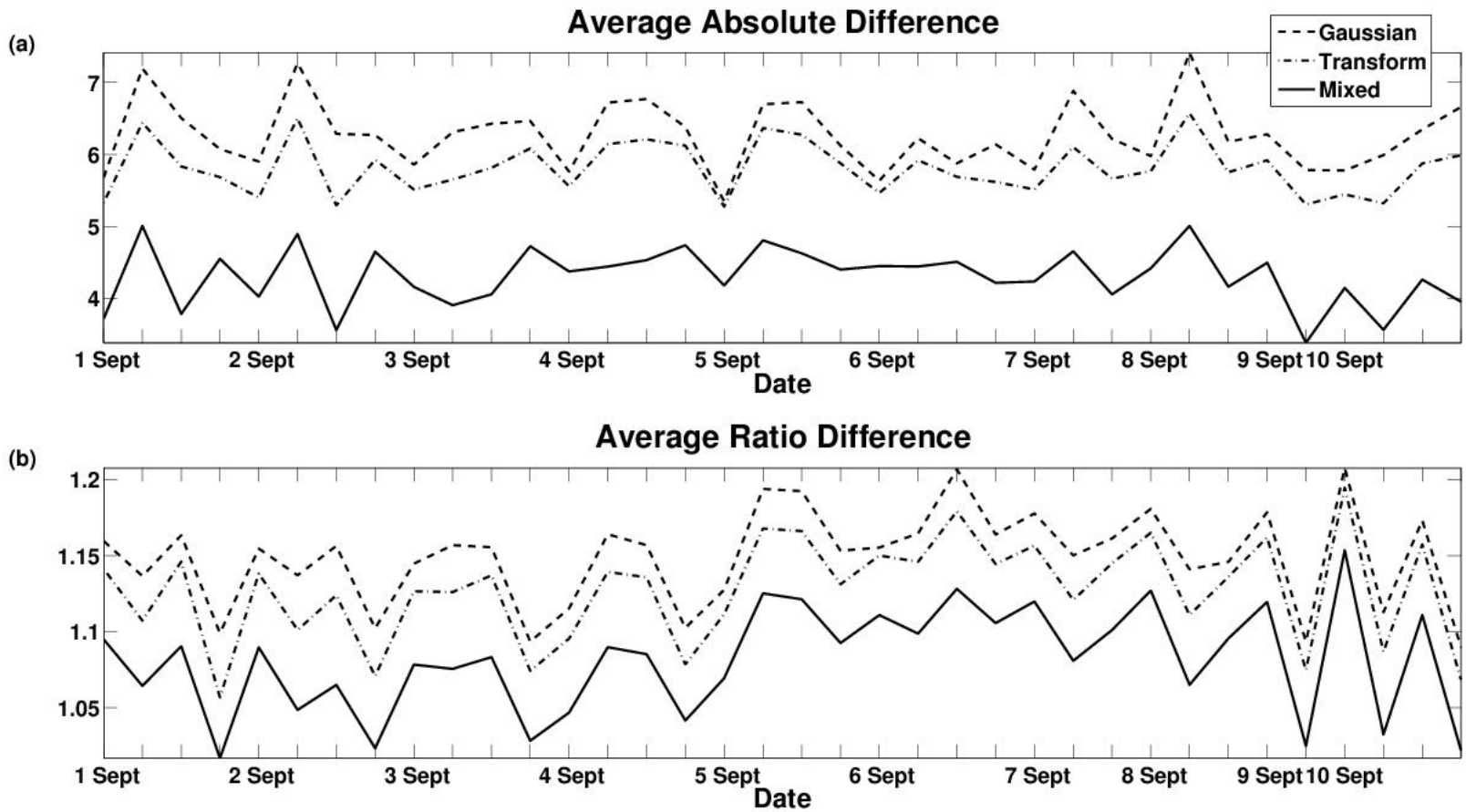
Mixed Lognormal-Gaussian Mixing Ratio – Temperature Retrievals



Field of View for the experiments presented in Kliewer *et al.*, (2015b) where a non-Gaussian signal for moisture had been detected near the coast of Japan for this year from GPS station data.

Kliewer, A.J., S.J. Fletcher, A.S. Jones and J.M. Forsythe, 2015b: Comparison of Gaussian, Logarithmic transform and Mixed Gaussian-Lognormal distribution-based 1DVAR Microwave Temperature-Water Vapor Mixing Ratio. Under Revision for *Q. J. Roy. Meteorol. Soc.*

