Introducing CDAT: CIRA Data Assimilation Testbed

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

Cooperative Institute for Research in the Atmosphere, Colorado State University

Plan of talk

- Introduction to the mixed Gaussian-lognormal Distribution
- Properties of the mixed distribution
- Plots of the mixed distributions
- Applying the mixed distribution to a variational formulation
- Application of the mixed distribution to microwave brightness temperature based temperature-humidity retrievals
- Comparison with Gaussian and the logarithmic transform approach.
- CIRA Data Assimilation Testbed - CDAT

Mixed Gaussian-Lognormal distribution

The mixed distribution in its bivariate formulation is defined by

$$MX(\mu_G, \mu_L, \sigma_G, \sigma_L, \rho_{mx}) \equiv \frac{1}{\sqrt{|\boldsymbol{\Sigma}_{mx}| 2\pi x_2}} \exp\left\{-\frac{1}{2}\begin{pmatrix} x_1 - \mu_G \\ \ln x_2 - \mu_L \end{pmatrix}^T \boldsymbol{\Sigma}_{\mathrm{mx}}^{-1} \begin{pmatrix} x_1 - \mu_G \\ \ln x_2 - \mu_L \end{pmatrix}\right\}$$

Where

$$\boldsymbol{\Sigma}_{mx} = \begin{pmatrix} VAR(X_1) & COV(X_1, \ln X_2) \\ COV(X_1, \ln X_2) & VAR(\ln X_2) \end{pmatrix}$$

Note that the variance of the lognormal component is with respect to $\ln X_2$, and that the covariance between the Gaussian and the lognormal random variables is between X_1 and $\ln X_2$.

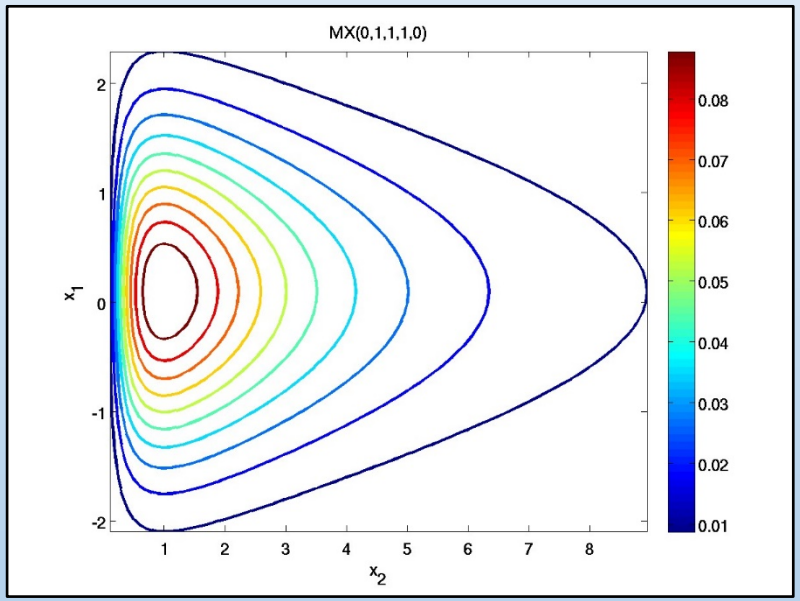
Properties of the Mixed Distribution

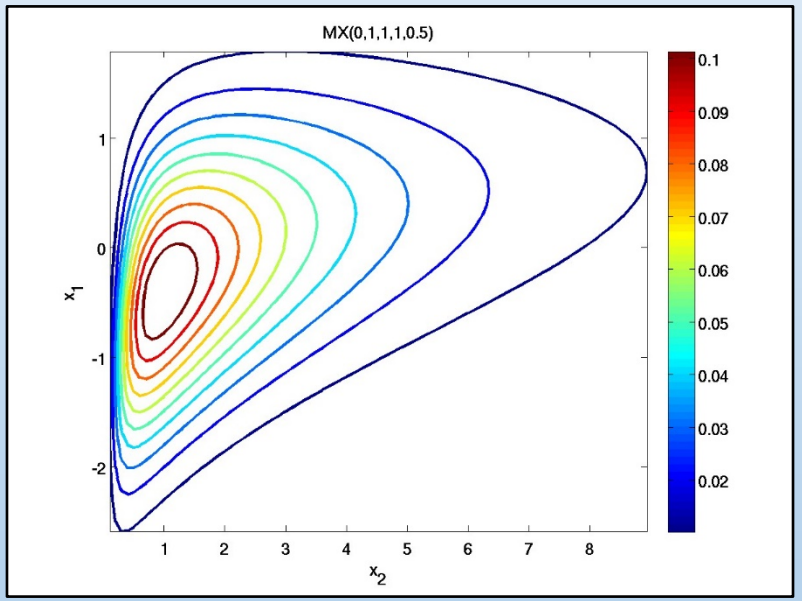
An important property of the mixed distribution is the definitions of the three descriptive statistics. The mean for each component can be found through forming the marginal and joint pdfs which can be shown to be Gaussian and lognormal, or vice-versa. Therefore the mean, mode and median are given by

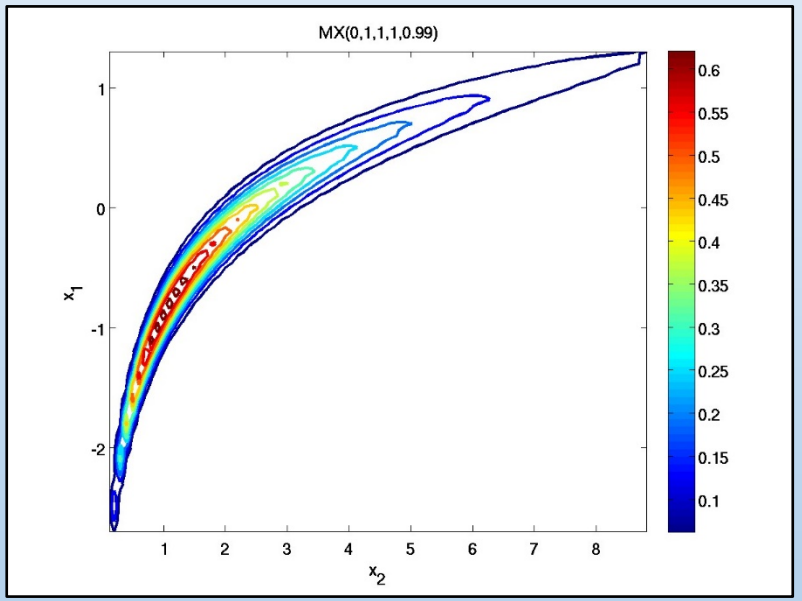
$$mean \equiv \begin{pmatrix} \mu_G \\ \exp\left\{\mu_L + \frac{\sigma_L^2}{2}\right\} \end{pmatrix}, \qquad median \equiv \begin{pmatrix} \mu_G \\ \exp\{\mu_L\} \end{pmatrix},$$

$$mode \equiv \begin{pmatrix} \mu_G - \rho\sigma_G\sigma_L \\ \exp \mu_L - \sigma_L^2 \end{pmatrix}$$

