The GIGG-EnKF: Ensemble Kalman Filtering for highly skewed non-negative uncertainty distributions.

Craig H. Bishop[1]

[1]*Marine Meteorology Division, Naval Research Laboratory, Monterey, USA*

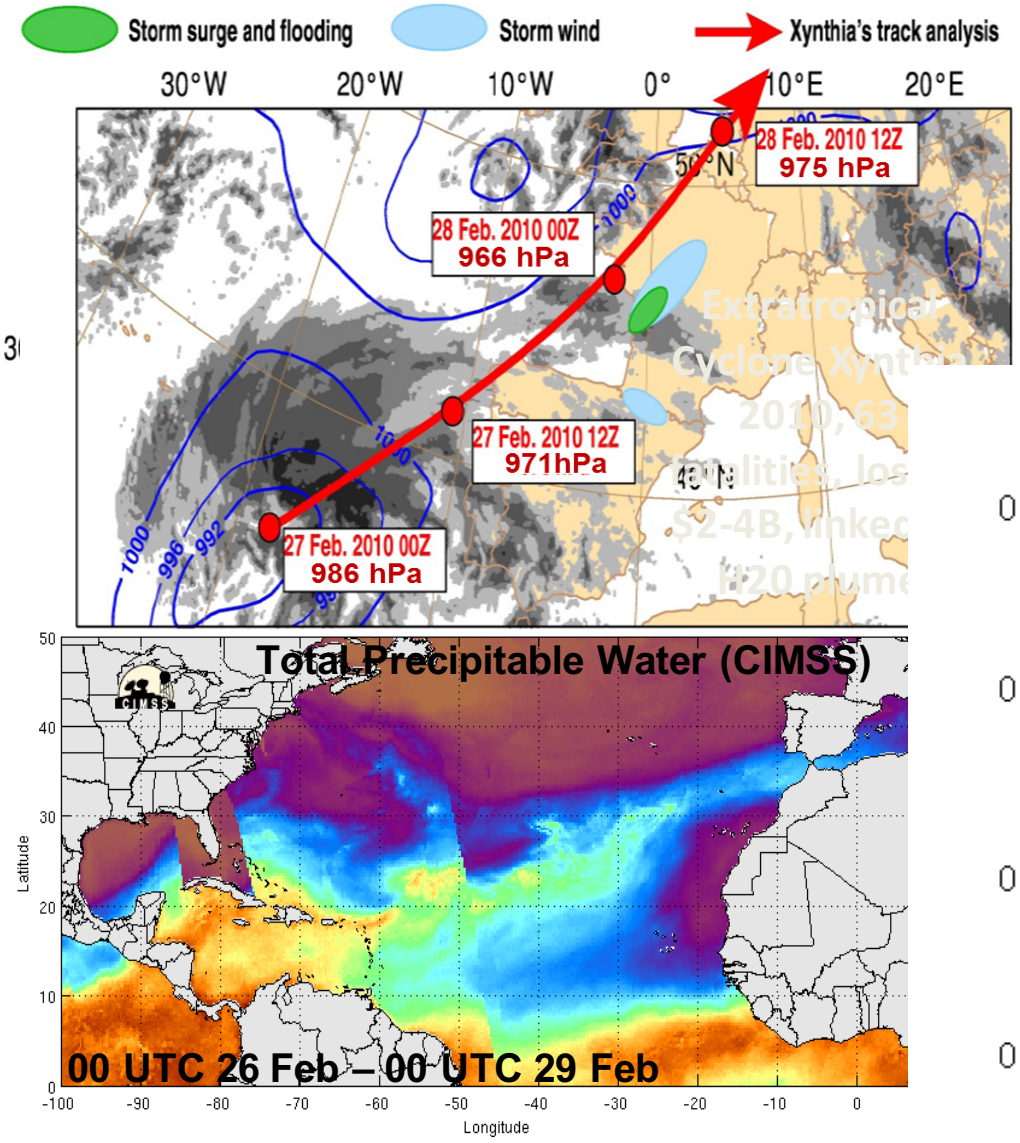
Challenges

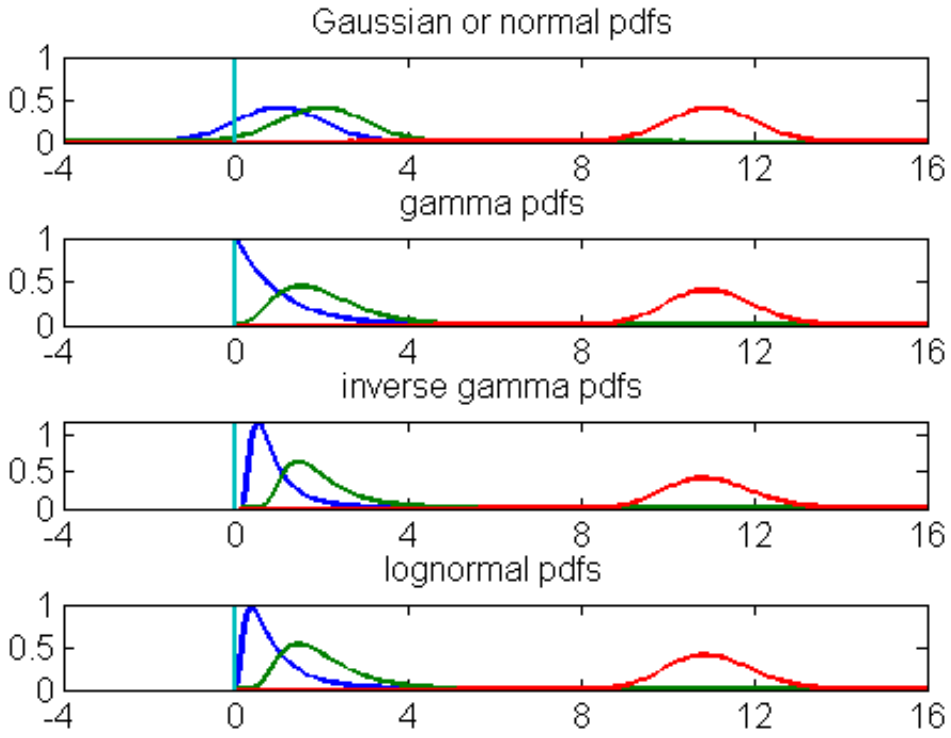
- **How to deal with non-Gaussian bounded variables with skewed uncertainty distributions such as aerosol and water vapor?**
- **How to fix the problem that ensembles of analyzed model variables obtained by EnKFs do not map onto the corresponding ensembles of analyzed observed variables?**
- **Future work: Simple extension to variables like precipitation and cloud whose prior pdfs are best modelled as sum of delta function at zero plus gamma pdfs.**

Outline

- The Gamma Inverse-Gamma and Gaussian (GIGG) variation on the EnKF. (In this talk and on-line in QJRMS)
- The benefit of using tools from 4DVAR to improve the mapping from obs space to model space in EnKFs
- Future work: The remarkable gamma function based Dirac delta function and its potential use in ensemble data assimilation of clouds and precipitation.

The challenge of non-Gaussian variables





Why not take log of ob and fcst?

- Taking the log of an unbiased observation creates a biased observation.

