U.S. NAVAL RESEARCH LABORATORY

# Episodic, non-linear and non-Gaussian: data assimilation for bounded semi-positive definite variables like clouds using the Gamma Inverse-Gamma (GIG) variation on the EnKF


**Craig H. Bishop**

*Marine Meteorology Division, Naval Research Laboratory, Monterey, USA*

**Derek J. Posselt**

*Jet Propulsion Laboratory, California Institute of Technology, Pasadena, USA*

U.S. NAVAL RESEARCH LABORATORY

# Overview

- Aim: Outline some improvements to GIGG-EnKF (Bishop, 2016, QJRMS.
  1. Background
  2. GIG-EnKF solution to Bayes' theorem for gamma prior and inverse-gamma-likelihood is now precise as K=>infinity - previously just approximate. ***Significance: Rigorous basis for GIG for all moments***
  3. Test of standard GIG for tropical cyclone surface wind energy assimilation problem: ***Significance: Standard GIG better than EAKF/EnKF for this problem.***
  4. Local iterative regression to account for non-linearity in observation operator. ***Significance: Greatly reduces analysis error.***
  5. Rigorous approach for dealing with on-off variables (rain, cloud, fire, etc) with gamma based delta function. ***Significance: Justifies ignoring dry members when rain is observed.***

# Background: A typical EnKF serial observation assimilation scheme

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

Step 1: Do univariate Gaussian assimilation of $y_j$ to obtain $y_{ji}^a$, $i = 1, 2, \ldots, K$

Step 2: Find corresponding analysis ensemble for observations and model variables

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

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

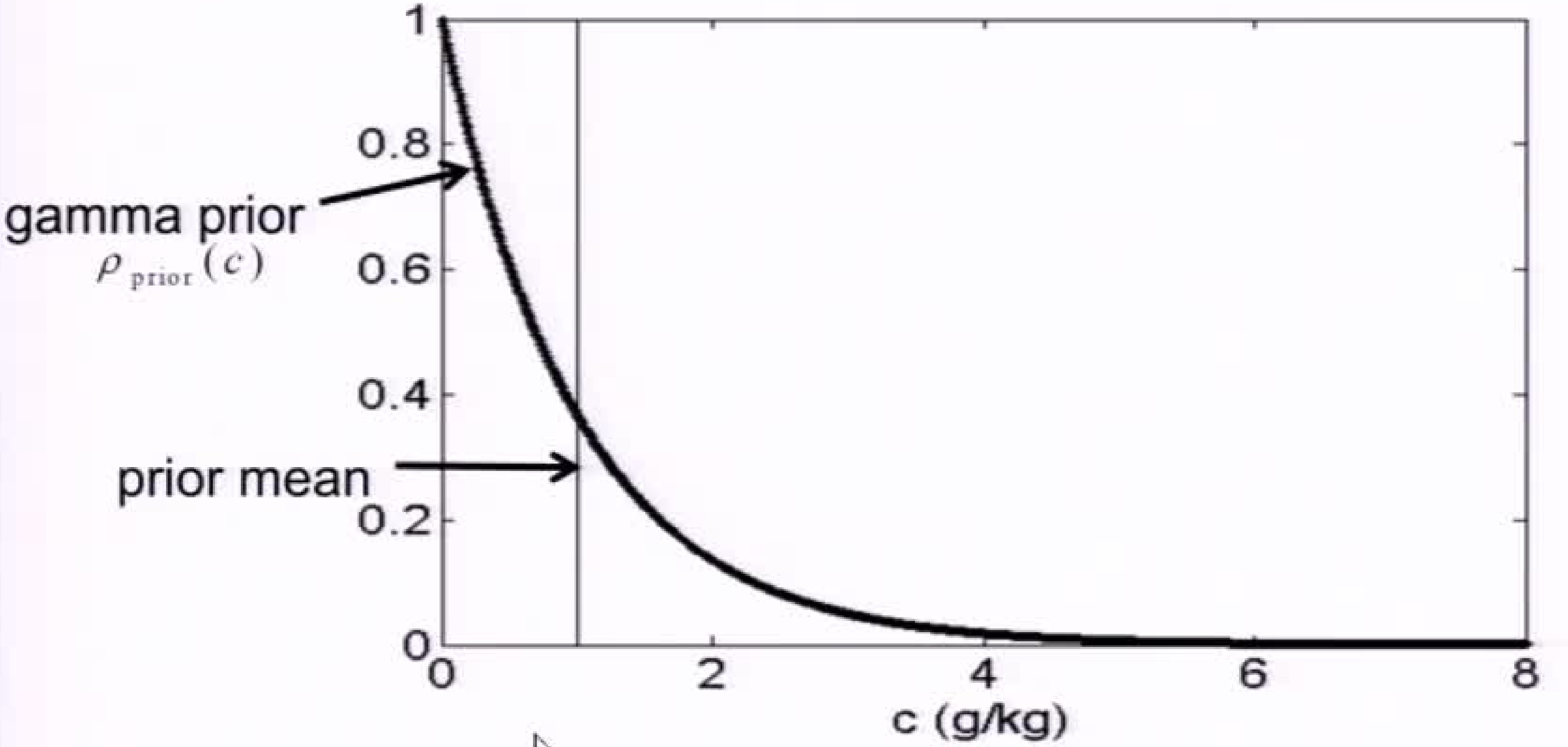
Step 3: Let the analysis ensemble be the prior ensemble for the next observation