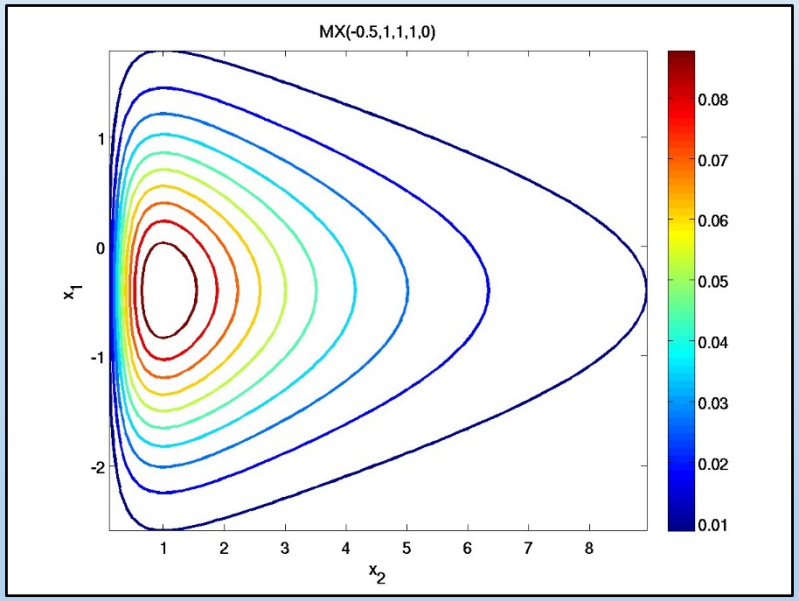
Plots of the Mixed Distribution

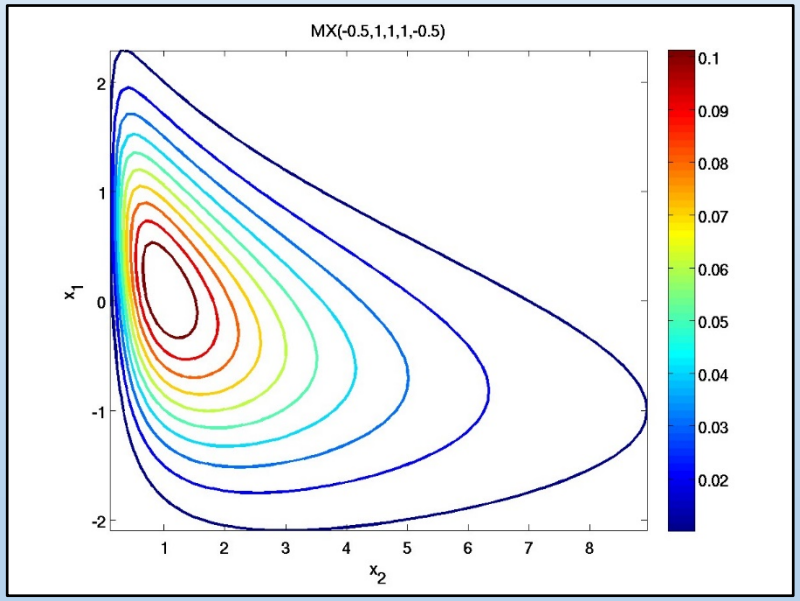


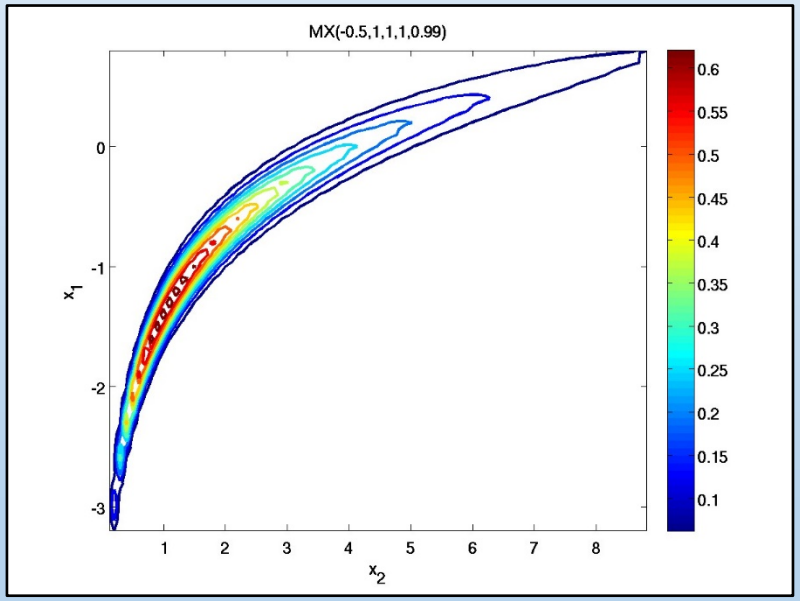




Plots of the Mixed Distribution







Applying the Mixed Distribution to VAR

To be able to apply the mixed distribution to a variational formulation we require the definitions for the errors along with the multivariate version of the mixed distribution. The background and observational errors are given by

$$\boldsymbol{\varepsilon_b} \equiv \begin{pmatrix} \boldsymbol{x^t_{p_1} - x^b_{p_1}} \\ \dfrac{\boldsymbol{x^t_{q_1}}}{\boldsymbol{x^b_{q_1}}} \end{pmatrix}, \qquad \boldsymbol{\varepsilon_o} \equiv \begin{pmatrix} \boldsymbol{y_{p_2} - h_{p_2}(x)} \\ \dfrac{\boldsymbol{y_{q_2}}}{\boldsymbol{h_{q_2}(x)}} \end{pmatrix}$$

Where there are different number of Gaussian and observational background and observational errors, and that $N = p_1 + q_1$ and $N_o = p_2 + q_2$.

Applying the Mixed Distribution to VAR

The multivariate version of the mixed distribution is defined by

$$MX(\boldsymbol{\mu}_{mx}, \boldsymbol{\Sigma}_{\mathrm{mx}}) \equiv \frac{1}{\sqrt{|\boldsymbol{\Sigma}_{mx}|(2\pi)^{\frac{N}{2}}}} \prod_{i=p+1}^{N} \frac{1}{x_i} \exp\left\{-\frac{1}{2}\begin{pmatrix} \boldsymbol{x}_p - \boldsymbol{\mu}_p \\ \ln \boldsymbol{x}_2 - \boldsymbol{\mu}_q \end{pmatrix}^T \boldsymbol{\Sigma}_{\mathrm{mx}}^{-1} \begin{pmatrix} \boldsymbol{x}_p - \boldsymbol{\mu}_p \\ \ln \boldsymbol{x}_2 - \boldsymbol{\mu}_q \end{pmatrix}\right\}$$