  If the pdf of observations y^o given the true value y is given by

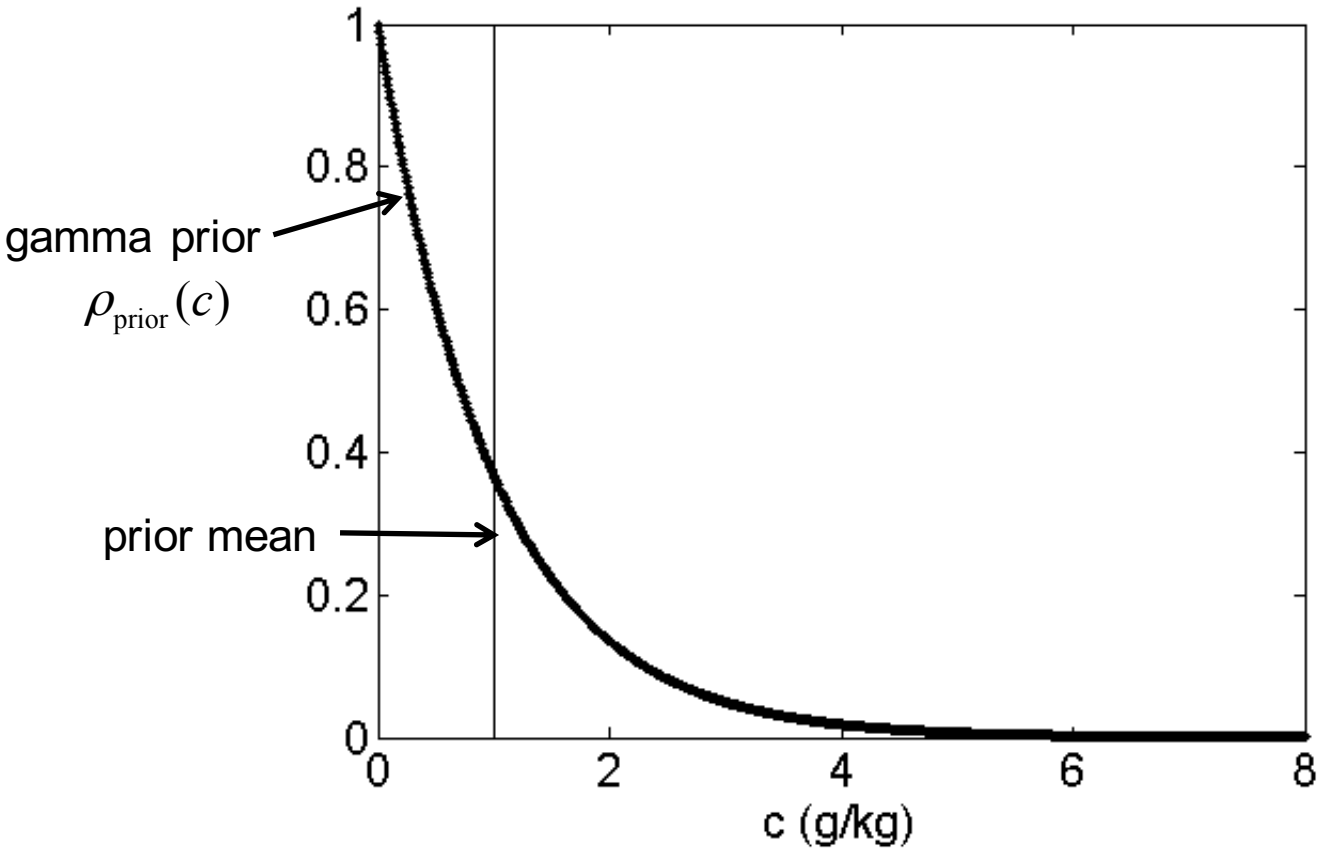
$$L\left(y^o \mid y\right) = \frac{1}{y^o \sigma \sqrt{2\pi}} \exp\left[- \frac{\left[\ln\left(y^o\right) - \ln\left(y\right)\right]^2}{2\sigma^2} \right] \text{ then}$$

$$\left\langle \ln\left(y^o\right) \right\rangle = \ln\left(y\right) \text{(unbiased), but then } \left\langle y^o \right\rangle = y \exp\left(\frac{\sigma^2}{2} \right) \text{(biased).}$$

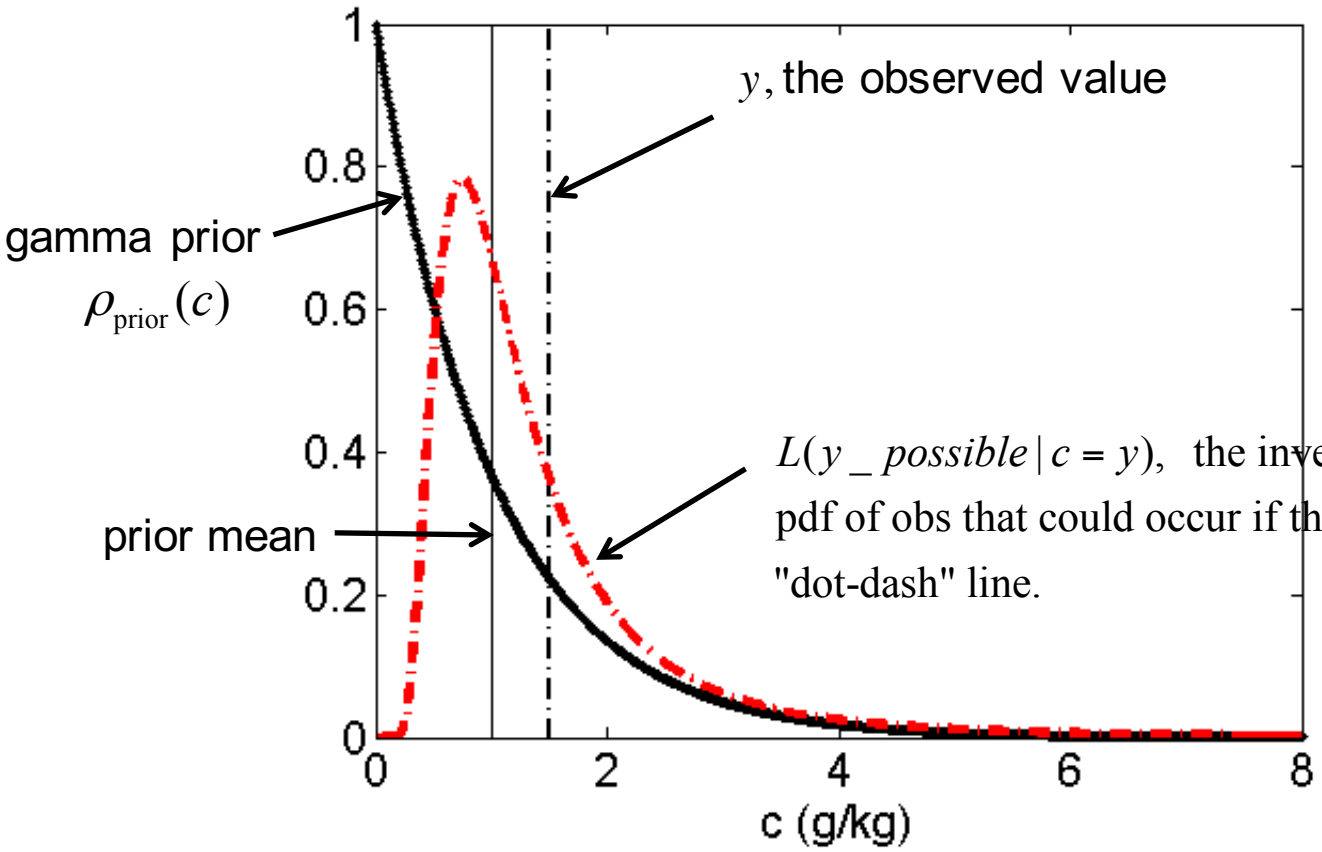
  Conversely, if $\left\langle y^o \right\rangle = y$ (unbiased) then $\left\langle \ln\left(y^o\right) \right\rangle = f\left(y, \sigma^2\right) \neq \ln\left(y\right)$ (biased).