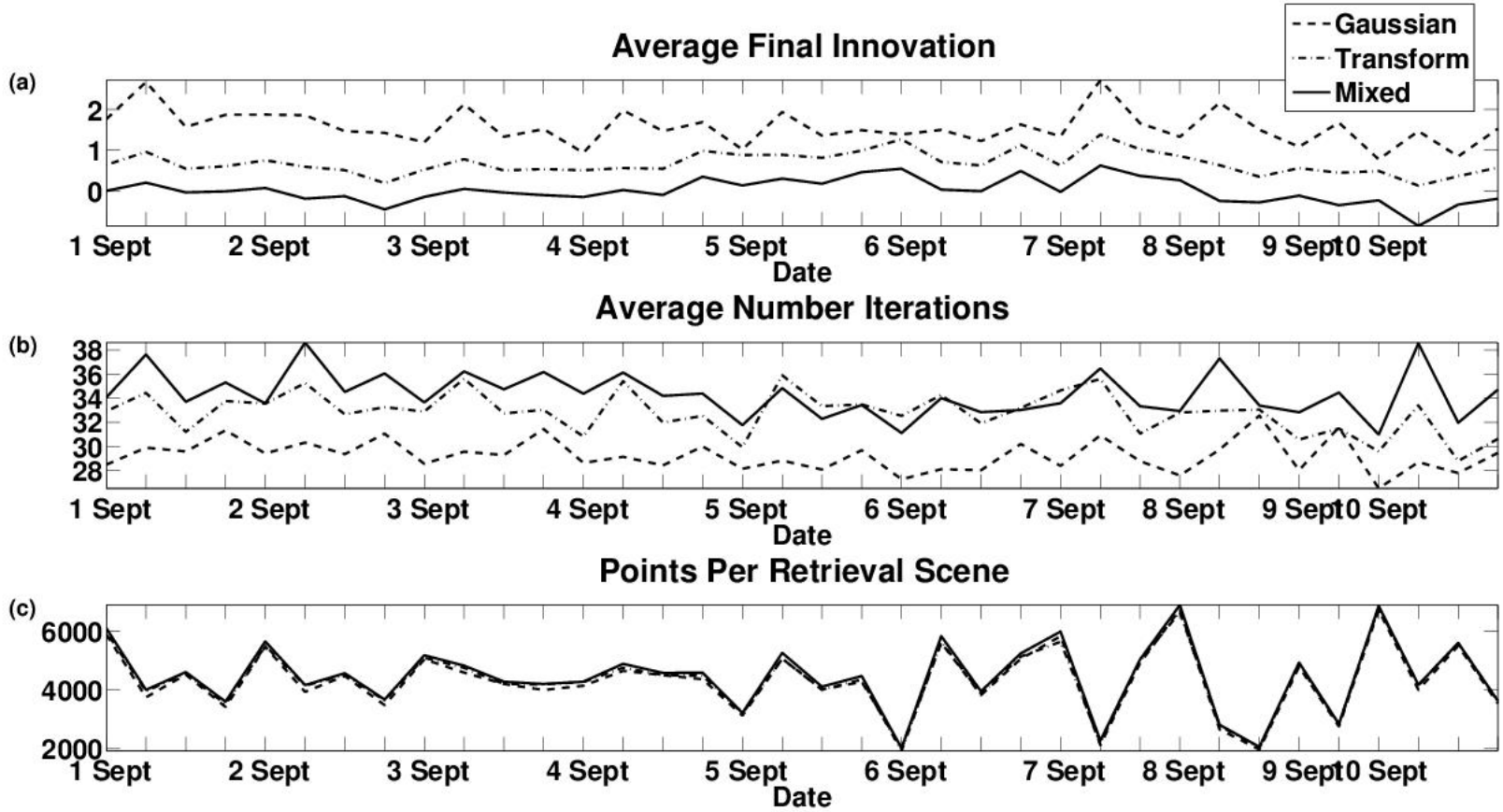
Mixed Lognormal-Gaussian Mixing Ratio – Temperature Retrievals



Comparisons of the three retrieval methods against the Microwave Surface and Precipitation Products Systems (MSPPS) TPW product. Solid is the mixed approach, dot-dashed is the transform and the dashed is the Gaussian.

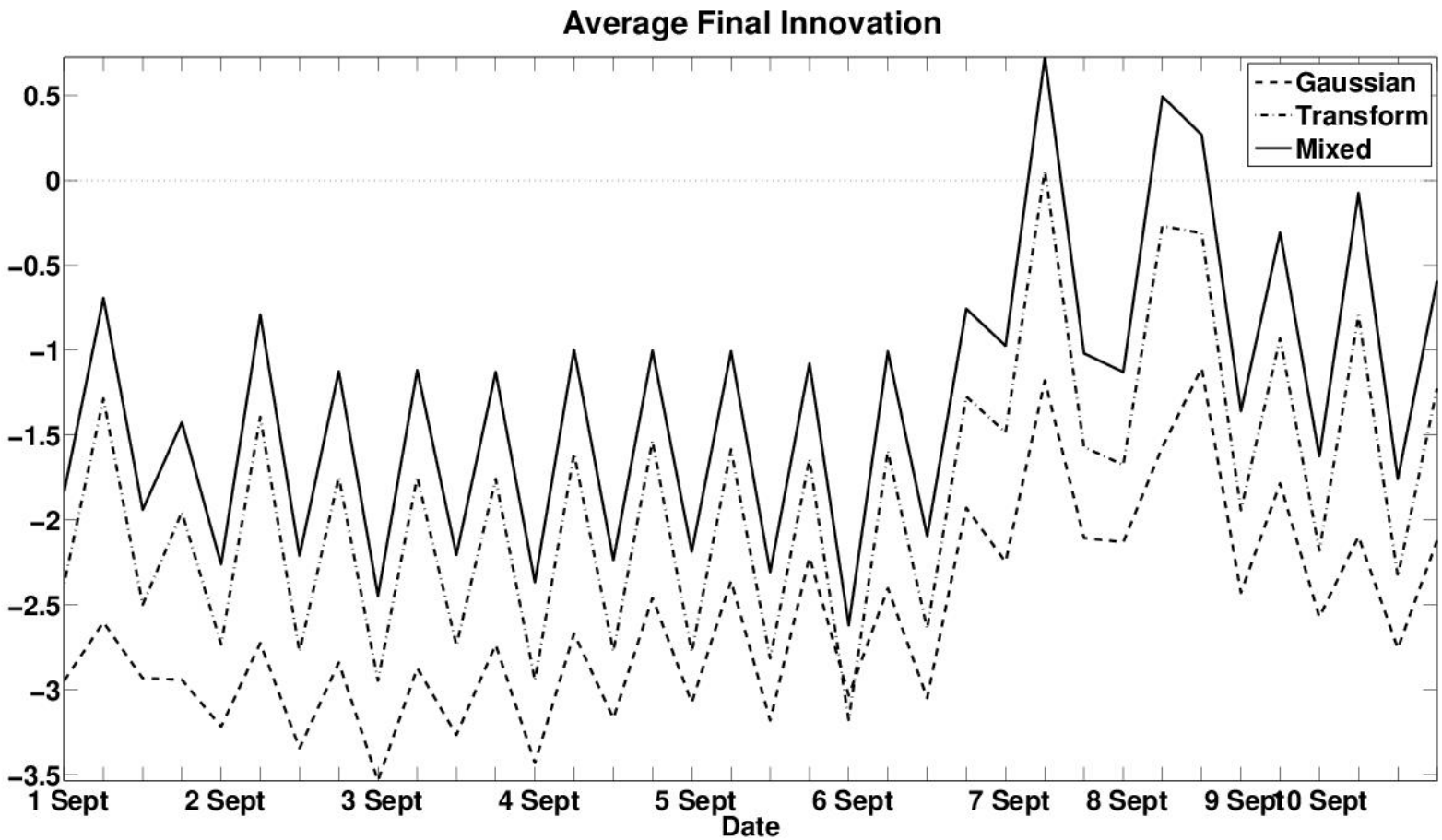
Mixed Lognormal-Gaussian Mixing Ratio – Temperature Retrievals