Where

$$\boldsymbol{\mu}_{\mathrm{mx}} \equiv \begin{pmatrix} \boldsymbol{\mu}_p \\ \boldsymbol{\mu}_q \end{pmatrix} \quad \boldsymbol{\Sigma}_{\mathrm{mx}} \equiv \begin{pmatrix} \boldsymbol{\Sigma}_{pp} & \boldsymbol{\Sigma}_{pq} \\ \boldsymbol{\Sigma}_{qp} & \boldsymbol{\Sigma}_{\mathrm{qq}} \end{pmatrix}$$

which leads to the mode of the multivariate mixed distribution is given by

$$\boldsymbol{x}_{\boldsymbol{mode}} = \begin{pmatrix} \boldsymbol{\mu}_p - \langle \boldsymbol{\Sigma}_{pq}, \mathbf{1}_q \rangle \\ \exp\{\boldsymbol{\mu}_q - \langle \boldsymbol{\Sigma}_{qq}, \mathbf{1}_q \rangle\} \end{pmatrix}$$

Applying the Mixed Distribution to VAR

If we now follow the standard log-likelihood approach for variational data assimilation through Bayes theorem then we obtain the 3DVAR cost function for the mixed distribution as

$$
\begin{aligned}
&J_{mx}(\boldsymbol{x}^{\boldsymbol{t}}) \\
&= \frac{1}{2}\begin{pmatrix} \boldsymbol{x}_{p_1}^t - \boldsymbol{x}_{p_1}^b \\ \ln \boldsymbol{x}_{q_1}^t - \ln \boldsymbol{x}_{q_1}^b \end{pmatrix}^{\mathrm{T}} \boldsymbol{B}_{\mathrm{mx}}^{-1} \begin{pmatrix} \boldsymbol{x}_{p_1}^t - \boldsymbol{x}_{p_1}^b \\ \ln \boldsymbol{x}_{q_1}^t - \ln \boldsymbol{x}_{q_1}^b \end{pmatrix} + \left\langle \begin{pmatrix} \boldsymbol{x}_{p_1}^t - \boldsymbol{x}_{p_1}^b \\ \ln \boldsymbol{x}_{q_1}^t - \ln \boldsymbol{x}_{q_1}^b \end{pmatrix}, \begin{pmatrix} \boldsymbol{0}_{p_1} \\ \boldsymbol{1}_{q_1} \end{pmatrix} \right\rangle \\
&+ \frac{1}{2}\begin{pmatrix} \boldsymbol{y}_{p_2} - \boldsymbol{h}_{p_2}(\boldsymbol{x}^t) \\ \ln \boldsymbol{y}_{q_2} - \ln \boldsymbol{h}_{q_2}(\boldsymbol{x}^t) \end{pmatrix}^{T} \boldsymbol{R}_{mx}^{-1} \begin{pmatrix} \boldsymbol{y}_{p_2} - \boldsymbol{h}_{p_2}(\boldsymbol{x}^t) \\ \ln \boldsymbol{y}_{q_2} - \ln \boldsymbol{h}_{q_2}(\boldsymbol{x}^t) \end{pmatrix} \\
&+ \left\langle \begin{pmatrix} \boldsymbol{y}_{p_2} - \boldsymbol{h}_{p_2}(\boldsymbol{x}^t) \\ \ln \boldsymbol{y}_{q_2} - \ln \boldsymbol{h}_{q_2}(\boldsymbol{x}^t) \end{pmatrix}, \begin{pmatrix} \boldsymbol{0}_{p_2} \\ \boldsymbol{1}_{q_2} \end{pmatrix} \right\rangle
\end{aligned}
$$

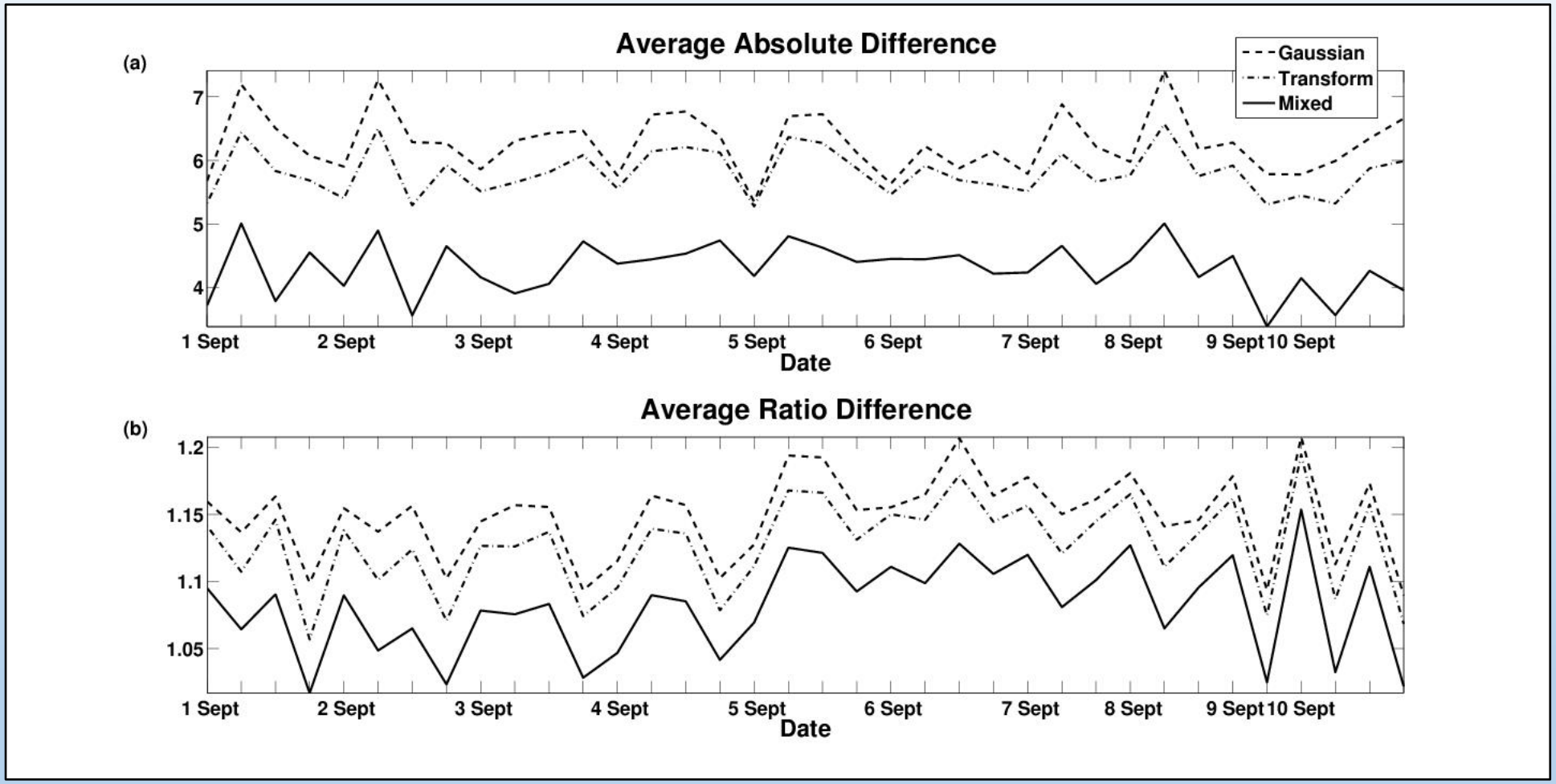
Application of the Mixed Distribution

The CIRA 1-Dimensional Optimal Estimator (C1DOE) is a 1DVAR retrieval system for mixing-ratio and temperature from microwave brightness temperatures. Its original version is a Gaussian fits all formulation.