- The log-normal distribution is a very poor approximation for variables that can be very close to zero or zero – like precipitation.
- In contrast, the gamma pdf is excellent for such variables – it even has a Dirac delta funtion form which can accommodate finite probabilities of zero.

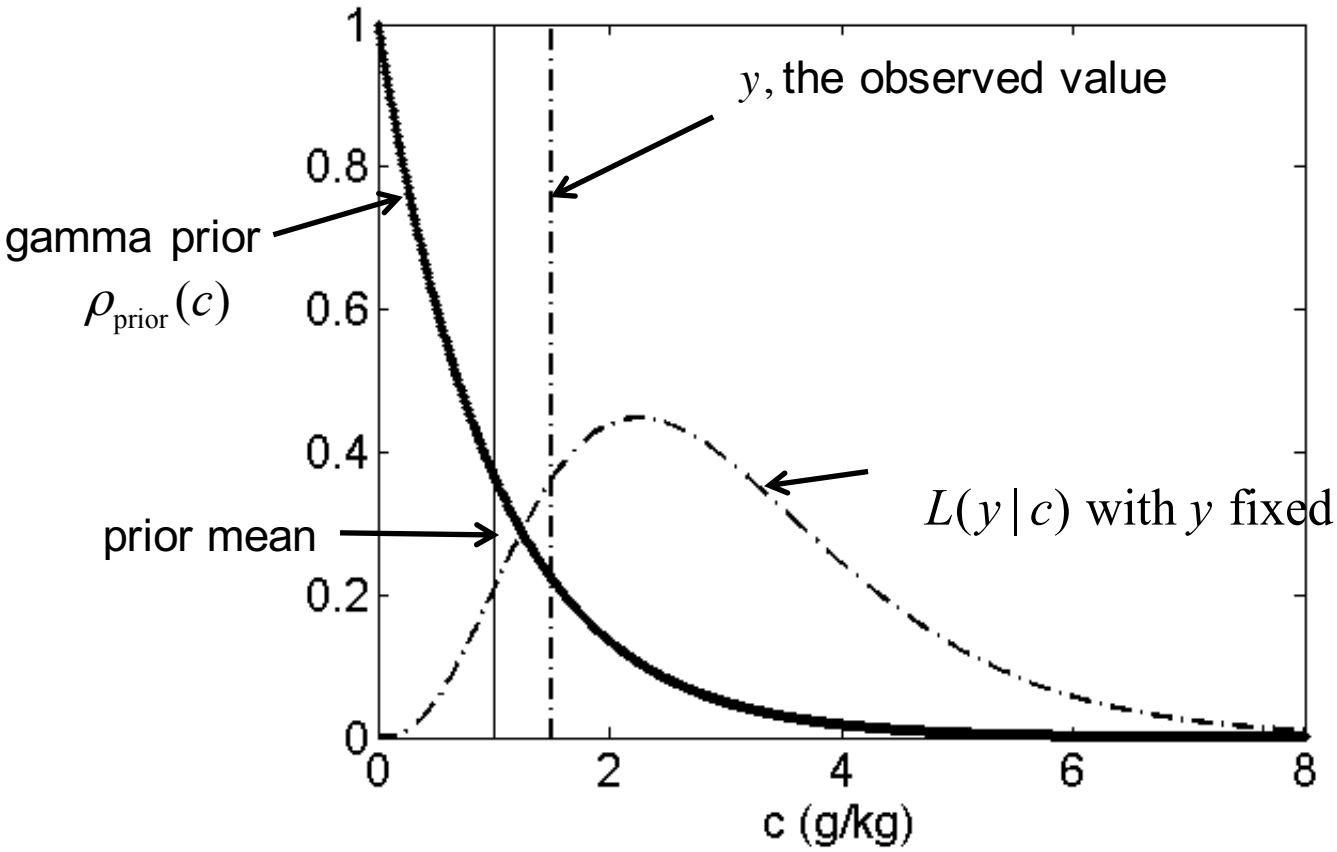
A prior Gamma distribution



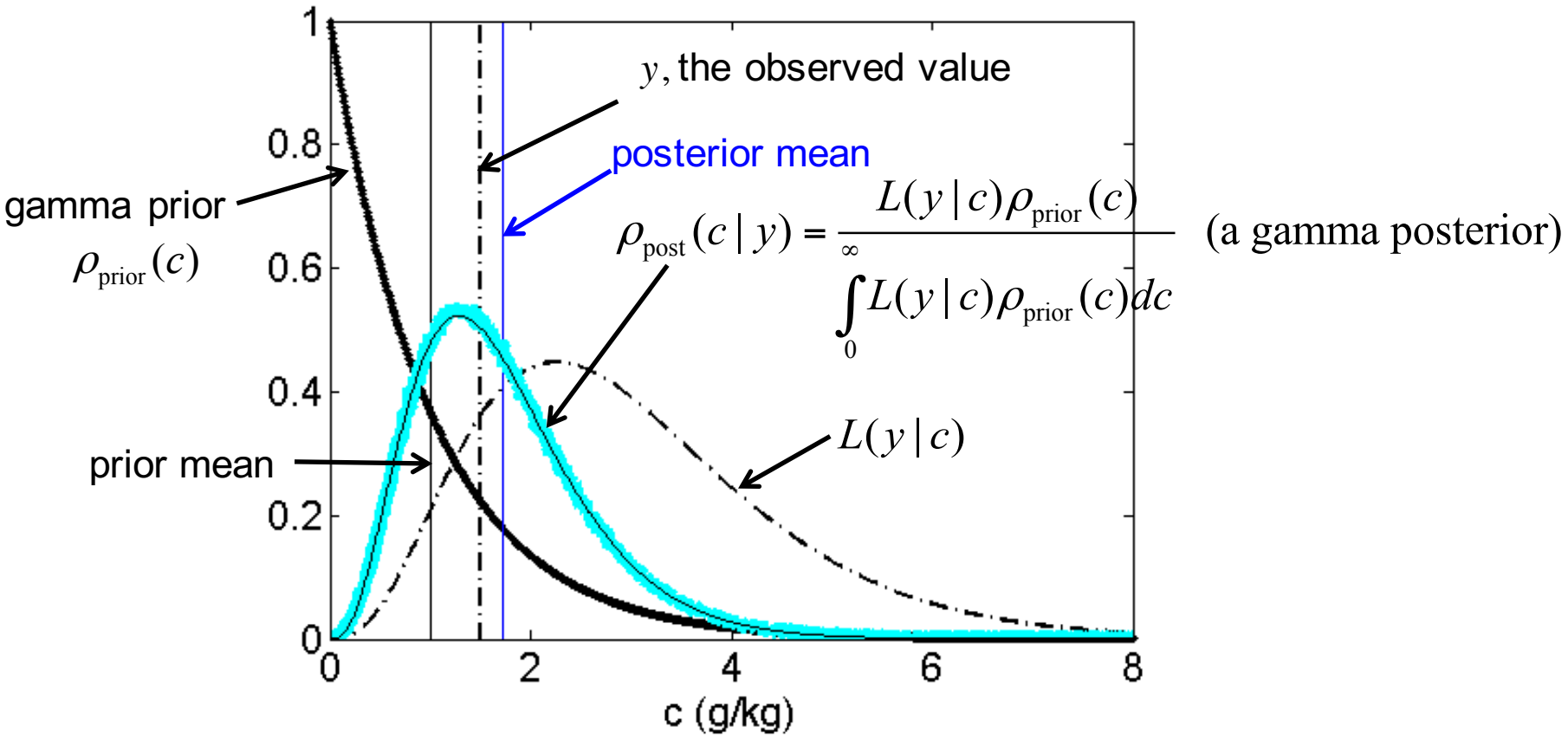
Inverse-Gamma pdf of obs given truth



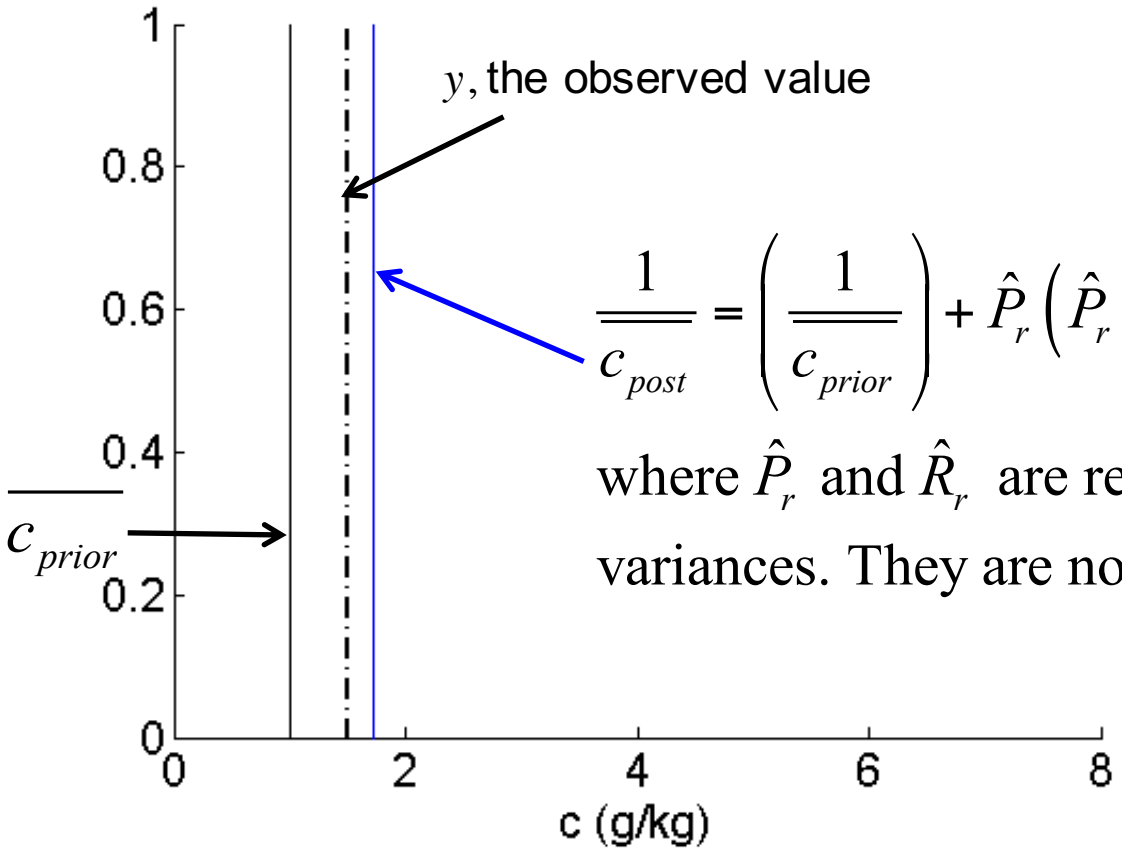
The likelihood function



The posterior pdf is then a gamma



Equation for posterior mean

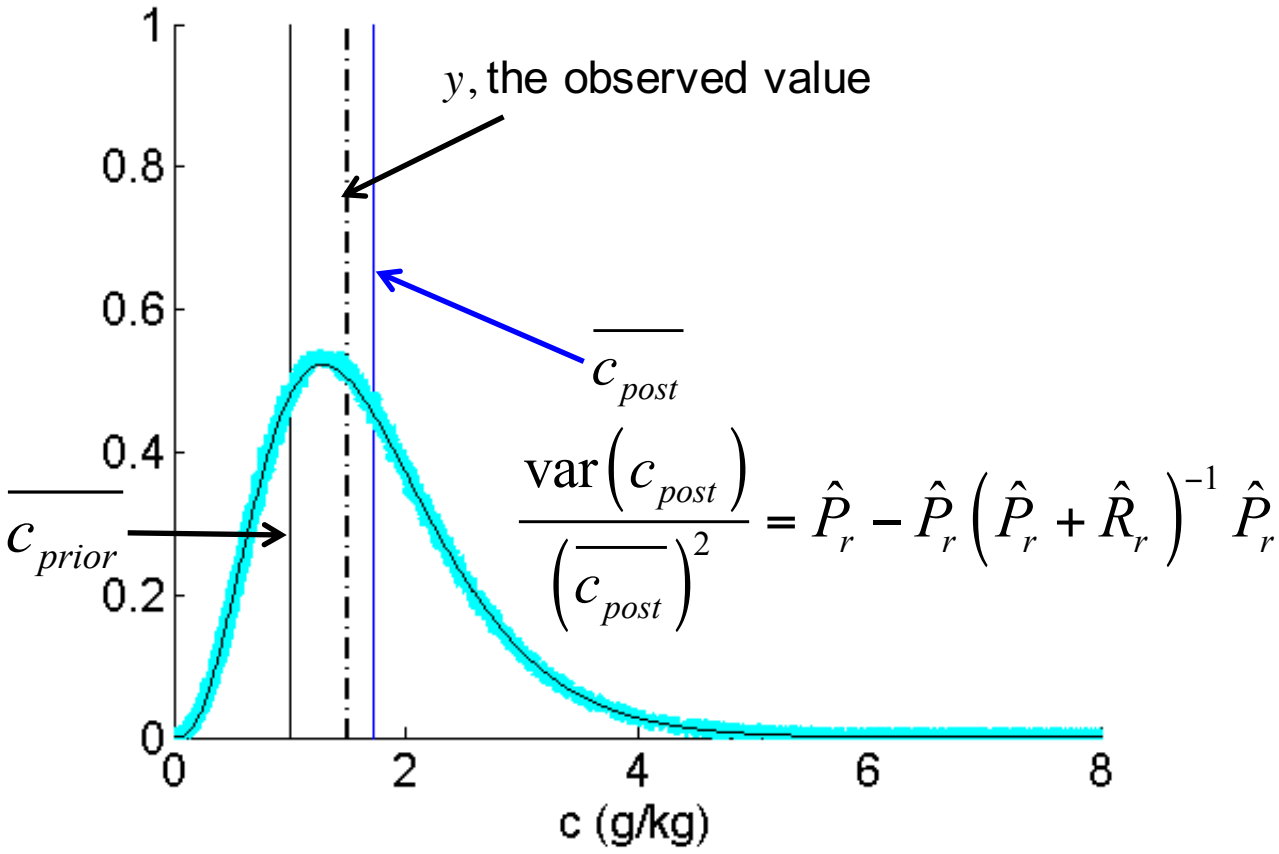


$$\frac{1}{c_{post}} = \left(\frac{1}{c_{prior}}\right) + \hat{P}_r\left(\hat{P}_r + \hat{R}_r\right)^{-1}\left[\frac{1}{y} - \left(1 + \hat{R}_r\right)\left(\frac{1}{c_{prior}}\right)\right]$$

where $\hat{P}_r$ and $\hat{R}_r$ are relative prior and likelihood variances. They are non-dimensional.

Posterior mean equation has Kalman like gain but everything else is inverted !