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

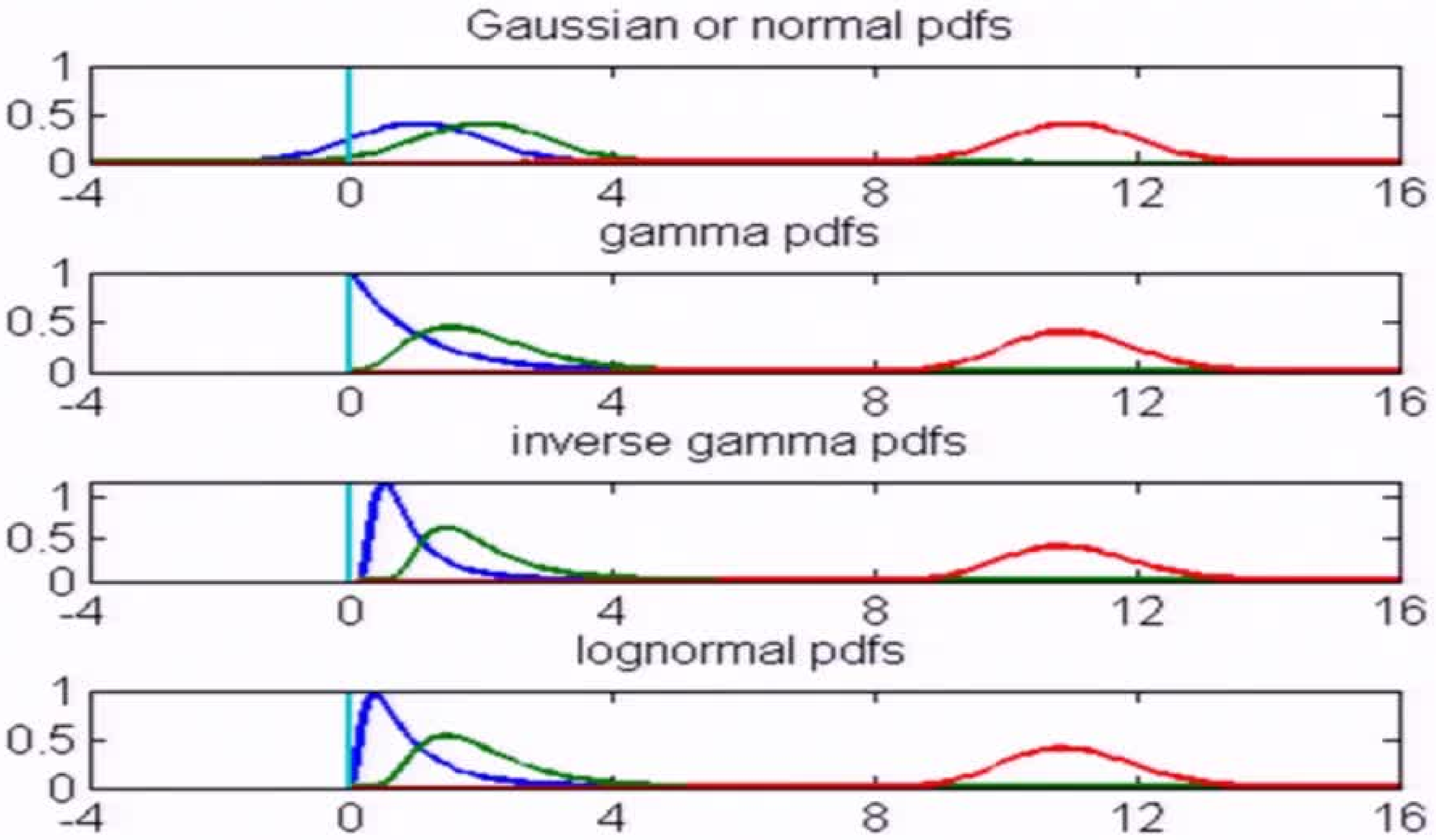
$x_{\mu i}^f = x_{\mu i}^a$, for $\mu = 1, 2, \ldots, n;\; i = 1, 2, \ldots, K$