We have now implemented the mixed distribution approach where we are assuming lognormal errors for the mixing-ratio, and Gaussian for the temperature.

Along with the mixed distribution and Gaussian fits all approaches we have also implemented the logarithmic transform approach for mixing-ratio (Kliewer et al 2016).

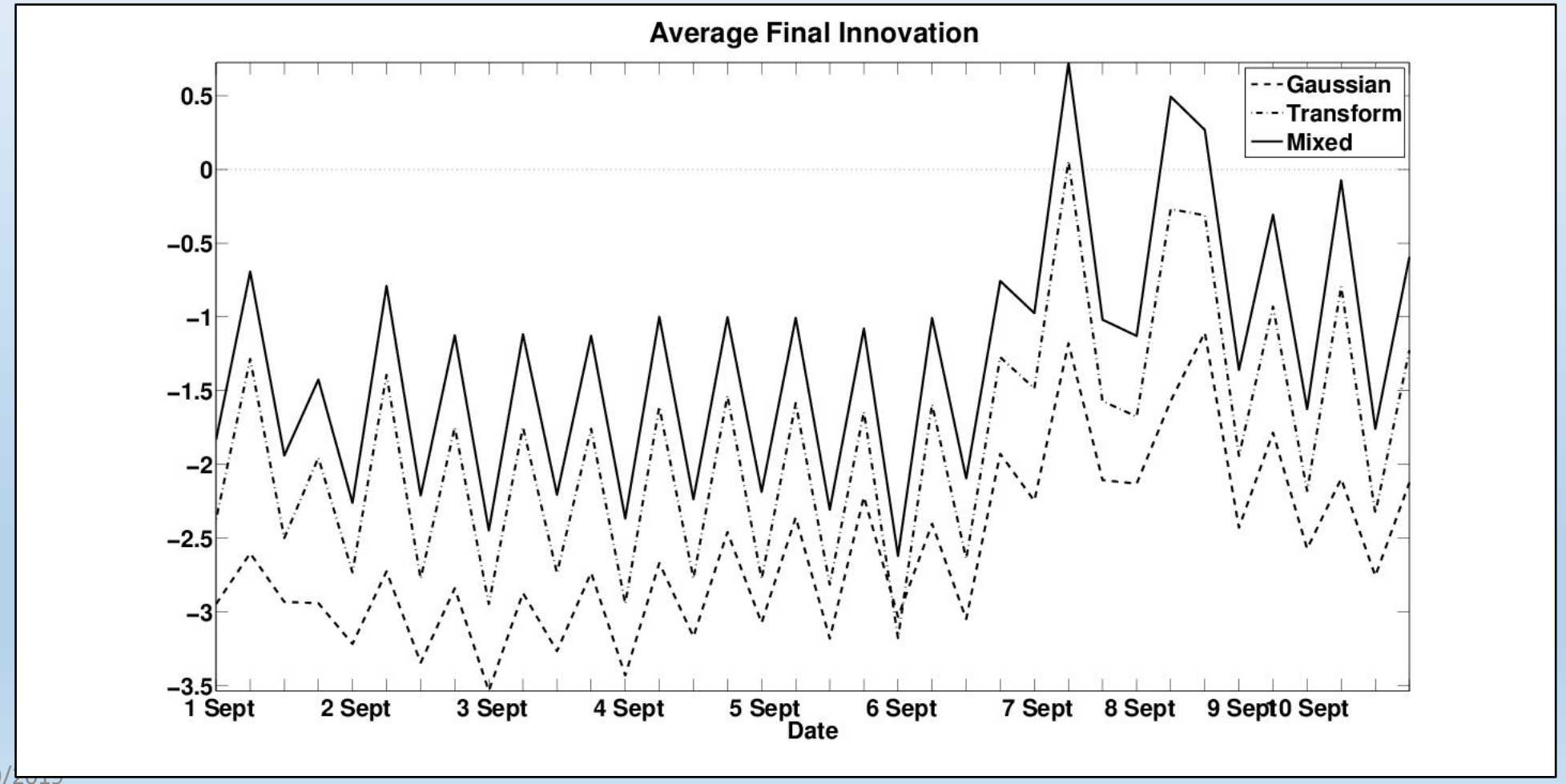
Application of the Mixed Distribution



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.

Application of the Mixed Distribution

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



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 sets of detection algorithms which will assess the if the background state is behaving in a non-Gaussian way, specifically if there is a indication of a lognormal behaviour.

CIRA Data Assimilation Testbed (CDAT)

There are a variety of tests for the Gaussianity assumption of a random variable. Due to their power and formulations the following were chosen:

Shapiro-Wilk: compares sample standard deviation and normal probability plot information to Gaussian distribution