Eq for relative variance of posterior



Eq for relative variance of posterior is almost identical to Kalman filter

Perturbed obs GIGG-EnKF (GIG case)

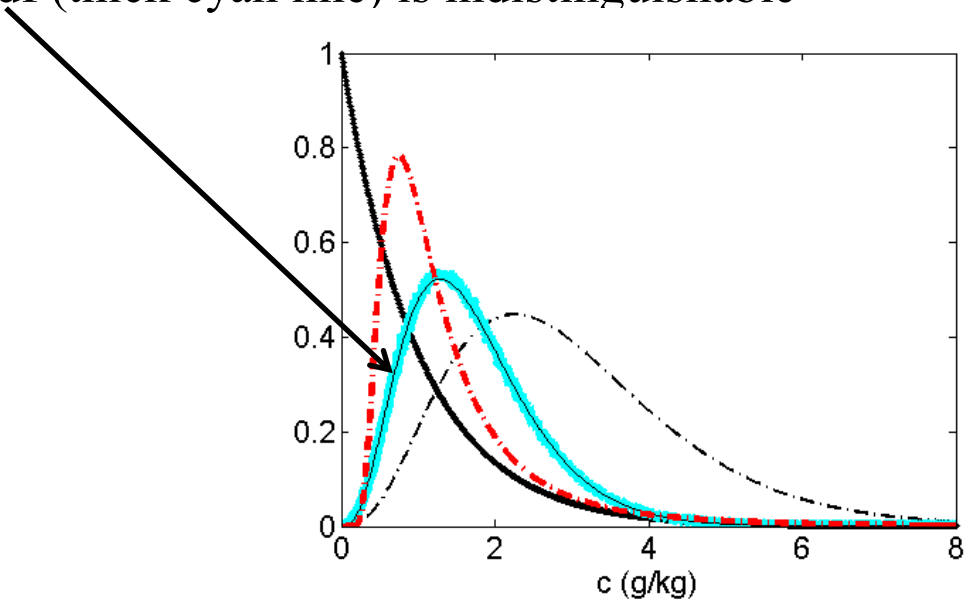
Drawing obs from gamma pdf with $L(y\,|\,c)$ (dot-dash line) shape and using them in

$$\frac{c_i^a - \overline{c_{post}}}{\overline{c_{post}}} = \frac{\left(c_i^f - \overline{c_{prior}}\right)}{\sqrt{\left(\overline{c_{prior}}\right)^2 + \mathrm{var}\left(c_{prior}\right)}} + \hat{P}_r\left(\hat{P}_r + R_r\right)^{-1}\left[\frac{\left(c_i^L - \left\langle c^L\right\rangle\right)}{\sqrt{\left\langle c^L\right\rangle^2 - 2\,\mathrm{var}\left(c^L\right)}} - \frac{\left(c_i^f - \overline{c_{prior}}\right)}{\sqrt{\left(\overline{c_{prior}}\right)^2 + \mathrm{var}\left(c_{prior}\right)}}\right]$$

where

$$\frac{1}{\overline{c_{post}}} = \left(\frac{1}{\overline{c_{prior}}}\right) + \hat{P}_r\left(\hat{P}_r + \hat{R}_r\right)^{-1}\left[\frac{1}{y} - \left(1 + \hat{R}_r\right)\left(\frac{1}{\overline{c_{prior}}}\right)\right]$$

successfully generates random sample whose pdf (thick cyan line) is indistinguishable from true pdf (thin black line).



Gamma Inverse-Gamma and Gaussian (GIGG) EnKF, fits seamlessly in DART

for $j = 1:p$; % where p is the number of observations

Step 1: Decide whether forecast and observation uncertainty associated with y_j^o is best approximated by GIG, IGG or Gaussian assumptions.

Step 2: if (GIG) then use (6), (7) and (12) to obtain $y_{ji}^a,\ i = 1,2,...,K$;

else if (IGG) then use (23), (24) and (30) to obtain $y_{ji}^a,\ i = 1,2,...,K$;

else if (Gaussian) then use (34) to obtain $y_{ji}^a,\ i = 1,2,...,K$;

Step 3: Find corresponding analysis ensemble for observations and model variables

$$y_{ki}^a = y_{ki}^f + \frac{\text{covar}\left(y_k^f, y_j^f\right)}{\text{var}\left(y_j^f\right)}\left(y_{ji}^a - y_{ji}^f\right), \text{ for } k = 1,2,...,p;\ i = 1,2,...,K$$