AMSU-B Channel 4 (183 ± 3GHz) (Water Vapor Channel in the troposphere) Final Innovations



Mixed Lognormal-Gaussian Mixing Ratio – Temperature Retrievals

AMSU-A Channel 6 (54.4GHz) (Temperature Channel in the troposphere) Final Innovations



Lognormal Based Quality Control Measures

The gross observational error quality control measure is roughly given by

$$|y - h(x_b)| < 3\sigma_o,$$

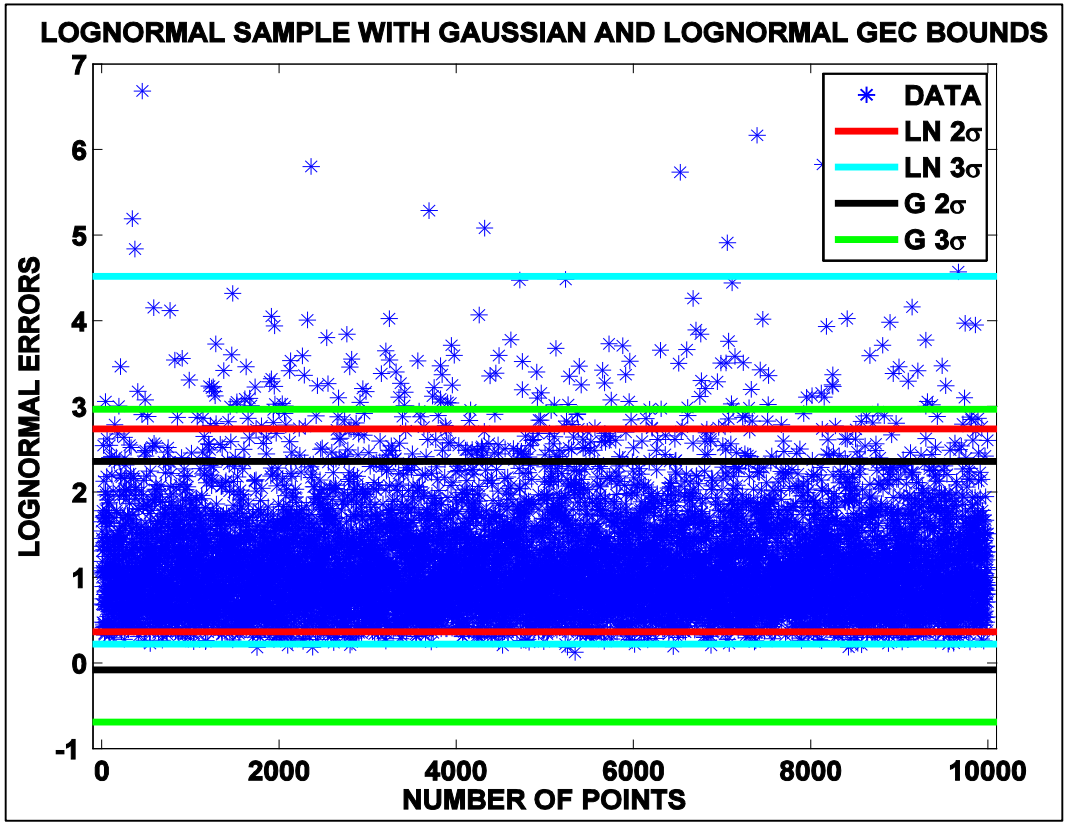
where σ_o is the observational error's standard deviation. This measure is accounting for approximately 99% of the distribution.

Given that this measure is consistent with percentiles, we would consider something similar for lognormally distributed errors.

If we apply the logarithmic transform to the lognormal random variable to make it a Gaussian random variable, then we could simply use the property of the **preservation of percentiles**, which means that equivalent lognormal gross error check could be

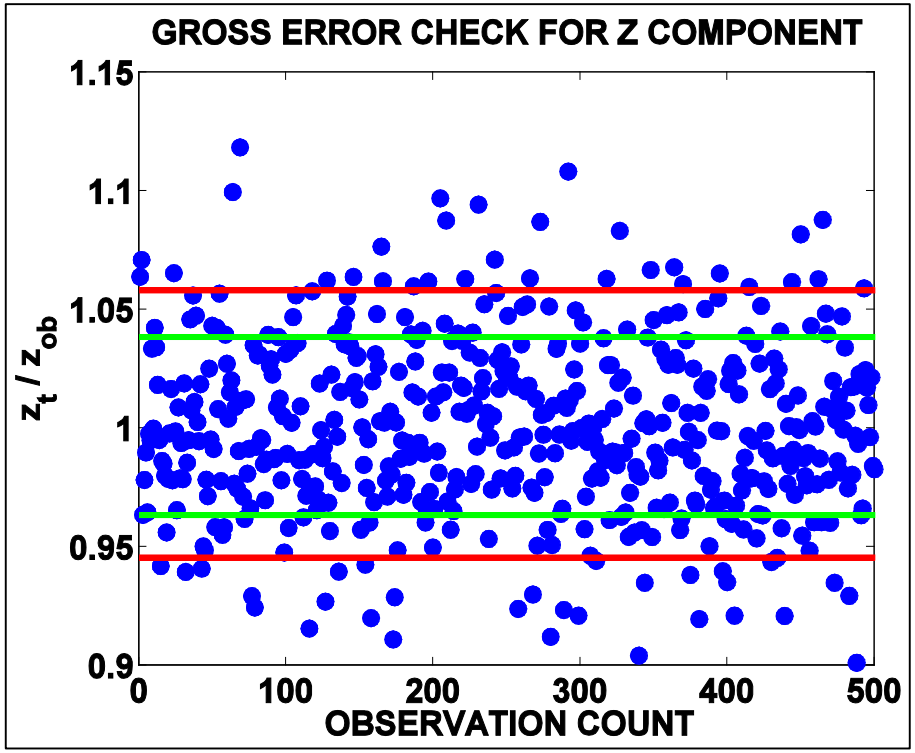
$$|\ln y - \ln h(x)| < 3\sigma_{o,L} \quad or \quad \exp\{-3\sigma_{o,L}\} < \frac{y}{h(x)} < \exp\{3\sigma_{o,L}\}.$$

Lognormal Based Quality Control Measures



Plot of a 10000 lognormal random numbers from $LN(0, 0.5)$ with the equivalent Gaussian and lognormal gross error check bounds.