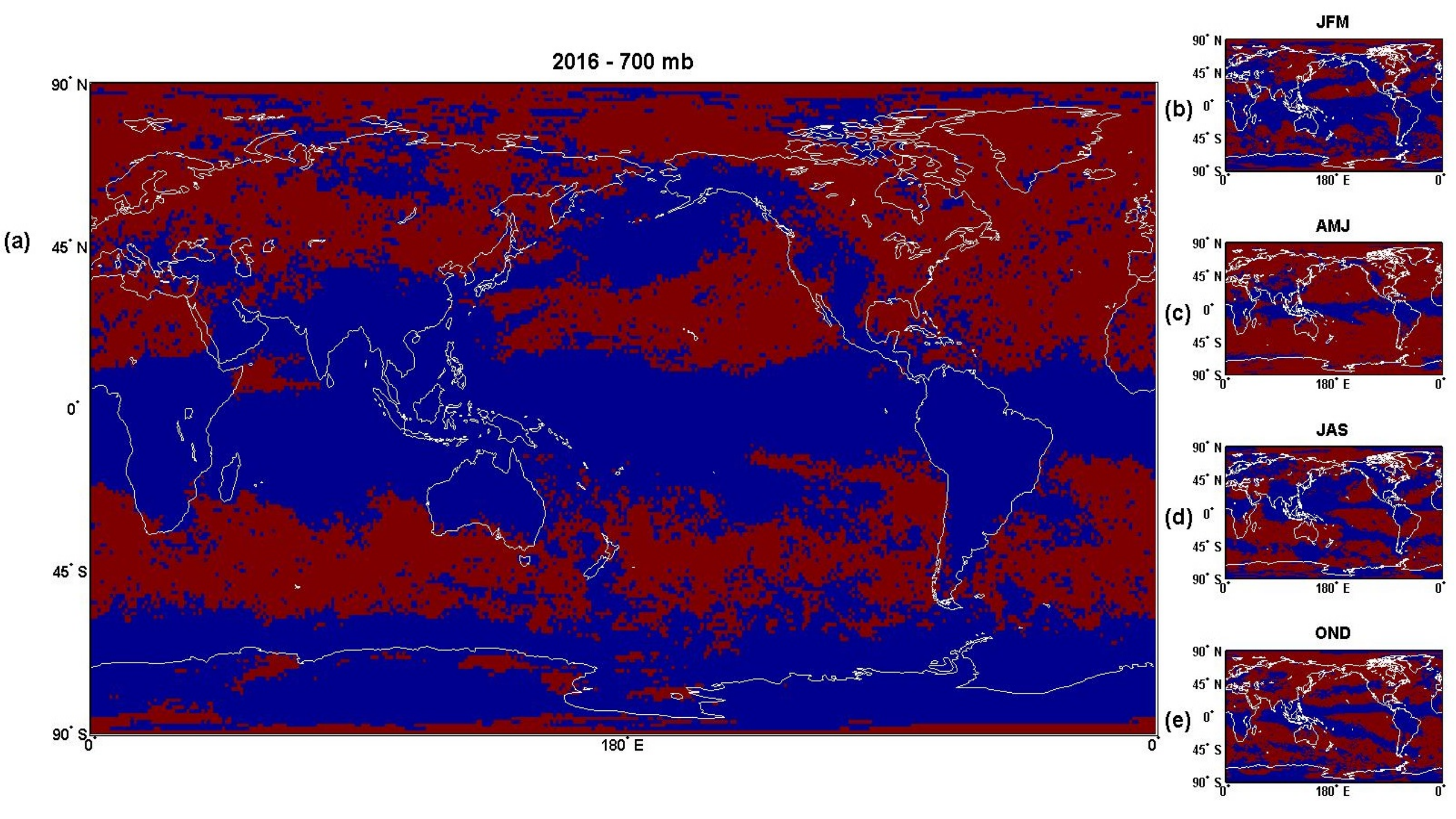
Jarque-Bera: compares skewness and kurtosis (3rd and 4th moments) of the sample to the Gaussian distribution

The variables are analyzed for lognormal characteristics with the following test:

χ^2 Goodness-of-Fit: checks observed vs. expected frequencies for a particular probability distribution, in this case, the lognormal distribution.

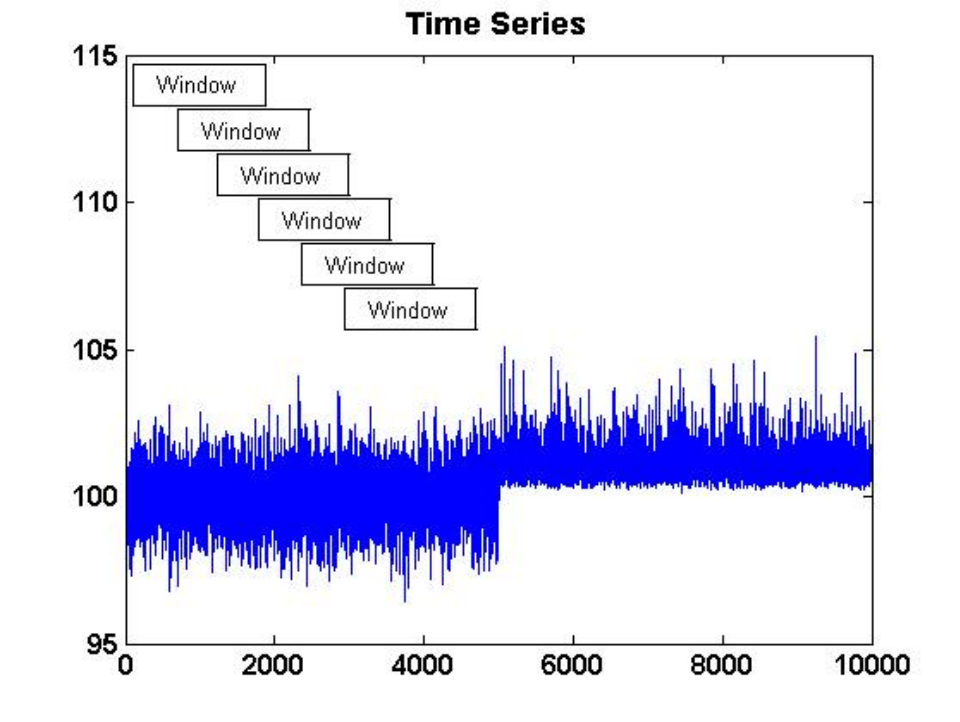
To combine the results of the previous three tests, we define the following:

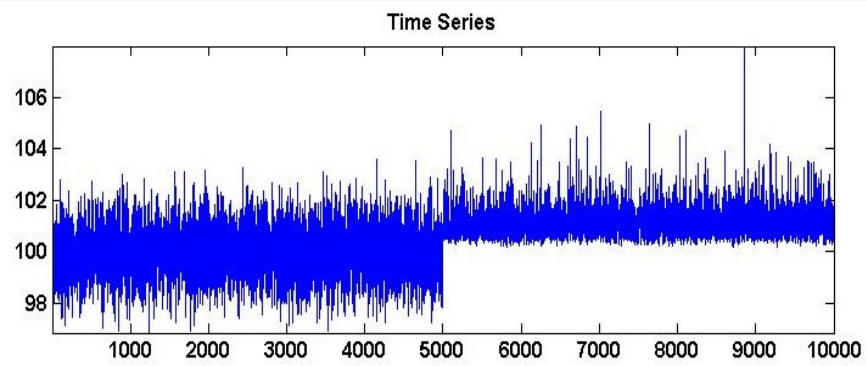
Composite Test: a positive result indicates that the Shapiro-Wilk and Jarque-Bera tests rejected the hypothesis that the data is Gaussian-distributed *and* that the X^2 test failed to reject the hypothesis the data is lognormally-distributed. A negative result indicates one of these conclusions is not met. This test represents agreement of all three tests. All tests performed at significance level **α=0.01**.