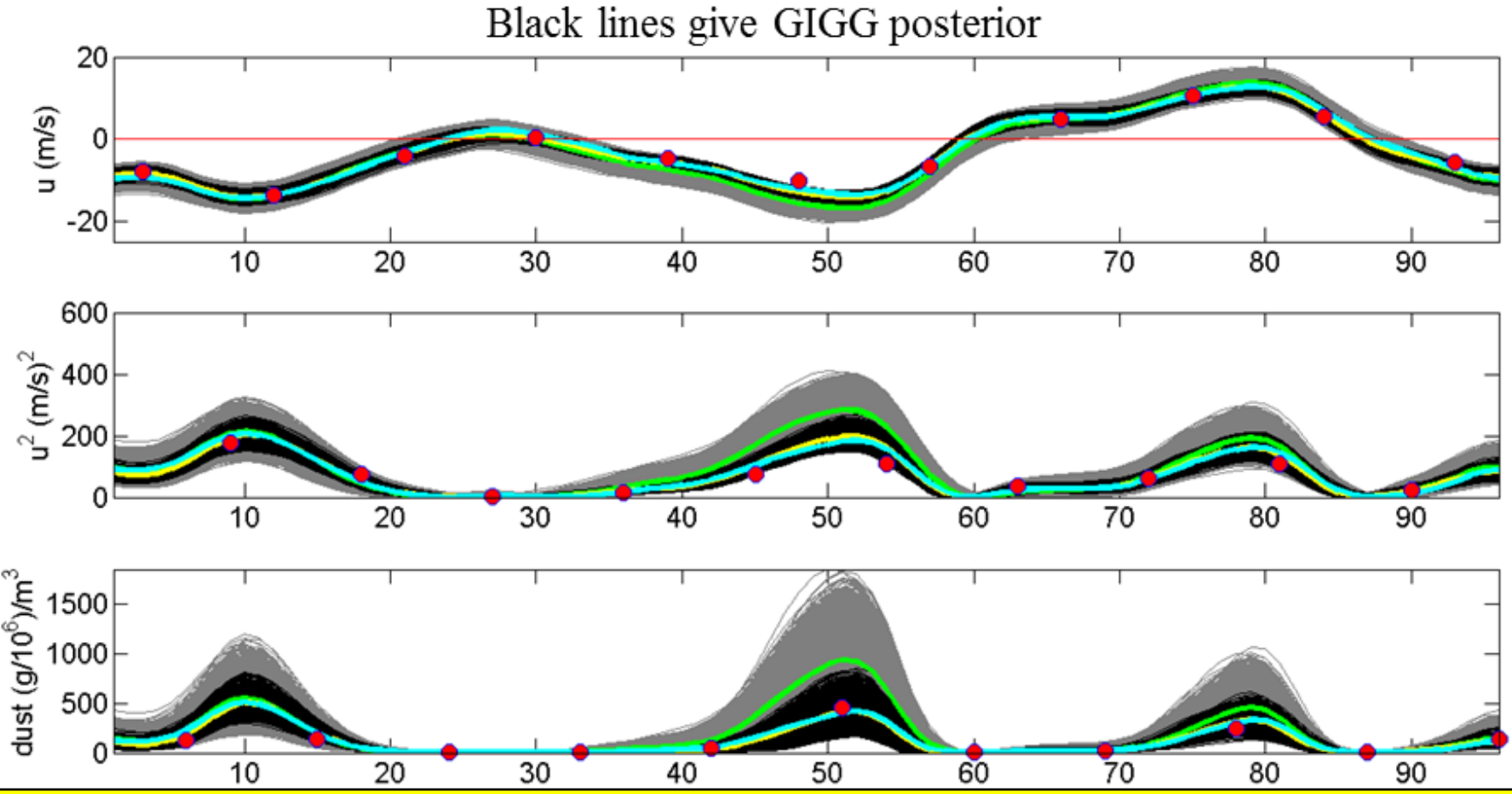
$$x_{\mu i}^a = x_{\mu i}^f + \frac{\text{covar}\left(x_\mu^f, y_j^f\right)}{\text{var}\left(y_j^f\right)}\left(y_{ji}^a - y_{ji}^f\right), \text{ for } \mu = 1,2,...,n;\ i = 1,2,...,K$$

Step 4: Let the analysis ensemble be the prior ensemble for the next observation

$y_{ki}^f = y_{ki}^a$, for $k = 1,2,...,p;\ i = 1,2,...,K$

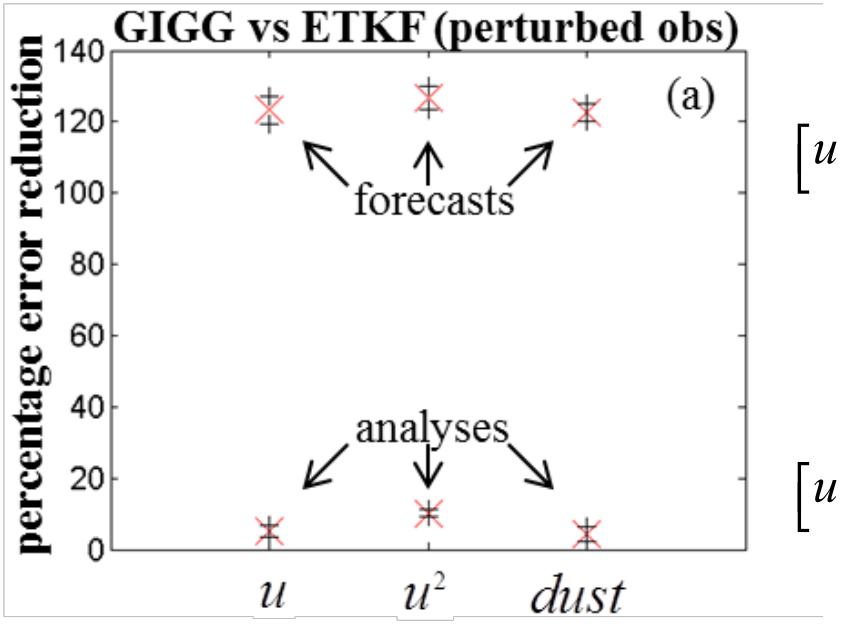
$x_{\mu i}^f = x_{\mu i}^a$, for $\mu = 1,2,...,n;\ i = 1,2,...,K$

end



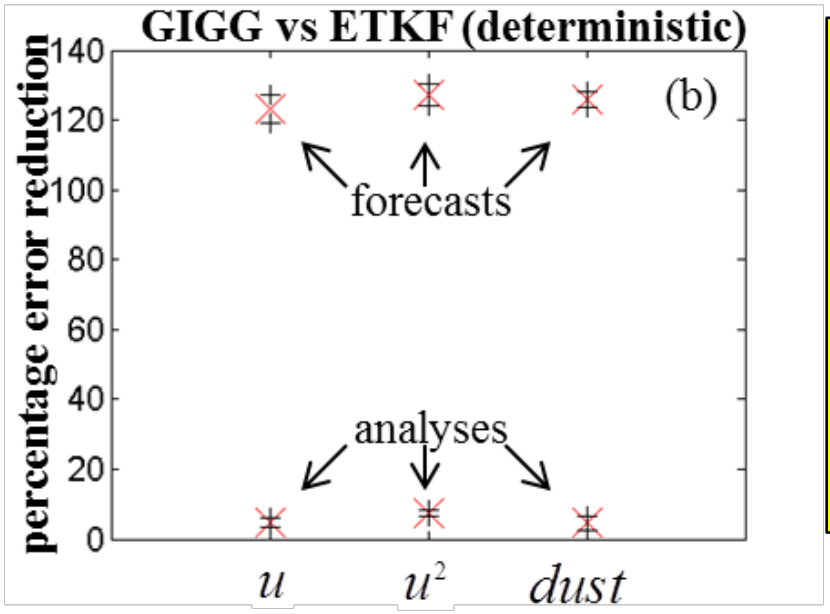
Abscissa gives the index of the grid point on the periodic model domain. Ordinate gives the value of the model variable. Grey lines depict the 10000 members of the prior ensemble forecast. Thick green line gives the prior forecast ensemble mean. Cyan line depicts the truth. Red dots indicate the error prone observations of the truth. Black lines depict the posterior analysis ensemble obtained from the GIGG filter. Yellow line depicts the posterior analysis ensemble mean. (The yellow line is largely obscured by the truth (cyan line)).