end

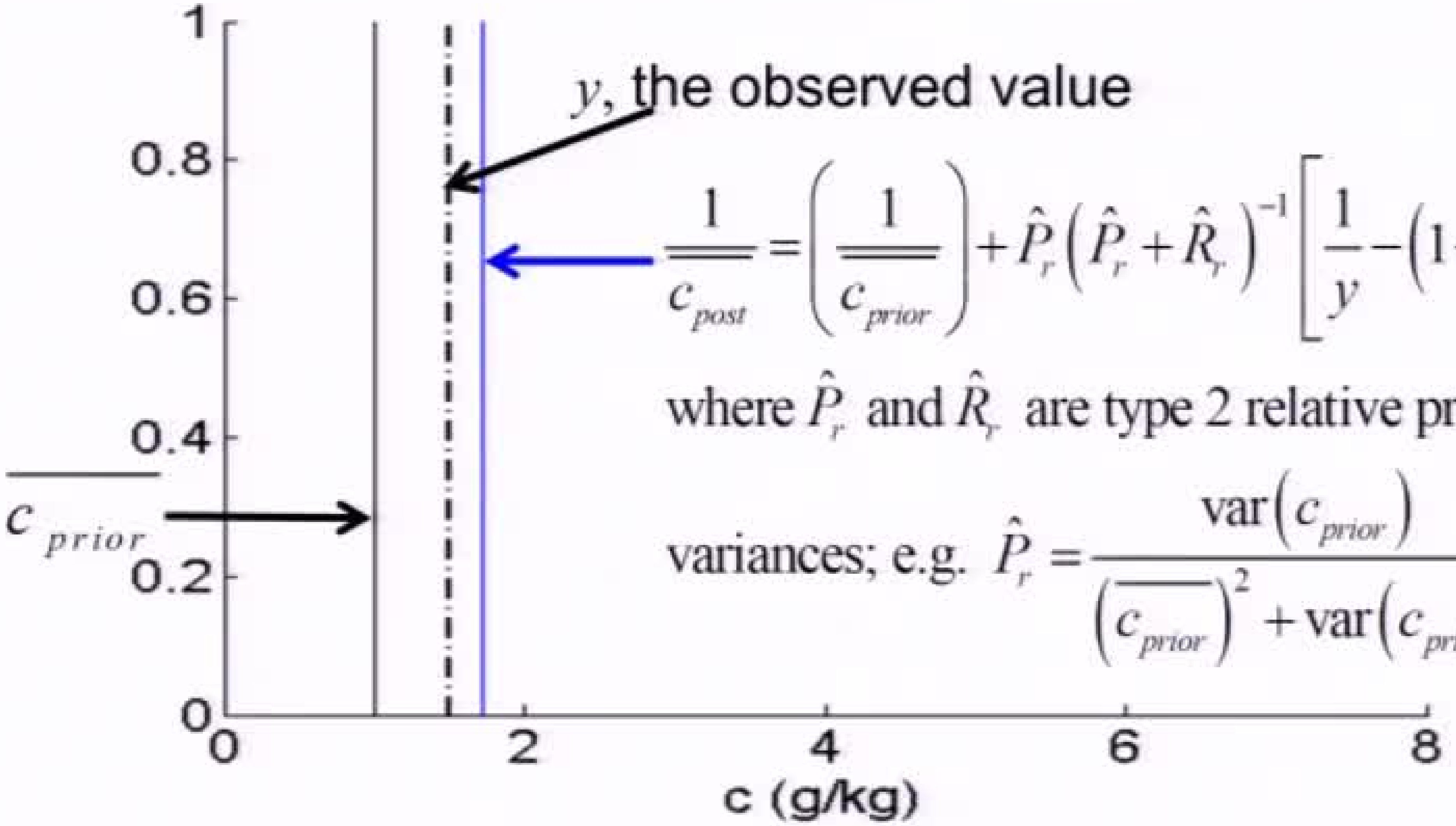
# A prior Gamma pdf



# Background: Gaussian pdfs versus bounded pdfs



# Equation for posterior mean

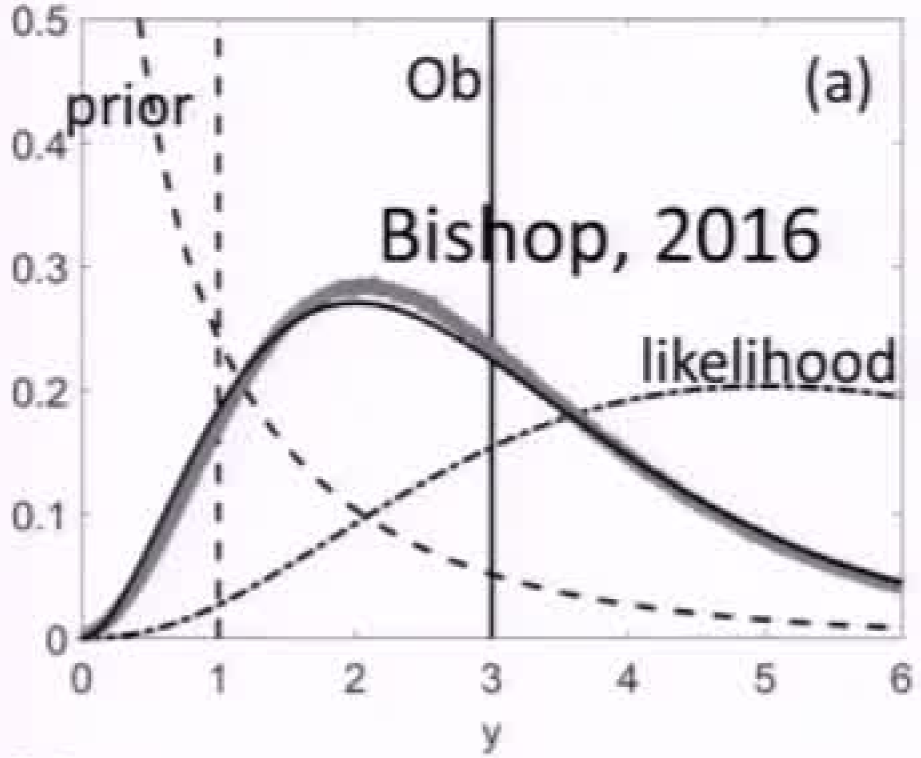


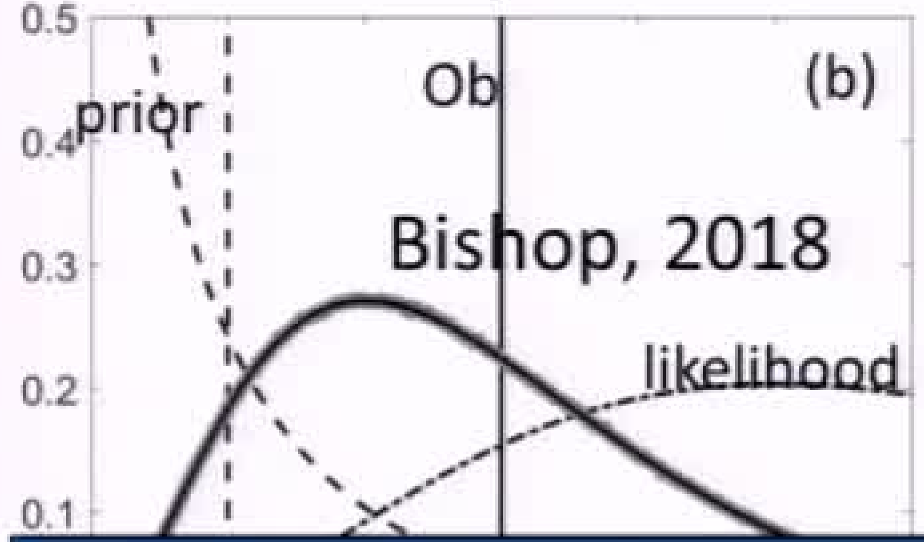

$$\frac{1}{\overline{c}_{post}} = \left(\frac{1}{\overline{c}_{prior}}\right) + \hat{P}_r\left(\hat{P}_r + \hat{R}_r\right)^{-1}\left[\frac{1}{y} - \left(1 + \hat{R}_r\right)\left(\frac{1}{\overline{c}_{prior}}\right)\right]$$

where $\hat{P}_r$ and $\hat{R}_r$ are type 2 relative prior and observation variances; e.g. $\hat{P}_r = \dfrac{\mathrm{var}\left(c_{prior}\right)}{\left(\overline{c}_{prior}\right)^2 + \mathrm{var}\left(c_{prior}\right)} = \dfrac{\sigma^2}{\mu^2 + \sigma^2}$

Posterior mean equation has Kalman like gain but everything else is inverted !
(See Bishop, 2016, QJRMS and Bishop, 2018 for details of GIG method)

# Improvement of GIG for high relative variances




Problem: Bishop's (2016, QJRMS) formulation (thick grey curve) departs from true posterior (thin solid line) when relative variance of prior (dashed curve) and likelihood (dot-dash curve) is large.

Solution: Use summation theorem for gamma pdfs to ensure that GIG-EnKF samples the true posterior. (Bishop 2018,hopefully)

$$y_i^a = \frac{\overline{y^a}}{\overline{y^f}} \frac{\left(P^r\right)^{-1}}{\left(\left(\tilde{P}^r\right)^{-1} + \left(\tilde{R}^r\right)^{-1}\right)} \left(y_i^f + y_i^{gig}\right), \quad y_i^{gig} \sim \Gamma\left(\left(\tilde{R}^r\right)^{-1} + 1, P^r \left\langle y^f \right\rangle\right),$$

$$\frac{1}{\overline{y^a}} = \left(\frac{1}{\overline{y^f}}\right) + \tilde{P}_r\left(\tilde{P}_r + \tilde{R}_r\right)^{-1}\left[\frac{1}{y} - \left(1 + \tilde{R}_r\right)\left(\frac{1}{\overline{y^f}}\right)\right],$$