Lognormal Based Quality Control Measures



An example from the Lorenz 63 model where the observations were generated from a lognormal random number generator multiplying the z_t with values for the mean and standard deviation of 0 and 0.0375 respectively. The green line is 2 SD, whilst the red is 3 SD.

Optimal Regions for Descriptive Statistics for Lognormal Based DA

During the implementation of a set of synthetic experiments with C1DOE to ascertain the accuracy of the three different approaches for a know true solution, it became apparent that something was affecting the performance of the modal based approach in that it could not beat the median approach even though we thought the problem had been set up so that the mode should have been the optimal estimator.

We ran a series of experiments with a univariate formulation with the lognormal cost function and used a Newton-Raphson solver to simulate the processes in C1DOE. During these experiments we discovered 4 important findings relative to the performance of a lognormal modal based variational problem (Fletcher *et al.,* 2015).

I. The three descriptive statistics, mean, median and mode, had regions relative to the apriori state such that they were the statistic that minimized the errors.

II. There existed a value for the a priori state for each statistic such that the minimum of the cost function was at the true state.

III. That if an a priori state was close to this optimal value then the Newton-Raphson solver became chaotic and incredibly sensitive to the first guess to the solver.

IV. The a priori state should not be the best approximation to the true state for the lognormal modal approach but rather to $E(\ln \boldsymbol{x}^t)$. This is because

$$LN(\mu, \sigma^2) \approx \frac{1}{x} \exp\left\{-\frac{1}{2}\left(\frac{\ln x \ - \mu}{\sigma}\right)^2\right\}$$

If the a priori state is close to the true state the median is optimal in minimizing the errors whilst if the a priori state is close μ then the mode is the optimal statistic.

Fletcher, S. J., A. J. Kliewer and A. S. Jones, 2018: Quantification of optimal choices of parameters in lognormal variational data assimilation and their chaotic behavior. Submitted to *Mathematical Geosciences*

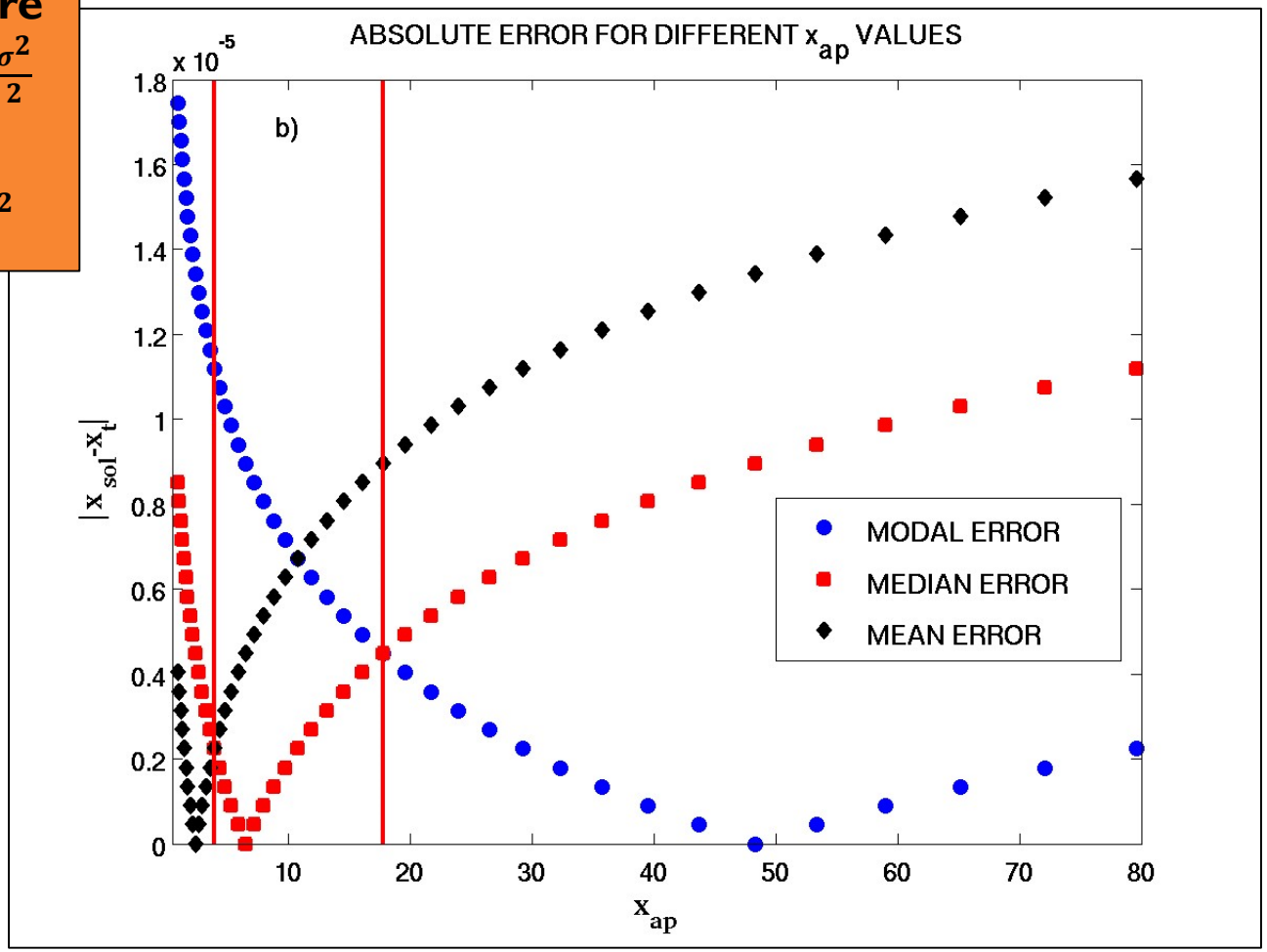
Optimal Regions for Descriptive Statistics for Lognormal Based DA