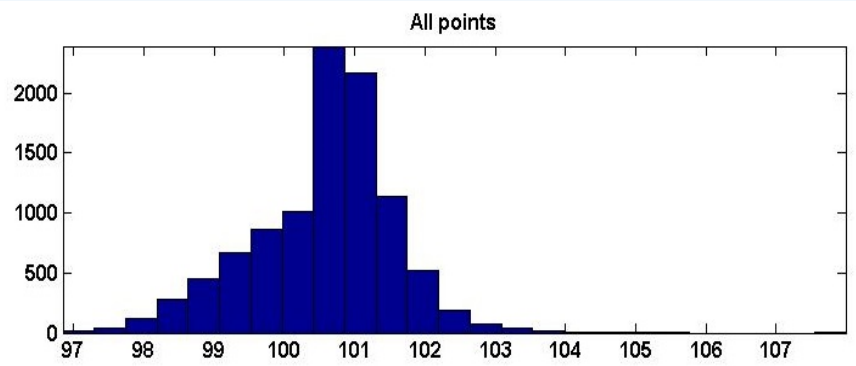

2016 - 700 mb
(a)
90° N
45° N
0°
45° S
90° S
0°
180° E
0°
JFM
(b)
AMJ
(c)
JAS
(d)
OND
(e)

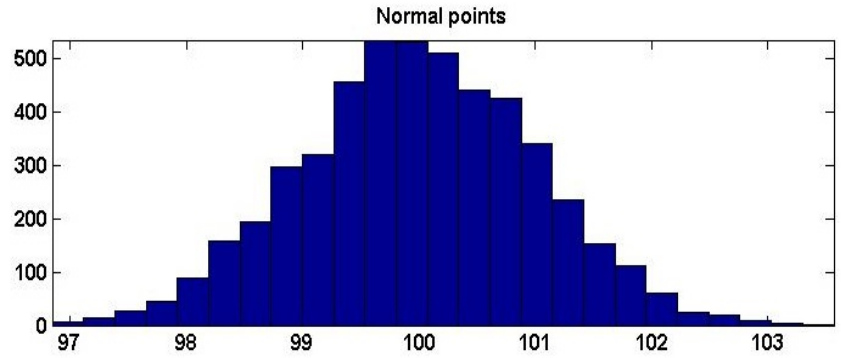
CIRA Data Assimilation Testbed (CDAT)

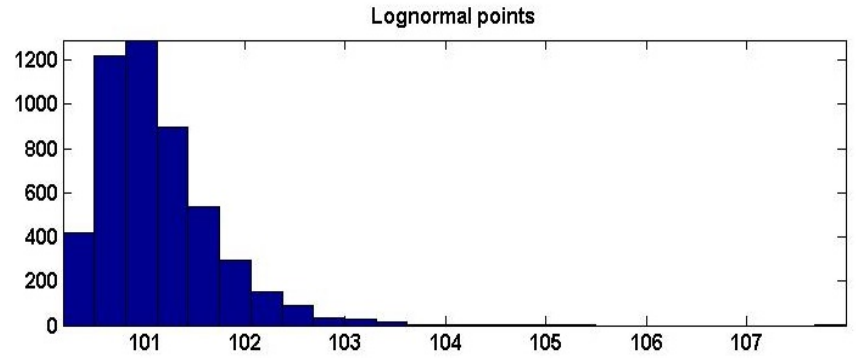
An online detection system has been developed for a time series that shifts from normally-distributed data points to being distributed lognormally. It is intended that this system be implemented in a real-time, online data assimilation system that is constantly receiving new data. A moving window incorporates the newest data points and applies the previously described composite test..