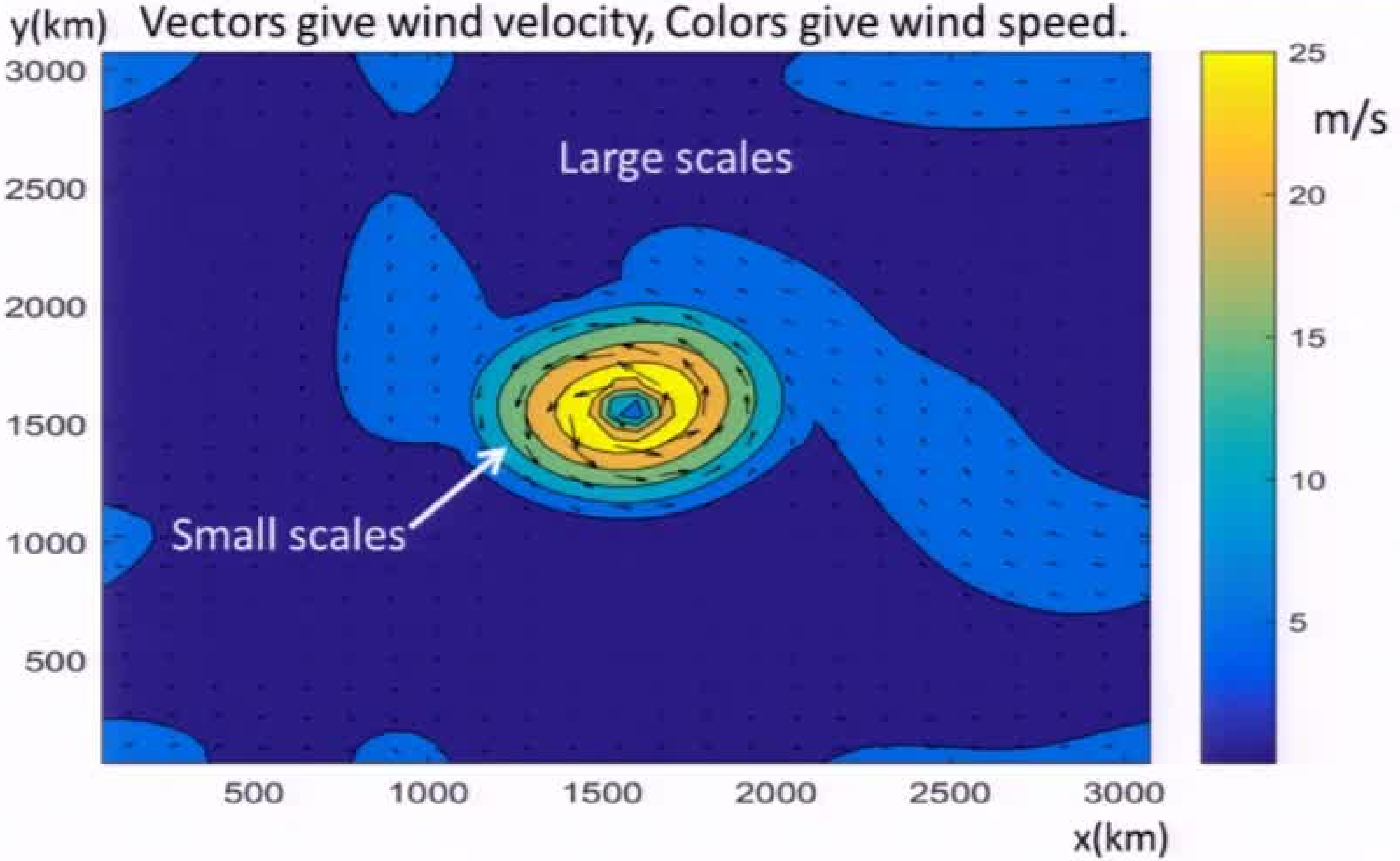
**GIG-EnKF now rigorous for all posterior moments**

U.S. NAVAL RESEARCH LABORATORY

# Overview

- Aim: Outline some improvements to GIGG-EnKF (Bishop, 2016, QJRMS).
  1. Background
  2. Solution to Bayes' theorem for gamma prior and inverse-gamma-likelihood is now precise as K=>infinity - previously just approximate. ***Significance: Rigorous basis for GIG***
  3. Test of standard GIG for tropical cyclone surface wind energy assimilation problem: ***Significance: Standard GIG better than EAKF/EnKF for this problem.***
  4. Local iterative regression to account for non-linearity in observation operator. ***Significance: Greatly reduces analysis error.***
  5. Rigorous approach for dealing with on-off variables (rain, cloud, fire, etc) with gamma based delta function. ***Significance: Justifies ignoring dry members when rain is observed.***