Optimal values are

Mean: $x_{ap} = x^t e^{-\frac{\sigma^2}{2}}$

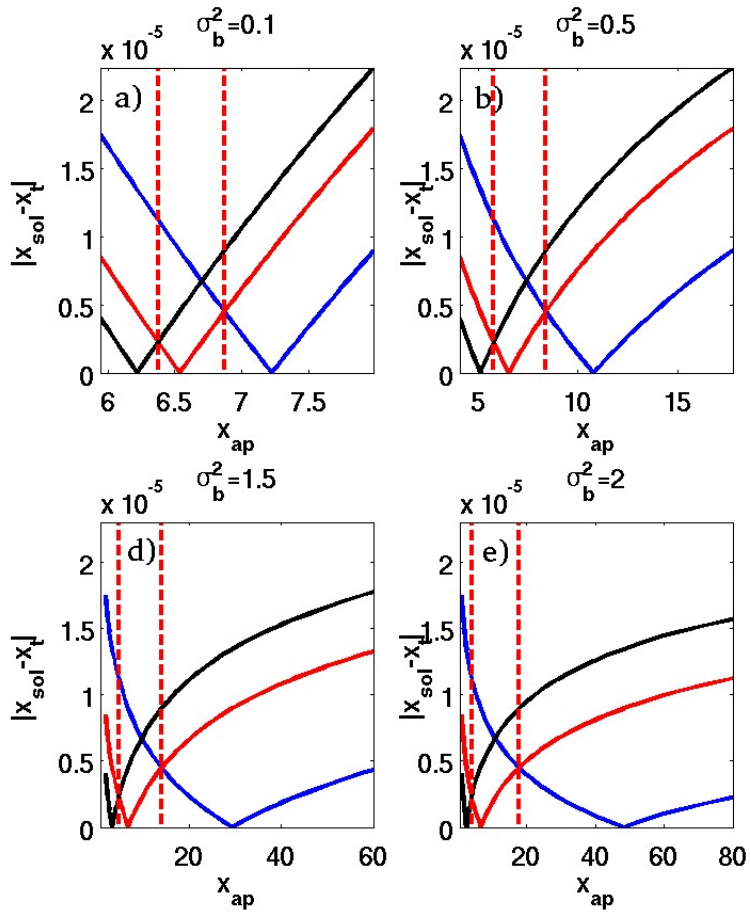
Median: $x_{ap} = x^t$

Mode: $x_{ap} = x^t e^{\sigma^2}$



Plot of absolute error for different values for the apriori state for different lognormal descriptive statistic cost functions.

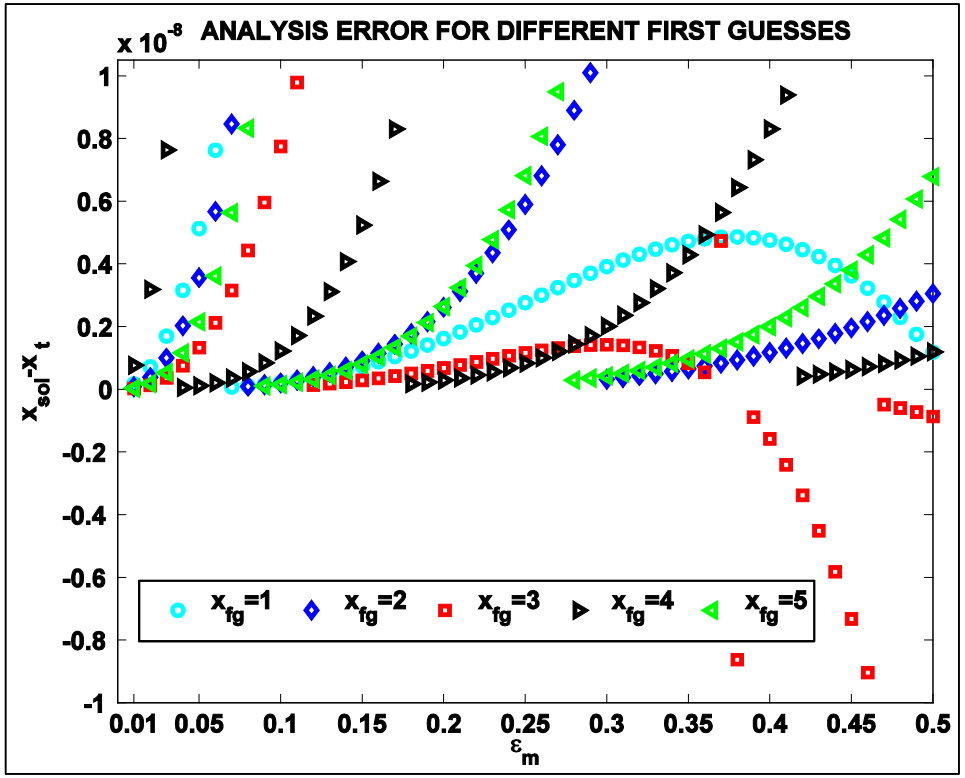
Optimal Regions for Descriptive Statistics for Lognormal Based DA



Impact of the variance on the range where the median minimizes the error.

Chaotic Signal with Newton-Raphson solver and Lognormal DA

It was when we introduced a measurement error to determine what the optimal value for the a priori state (or the observational and background variances) that we started to detect that there was something happening to the accuracy of the solution of the modal approach with the Newton-Raphson solver with respect to the choice of first guess to the solver!!