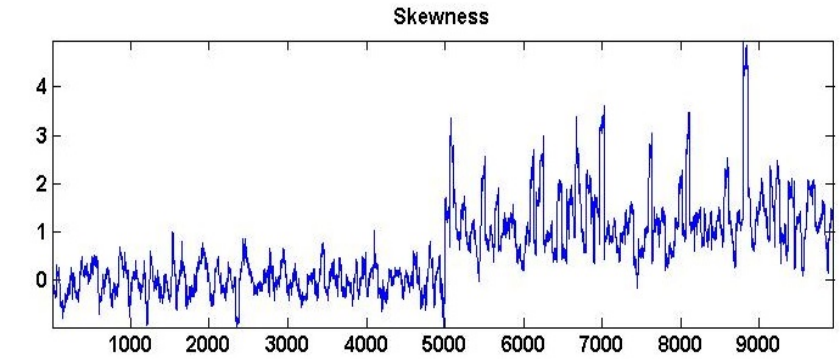


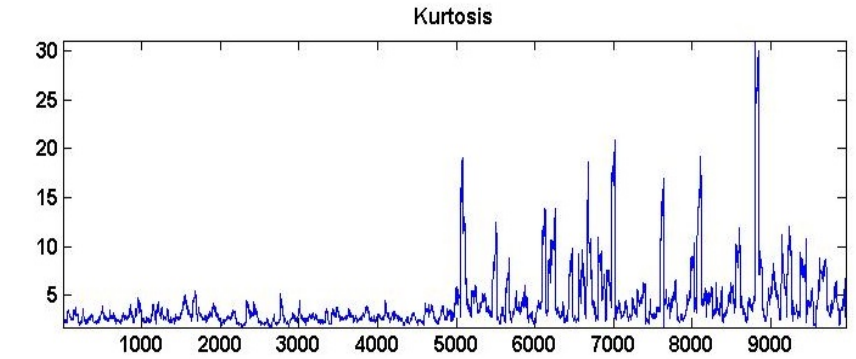




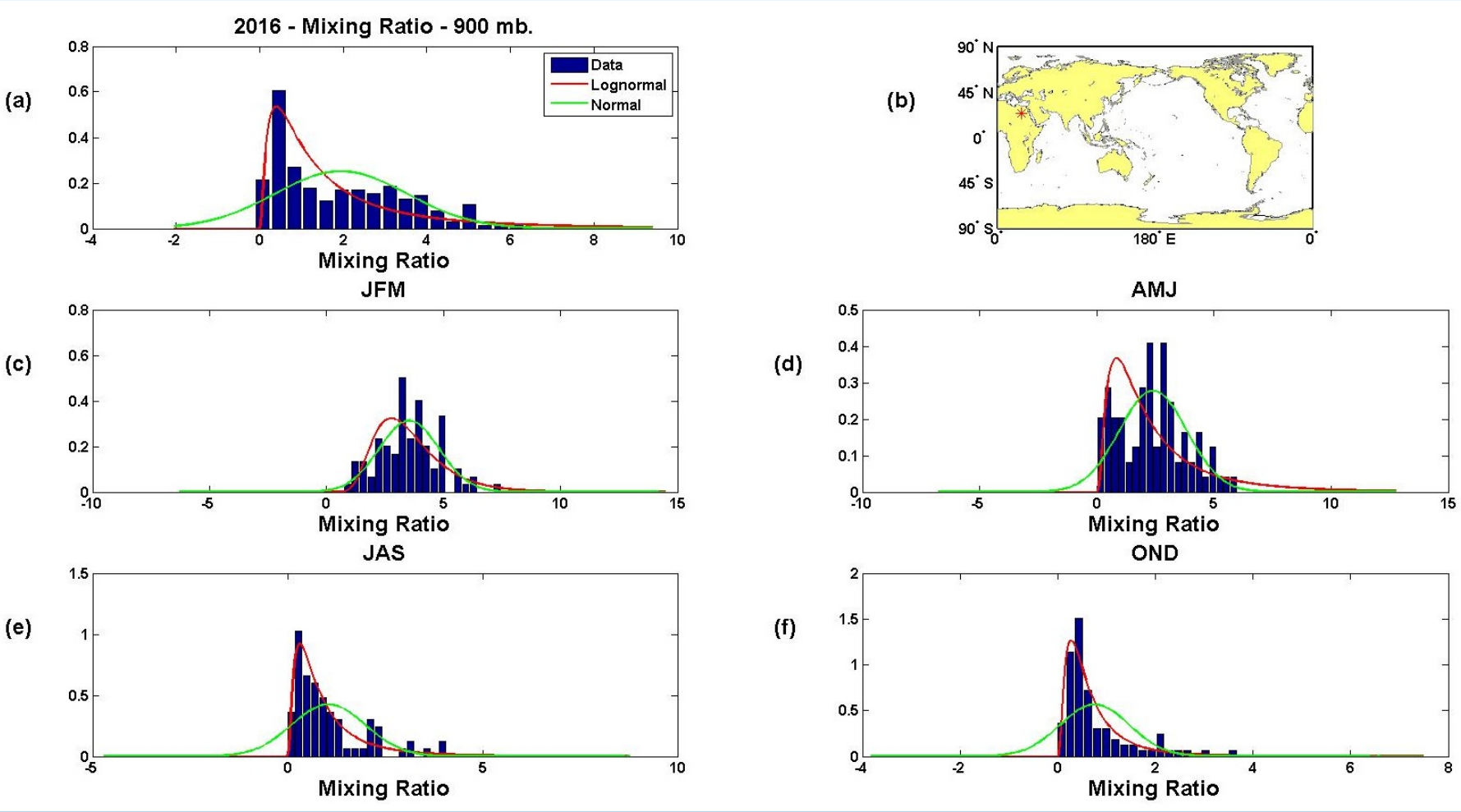




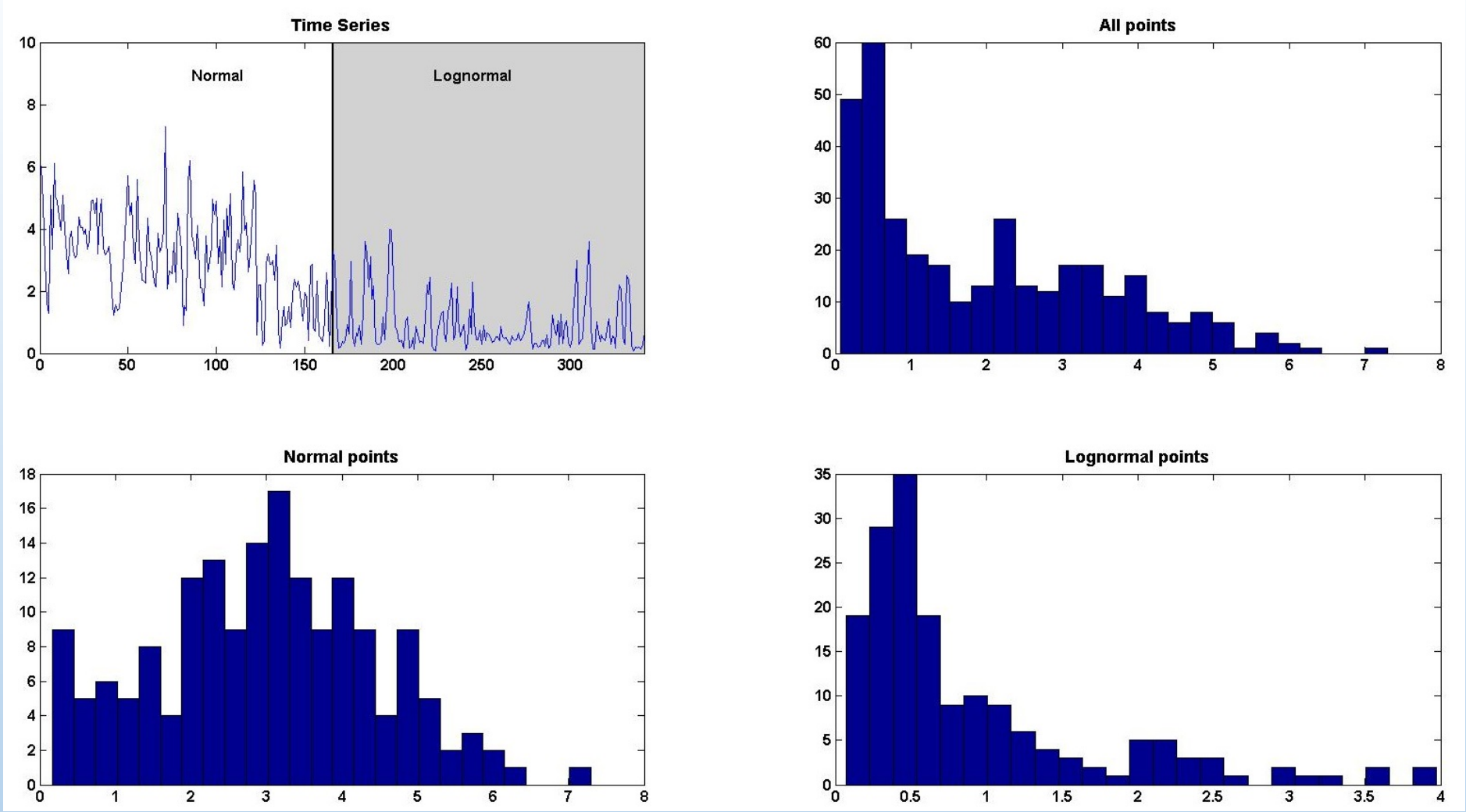




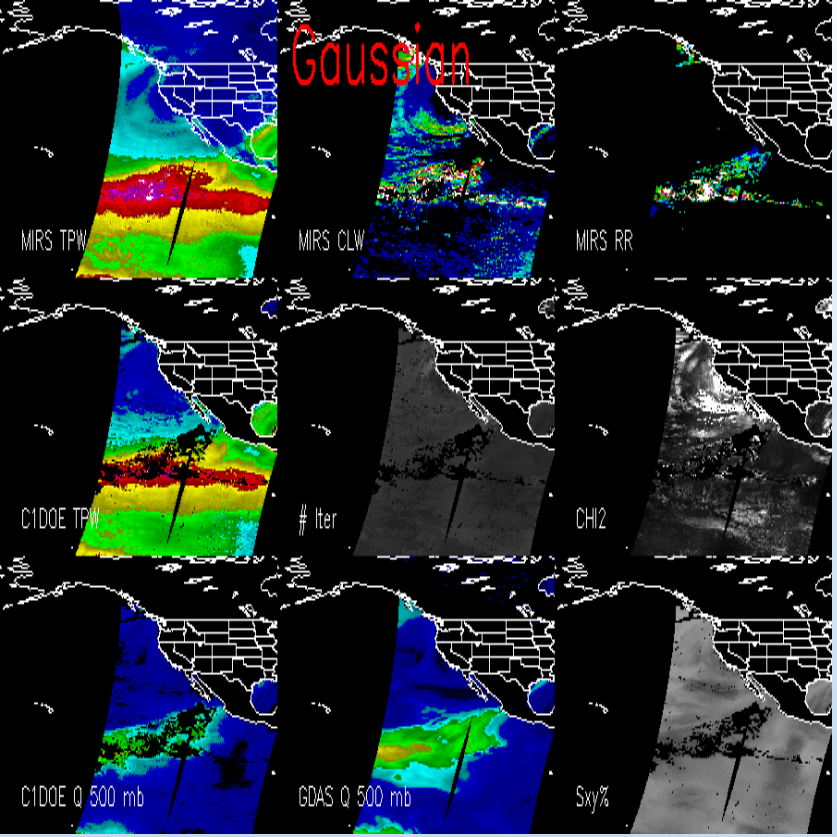
The online test correctly identifies the change point in the time series. Note the change in kurtosis and skewness of the window after the change point.