# Simple DA testbed for TC like surface winds

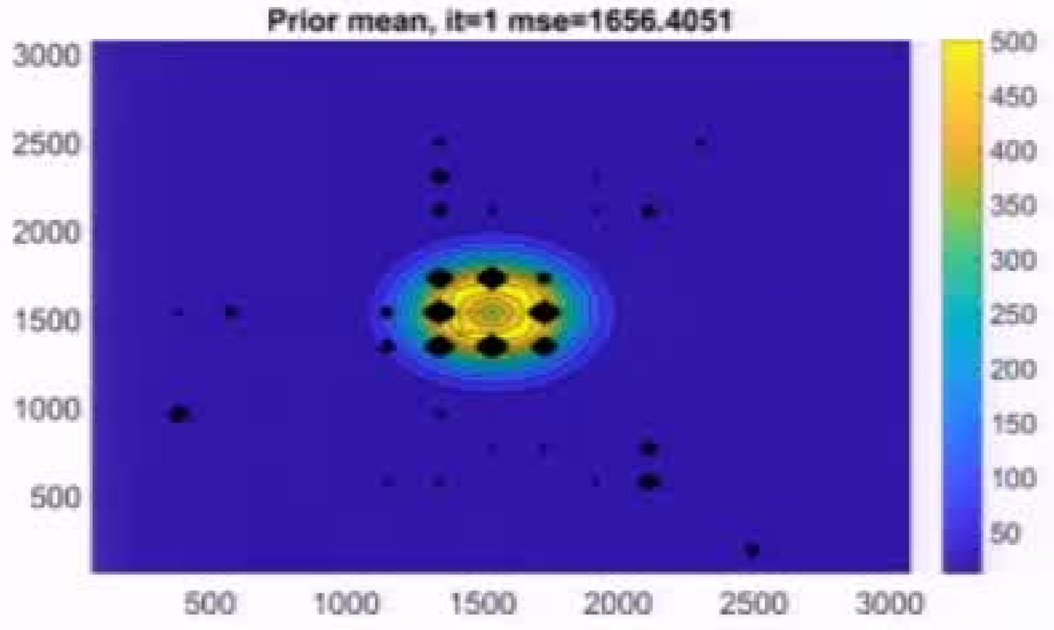


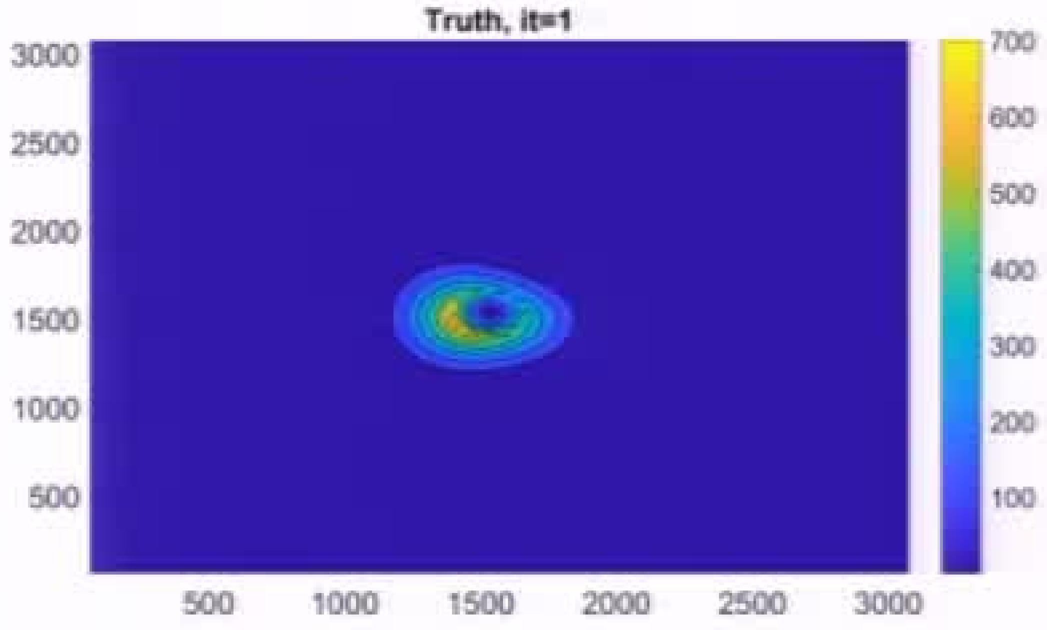

**Another random draw from the simple testbed's multi-scale pdf**

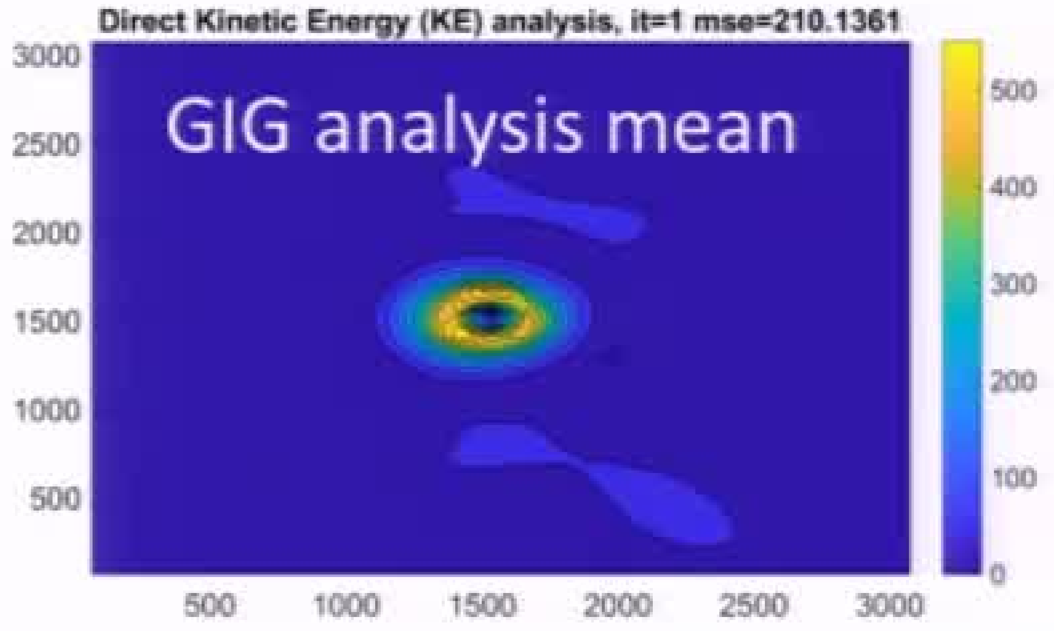
# Simple DA testbed for TC like surface winds

- Model states defined by random, multi-scale TC like (u,v) wind field.
- **Let observations be non-linear functions of u and v; e.g. Kinetic Energy, $KE=(u^2+v^2)/2$, tanh(*KE*) or Heaviside(*KE-constant*).**

# Prior mean, obs, truth and GIG analysis using a 3000 member ensemble (no localization required).




Observed variable is $KE=0.5(u^2+v^2)$. Distribution of random observations given truth is an inverse gamma pdf with a relative variance of 0.25.

# Background: The GIGG-EnKF serial observation assimilation scheme with *linear* regression

U.S. NAVAL RESEARCH LABORATORY

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-delta, GIG, IGG or Gaussian assumptions.

Step 2: if (GIG-delta) then use ... to obtain $y_{ji}^a,\ i = 1,2,\ldots,K$;

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

Note: GIG posterior mean eq *very* different to EnKF eq

$$\frac{1}{y^a} = \left(\frac{1}{y^f}\right) + \tilde{P}_r\left(\tilde{P}_r + \tilde{R}_r\right)^{-1}\left[\frac{1}{y} - \left(1 + \tilde{R}_r\right)\left(\frac{1}{y^f}\right)\right],$$

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

else if (Gaussian) then use ... to obtain $y_{ji}^a,\ i = 1,2,\ldots,K$ (EAKF/EnSRF/EnKF)