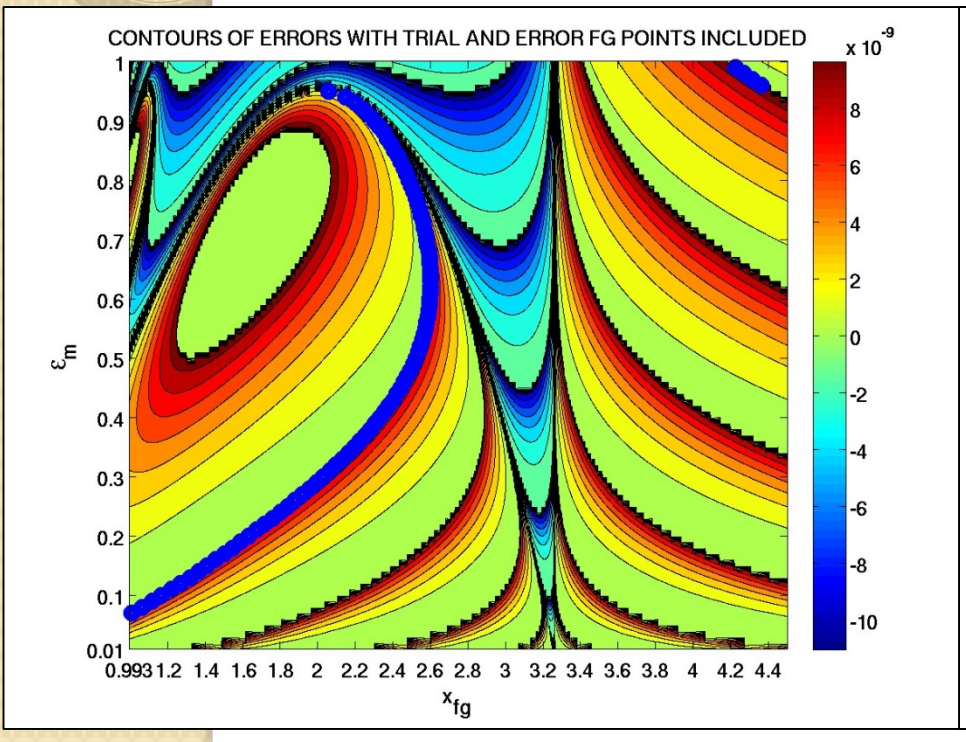


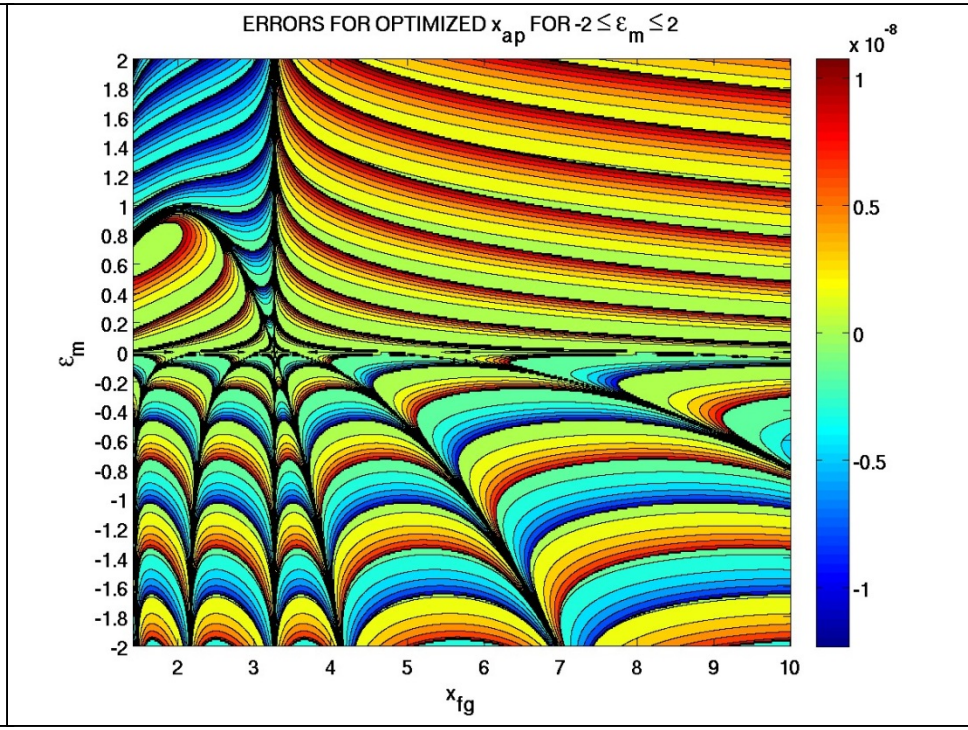
Chaotic Signal with Newton-Raphson solver and Lognormal DA