GIGG-EnKF outperforms EnKF/ETKF



$$\left[u^f\right]_{mse}^{rel} = \frac{1}{n}\sum_{i=1}^{96}\left[\frac{\left(\overline{u_i^f} - \left(u_i^f\right)^t\right)}{\frac{1}{2}\left\{\overline{u_i^f} + \left(u_i^f\right)^t\right\}}\right]^2$$

$$\left[u^f\right]_{per} = 100\left\{\frac{\left\langle\left[u^f\right]_{mse}^{rel}(ETKF)\right\rangle - \left\langle\left[u^f\right]_{mse}^{rel}(GIGG)\right\rangle}{\frac{1}{2}\left[\left\langle\left[u^f\right]_{mse}^{rel}(ETKF)\right\rangle + \left\langle\left[u^f\right]_{mse}^{rel}(GIGG)\right\rangle\right]}\right\}$$



Statistically significant improvement in all variables.

Massive improvement in mse of ensemble mean of non-linear forecasts.

Gamma Inverse-Gamma and Gaussian (GIGG) EnKF, fits seamlessly in DART

for $j = 1:p$; % where p is the number of observations

Step 1: Decide whether forecast and observation uncertainty associated with y_j^o is best approximated by GIG, IGG or Gaussian assumptions.

Step 2: if (GIG) then use (6), (7) and (12) to obtain $y_{ji}^a,\ i = 1,2,...,K$;

else if (IGG) then use (23), (24) and (30) to obtain $y_{ji}^a,\ i = 1,2,...,K$;

else if (Gaussian) then use (34) to obtain $y_{ji}^a,\ i = 1,2,...,K$;

Step 3: Find corresponding analysis ensemble for observations and model variables

$$y_{ki}^a = y_{ki}^f + \frac{\text{covar}\left(y_k^f, y_j^f\right)}{\text{var}\left(y_j^f\right)}\left(y_{ji}^a - y_{ji}^f\right), \text{ for } k = 1,2,...,p;\ i = 1,2,...,K$$

$$x_{\mu i}^a = x_{\mu i}^f + \frac{\text{covar}\left(x_\mu^f, y_j^f\right)}{\text{var}\left(y_j^f\right)}\left(y_{ji}^a - y_{ji}^f\right), \text{ for } \mu = 1,2,...,n;\ i = 1,2,...,K$$

Can do better than linear regression using tools from 4DVAR.

Step 4: Let the analysis ensemble be the prior ensemble for the next observation

$y_{ki}^f = y_{ki}^a$, for $k = 1,2,...,p;\ i = 1,2,...,K$

$x_{\mu i}^f = x_{\mu i}^a$, for $\mu = 1,2,...,n;\ i = 1,2,...,K$

end

Using tools from 4DVAR to improve GIGG-ENKF analysis

If the state is being estimated at the beginning of the DA window and/or the observation operator H is non-linear then $H\left(M\left(\mathbf{x}_i^{GIG-EAKF}\right)\right)$ will not be equal to $\mathbf{y}_i^{GIG-EAKF}$. To correct this inconsistency, we seek minima of the penalty function

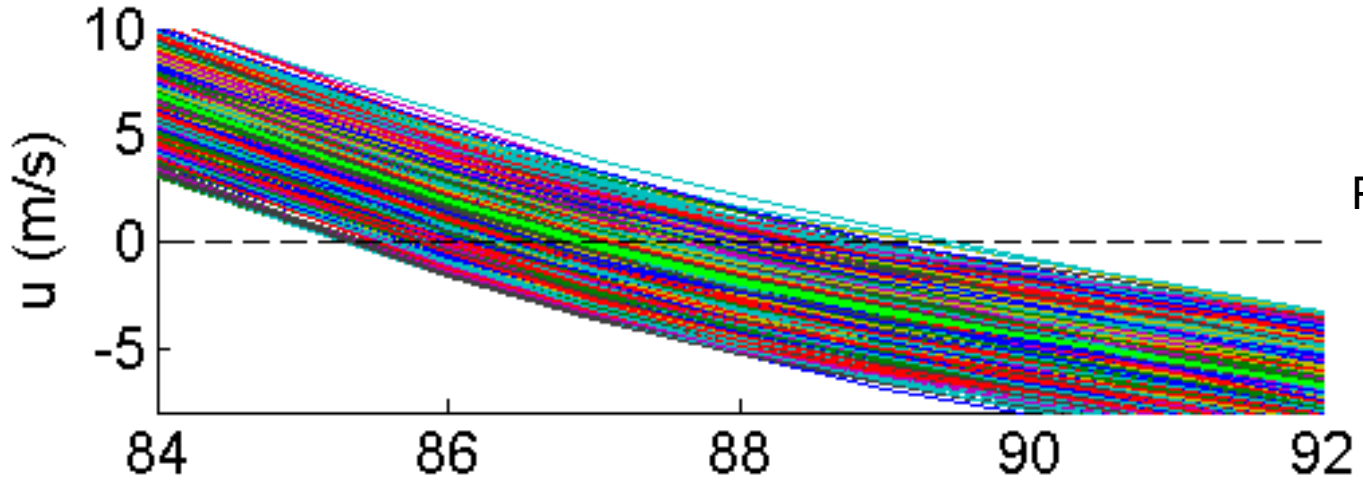
$$J_i\left(\mathbf{x}_i\right)=\frac{1}{2}\left(\mathbf{x}_i-\mathbf{x}_i^{GIGG-EAKF}\right)^T\left(\mathbf{P}_{GIGG-EAKF}^a\right)^{-1}\left(\mathbf{x}_i-\mathbf{x}_i^{GIGG-EAKF}\right)+$$

$$\frac{1}{2}\left[H\left(M\left(\mathbf{x}_i\right)\right)-\mathbf{y}_i^{GIGG-EAKF}\right]^T\mathbf{G}^{-1}\left[H\left(M\left(\mathbf{x}_i\right)\right)-\mathbf{y}_i^{GIGG-EAKF}\right]$$