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

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

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

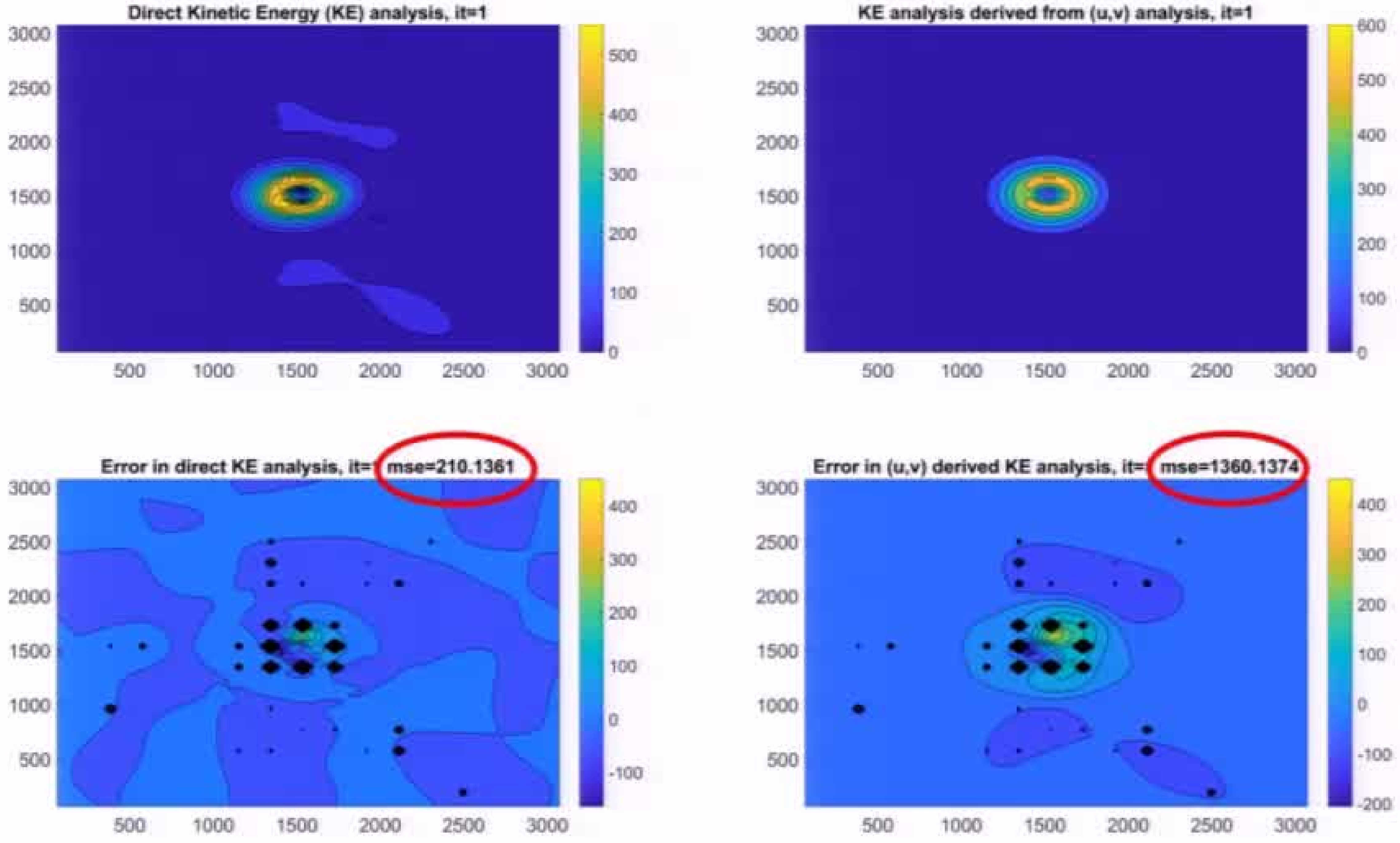
Step 4: Let the analysis ensemble be the prior ensemble for the next observation

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

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

end

# Treatment of non-linearity of Kinetic Energy ob operator (linear regression from ob to model space yields inconsistencies)




Observed variable is $KE=0.5(u^2+v^2)$. Standard GIG/EAKF uses linear regression to give an inconsistent analysis of $(u^a,v^a)$ and $(KE)^a$. Bottom left panel gives $(KE)^a$. Bottom right gives,

$$KE\left(u^a,v^a\right)=\frac{1}{2}\left(u^{a2}+v^{a2}\right)\neq\left(KE\right)^a$$

- which is far less accurate than $(KE)^a$.

U.S. NAVAL RESEARCH LABORATORY

# New method to account for non-linearity in ob-operator: The observation to model space consistency iteration

3.1: Define minimum list of variables $\left[\left(\mathbf{x}_i\right)_{h_j}\right]$ required to predict the ob $y_j$ that was just assimilated;

for example, in the KE example $\left(\mathbf{x}_i\right)_{h_j} = \begin{bmatrix} u_j \\ v_j \end{bmatrix}$ where $\left(u_j, v_j\right)$ are the wind components required

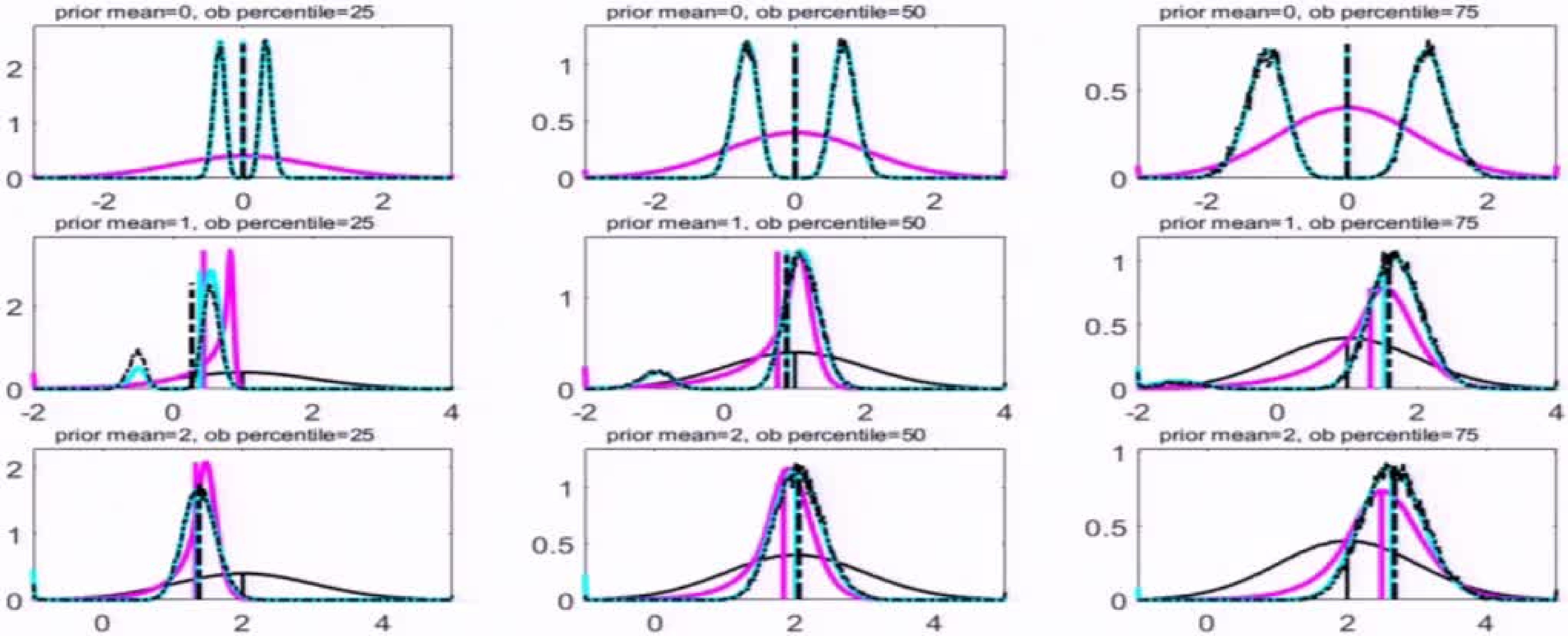
to predict the KE of the model state at $y_j$.