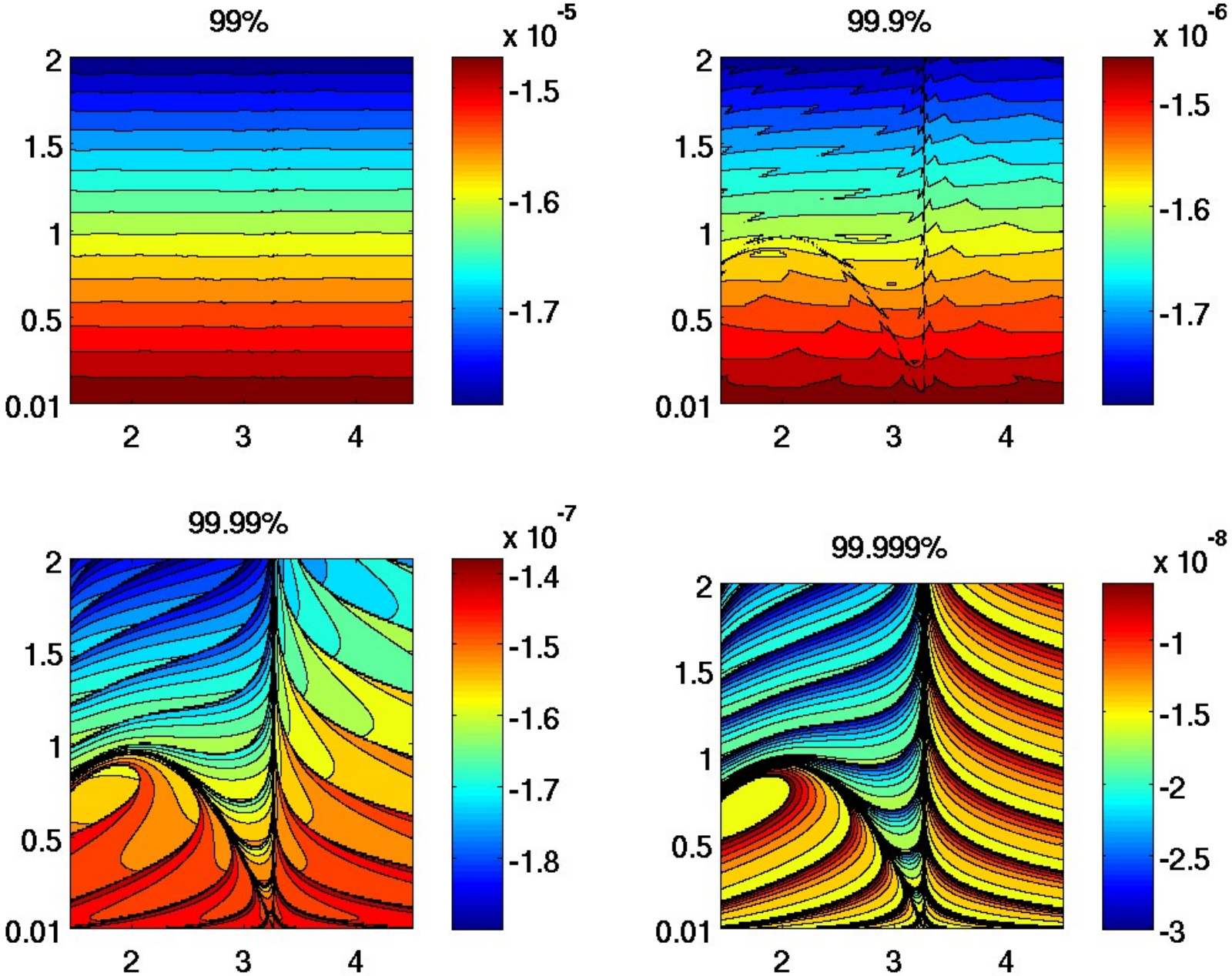
Upon a series of trial and error experiments with the first guess values we started to see that the point where we needed to start from to ensure accuracy of $\mathbf{10^{-8}}$in our solution or better was changing, but that this value at every decimal point was changing!!!

We would get to a value at each decimal point that if we went past that number at that decimal point we would lose 3 to 4 figures of accuracy!! To investigate this feature we ran do loops over first guess to the Newton-Raphson solver stepping by 0.001 and for measurement errors stepping at 0.001 from 0 to 1. This is what we found

Chaotic Signal with Newton-Raphson solver and Lognormal DA






99%
x 10^-5
-1.5
-1.6
-1.7
99.9%
x 10^-6
-1.5
-1.6
-1.7
99.99%
x 10^-7
-1.4
-1.5
-1.6
-1.7
-1.8
99.999%
x 10^-8
-1
-1.5
-2
-2.5
-3
2
1.5
1
0.5
0.01
2
3
4

Chaotic Signal with Newton-Raphson solver and Lognormal DA

This signal is not only present when the a priori state is optimized for measurement error. It is also present when the a priori state is optimized for representative errors as well as in the Gaussian formulation and the incremental lognormal formulation.

A different *fractal* appears for each and for different true states as well as different background and observational errors. It is also present when the background or the observational errors are optimized to correct for the other incorrect statistics to ensure that the minimum of the cost function is at the true state.

This patterns are referred to as **Newton-Fractals** and occur when the quadratic convergence breaks down.

Conclusions

- It is possible to go non-Gaussian in the variational framework, while not shown here it is also possible to go non-Gaussian with the PSAS system.
- We have to be careful when selecting which descriptive statistics we use to minimize the lognormal errors, which is dependent on how close the a priori state is to true state.
- There appears to be a sensitivity in the Newton-Raphson solver when the cost function is at the "optimal" values.
- It is possible to extend the Gaussian based observational quality controls to lognormal formulations.
- Possible to combine different distributions to create mixed distributions to model the covariances between random variables that are of different distributions.

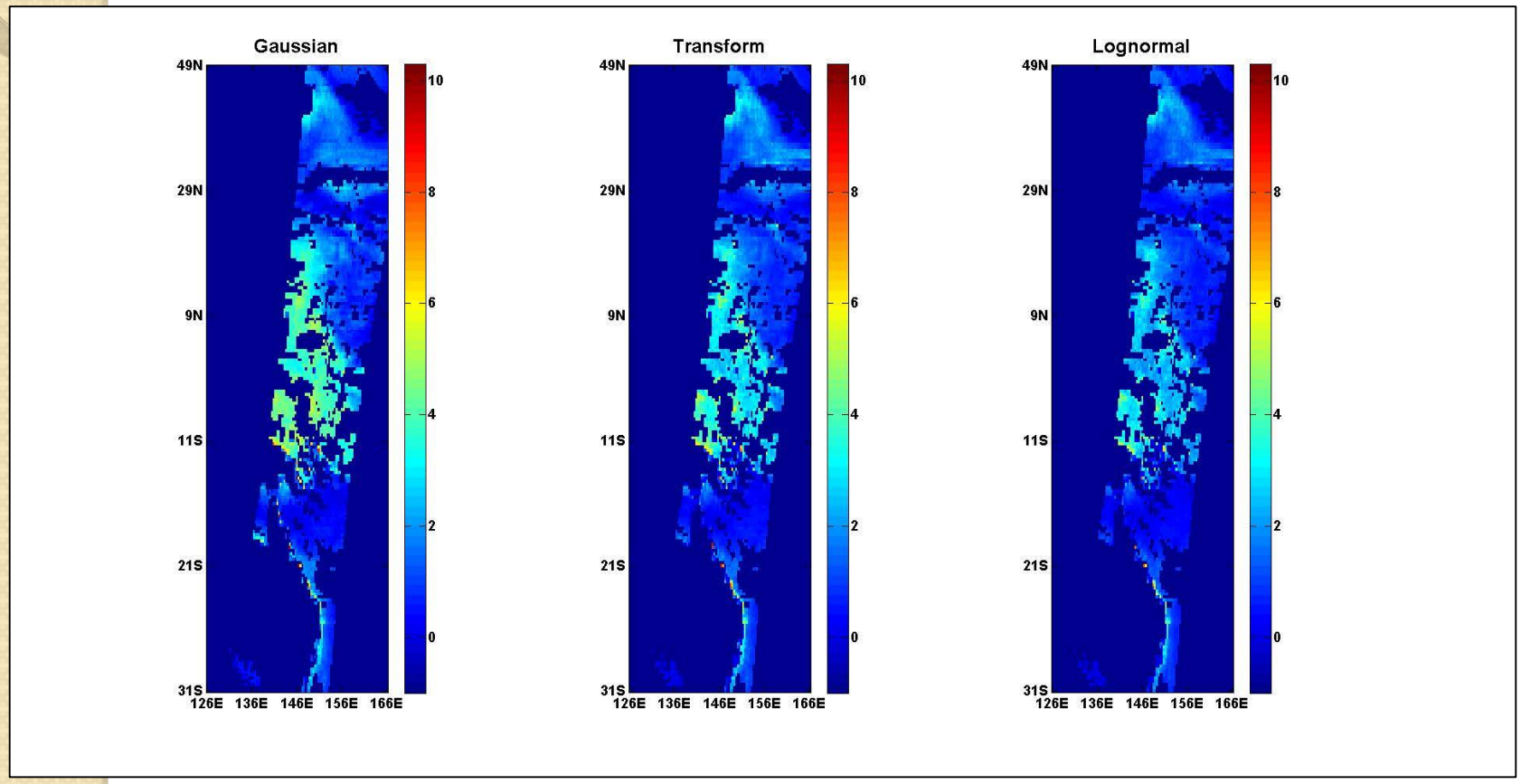
CIRA Data Assimilation Testbed (CDAT)