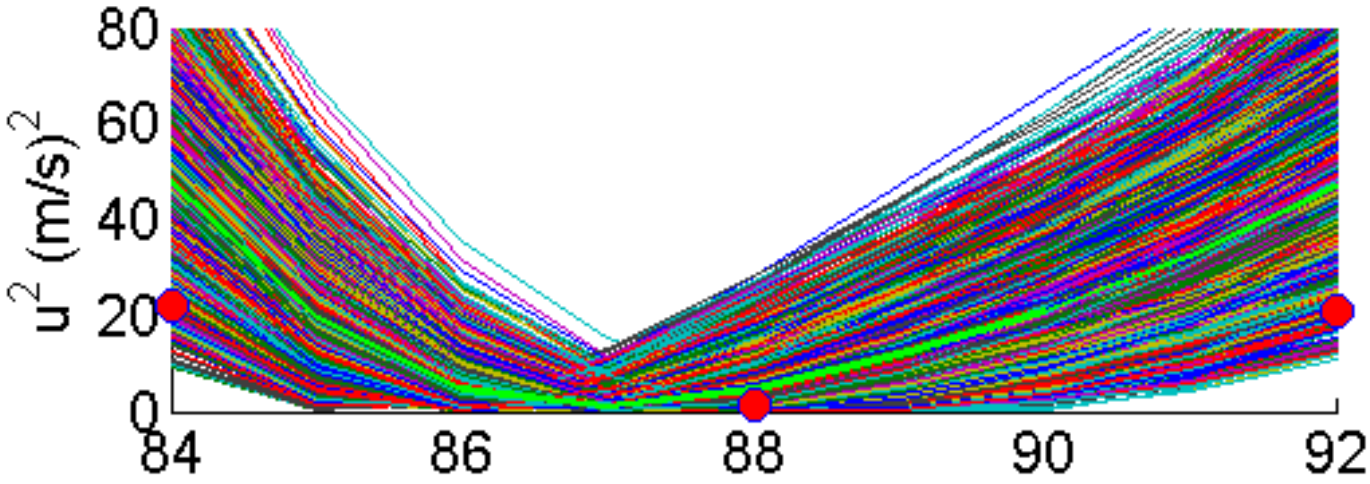
The matrix $\mathbf{P}_{GIGG-EAKF}^a$ is the covariance of the $\mathbf{x}_i^{GIGG-EAKF}$ distribution.

$\mathbf{G}=\varepsilon\left[diag\left(\mathbf{H}\mathbf{P}_{GIG-EAKF}^a\mathbf{H}^T\right)\right]$ where ε is a small positive scalar.

Need for GIGG-EnKF plus 4DVAR



Prior ensemble of zonal wind

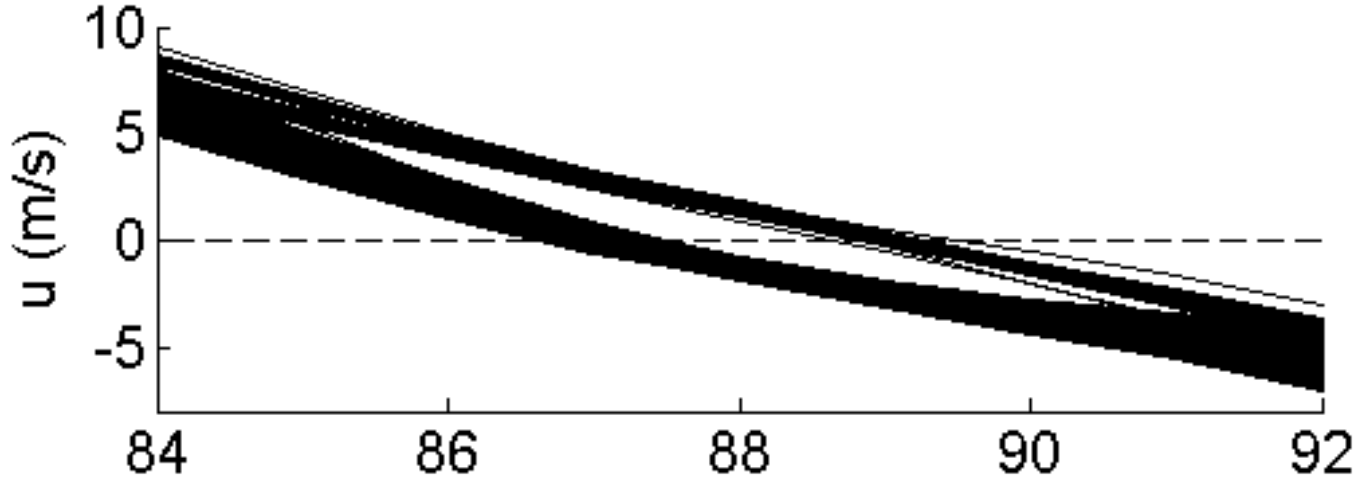


Prior ensemble of zonal wind squared

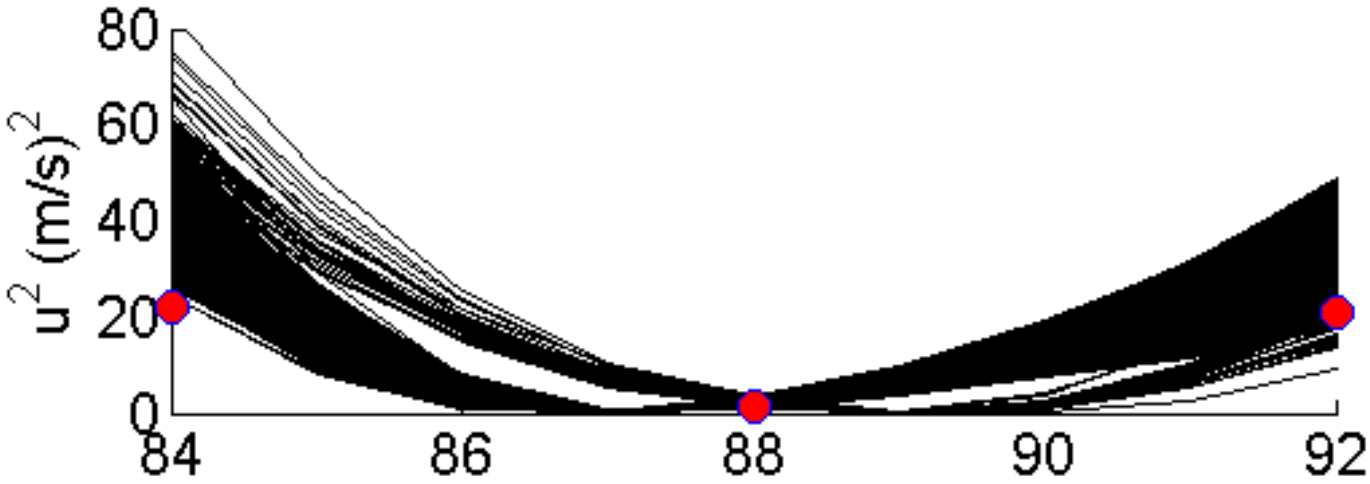
$y = 22.2\ m^2 s^{-2}$ $y = 1.27\ m^2 s^{-2}$ $y = 21.2\ m^2 s^{-2}$

Red dots are observations of the square of the zonal wind

Shouldn't the posterior look like this?



Posterior ensemble of zonal wind



Posterior ensemble of zonal wind squared

$y = 22.2\ m^2 s^{-2}$ $y = 1.27\ m^2 s^{-2}$ $y = 21.2\ m^2 s^{-2}$

Summary of GIGG-EnKF results

i. Much better than EnKF for skewed, semi-positive definite near zero uncertainty distributions.

ii. Log-normal approach unsatisfactory because it produces observation bias and cannot account for high probability densities near and at zero.

iii. GIGG-EnKF likely to be further improved by using tools from 4DVAR and/or replacing the linear regression step by a higher order regression.

In progress: Extension to variables like precipitation/cloud whose prior pdfs are best modelled as sum of delta function at zero plus gamma pdfs.