3.2: Find the usual GIG-EnKF model-space analysis $\left[\left(\mathbf{x}_i^a\right)_{h_j}\right]_{lin}$.

3.3: Starting with $\left(\mathbf{x}_i\right)_{h_j} = \left[\left(\mathbf{x}_i^a\right)_{h_j}\right]_{lin}$ minimize $J\left[\left(\mathbf{x}_i\right)_{h_j}\right] = \frac{1}{2}\left\{y_{ji}^a - h_j^{local}\left[\left(\mathbf{x}_i\right)_{h_j}\right]\right\}^2$

using ensemble-space constrained Newton iteration on gradient to obtain $\left(\mathbf{x}_i^a\right)_{h_j}$ (the minimizer).

# The observation to model space consistency iteration. Test in 1D model in which only $u^2$ is observed




Solid black line gives prior pdf of zonal wind (***u***) field

***u***$^2$ is observed at 25$^{th}$, 50$^{th}$ or 75$^{th}$ percentile of prior pdf of obs (left to right)

Dashed black line gives true posterior pdf of u field

Solid mauve line is GIG posterior pdf with *linear regression*

Solid cyan line is GIG posterior pdf with *non-linear observation to model space consistency iteration*

U.S. NAVAL RESEARCH LABORATORY

# New method to account for non-linearity in ob-operator: The observation to model space consistency iteration

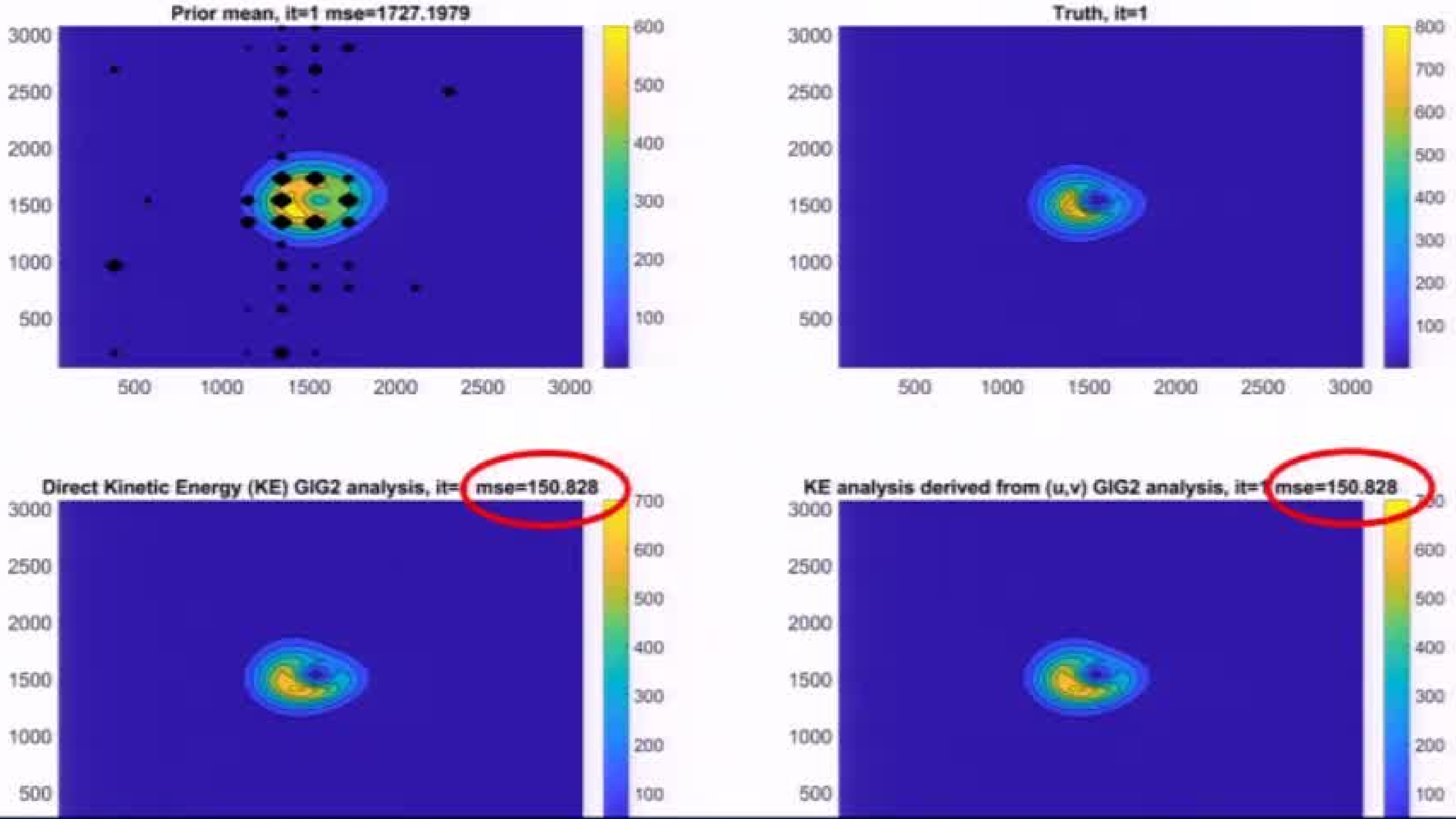
3.4: Update the rest of the model state using $\left(\mathbf{x}_i^a\right)_{h_j}$ and linear **multivariate** regression

$$x_{\mu i}^a = x_{\mu i}^f + \mathrm{covar}\left[x_{\mu i}^f, \left(\mathbf{x}^f\right)_{h_j}\right]\left\{\mathrm{covar}\left[\left(\mathbf{x}^f\right)_{h_j}, \left(\mathbf{x}^f\right)_{h_j}\right]\right\}^{-1}\left[\left(\mathbf{x}_i^a\right)_{h_j} - \left(\mathbf{x}_i^f\right)_{h_j}\right], \text{ for } \mu = 1,2,\ldots,n;\ i = 1,2,\ldots,K$$