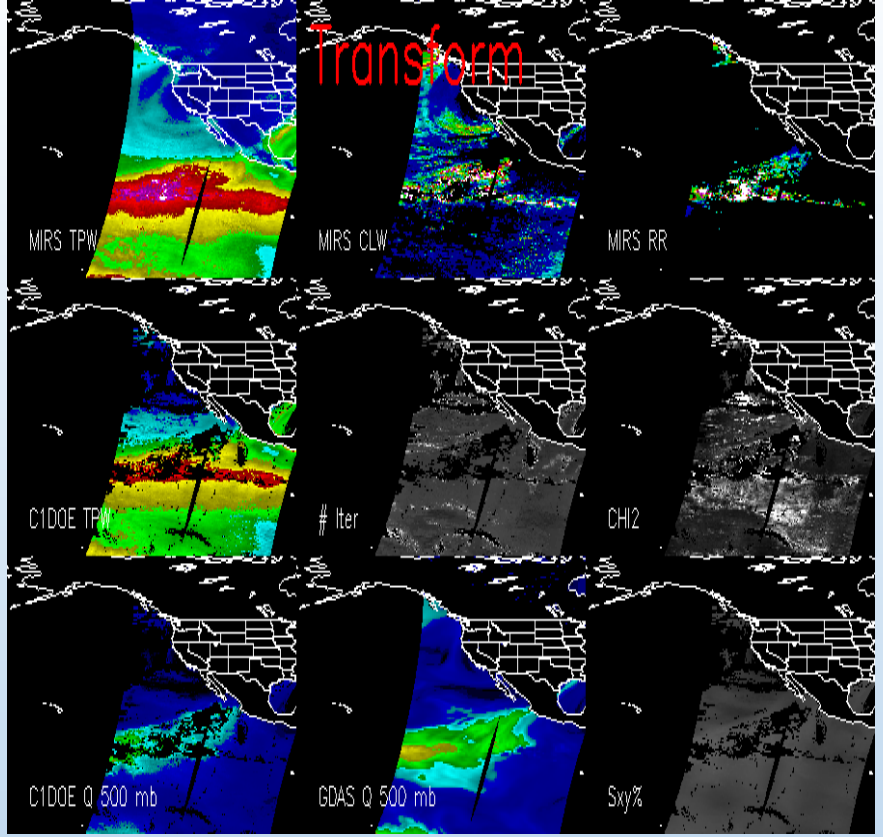


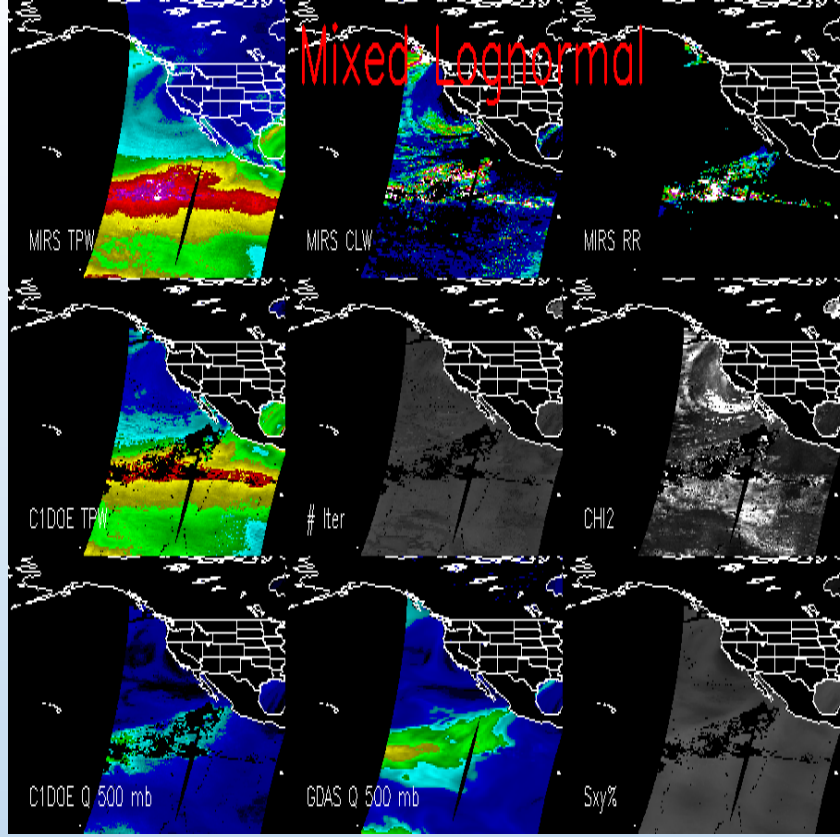
This system is demonstrated to work with NOAA GFS data. For a location in northern Africa, it is seen that for the first two "seasons" (JFM and AMJ) of water vapor mixing ratio are approximately normal with the later two (JAS and OND) following more of a lognormal distribution.



The online detection system determines a change from normal to lognormal data at the vertical line in the upper left panel. This shift corresponds almost directly with the shift in seasons noted in the previous slide. The upper right plot is a histogram of the entire time series, the lower left is of the normal data points (all data before change point), and the lower left is of the lognormal data (all data after change point).







The three cost functions are solved for daily using the observations from the Suomi-National Polar-Orbiting Operational Environmental Satellite System Preparatory Project (SNPP) Advanced Technology Microwave Sounder (ATMS) and are available in near real time currently set up over the west coast. Plots are from 1/1/2019

http://cdat.cira.colostate.edu/C1DOE_Graphics/C1DOE_Main.htm

Conclusions and Further Work

- Have presented a multivariate PDF that is a models the behavior of Gaussian and lognormal random variables simultaneously.
- Have shown that the Gaussian component of the mode of the distribution is a function of the covariances between the Gaussian and lognormal random variables.
- Present initial approaches to detect a change in the distribution of the background state through static statistical approach as well as a time series based approach.
- CDAT is up and running in its early stages with the three versions of the microwave retrievals.

8th International Symposium on Data Assimilation

Canvas Stadium, Colorado State University

Fort Collins, Colorado, USA

8th – 12th June, 2020

Venue: Canvas Stadium, Stadium Club Level