As part of a new NSF award to CIRA we are developing a website that will be running the three different versions of the retrieval system in near real time in different regions over the Earth.

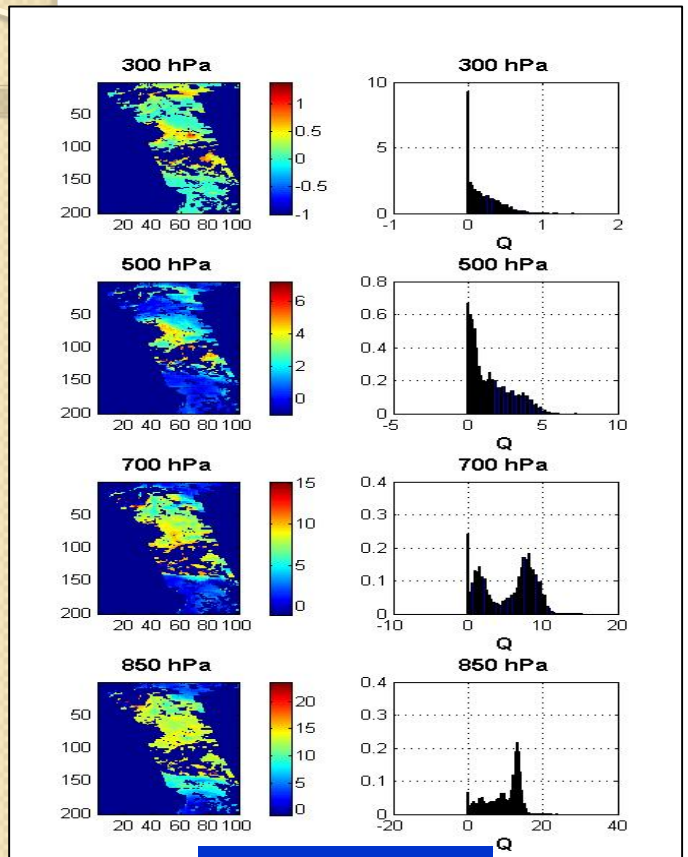
Along with the output from the three different retrievals systems, there will also be the output from the detection. The purpose of also showing the detection algorithm output on the website is for users to look at where a non-Gaussian signal has been detected and to look at the impact on the values from the retrieval systems to make decision of where the Gaussian fits all could be sub-optimal in those areas.

The website address will be http://.cdat.cira.colostate.edu

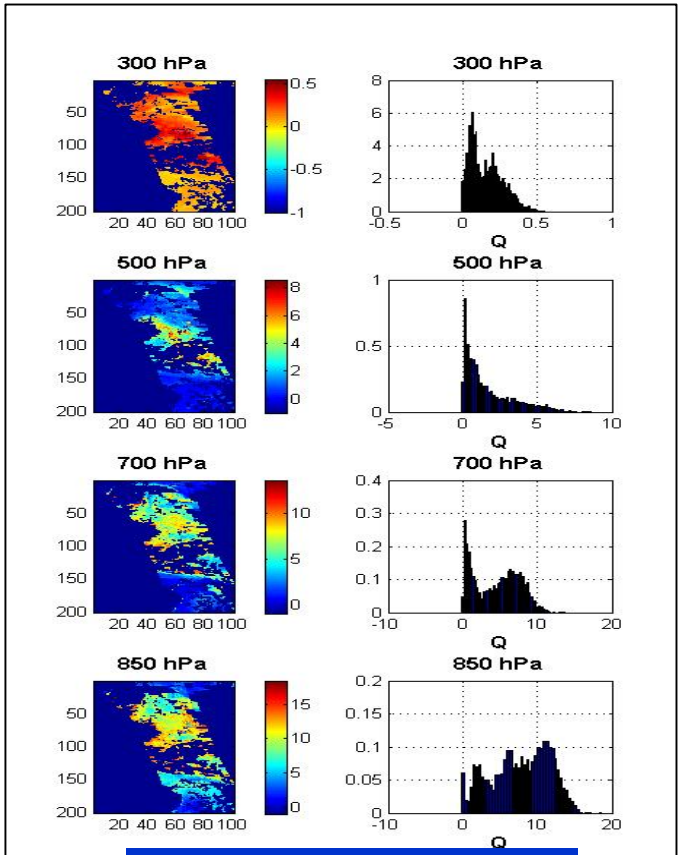
CDAT


Gaussian
Transform
Lognormal
49N
29N
9N
11S
21S
31S
126E 136E 146E 156E 166E
10
8
6
4
2
0

http://cdat.cira.colostate.edu



Gaussian Fits all



Mixed Gaussian-Lognormal