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

$$y_{ki}^a = \begin{cases} 0.5\left[\left(y_{ki}^a\right)_{lin} + h_k^{local}\left(\mathbf{x}_i^a\right)_{h_k}\right] & \text{if } \left(y_{ki}^a\right)_{lin} \geq 0 \\ h_k^{local}\left(\mathbf{x}_i^a\right)_{h_k} & \text{if } \left(y_{ki}^a\right)_{lin} < 0 \end{cases}, \quad \text{for } k = 1,2,\ldots,p;\ i = 1,2,\ldots,K$$

# The observation to model space consistency iteration. Test in 2D model

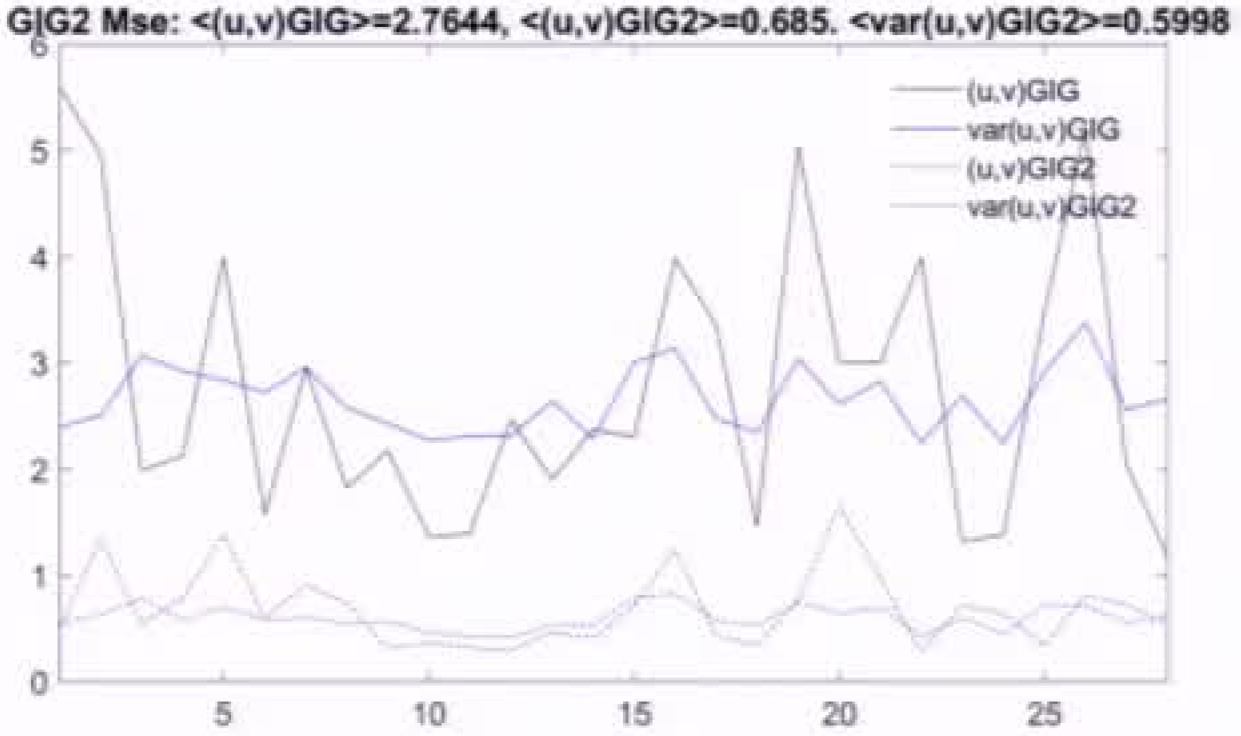


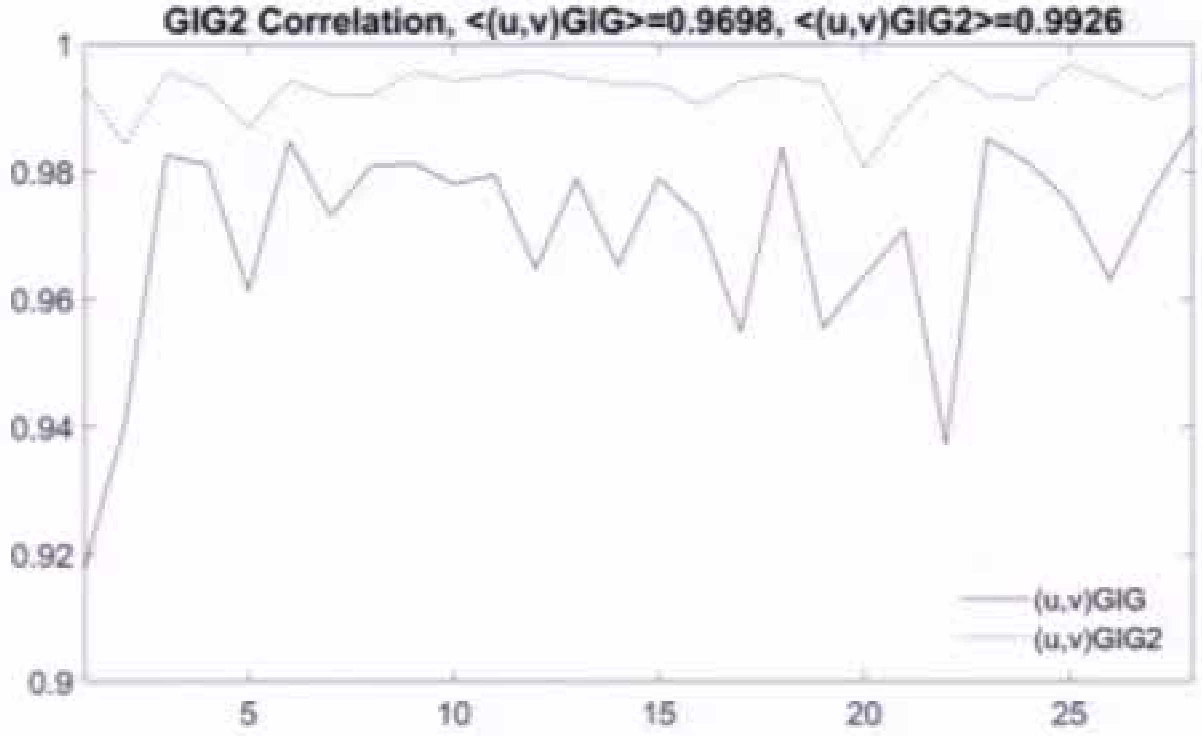

Observed variable is $KE=0.5(u^2+v^2)$.
Linear regression plus consistency iteration improves consistency of $(u^a,v^a)$ and $(KE)^a$. Bottom left panel gives $(KE)^a$. Bottom right gives,

$$KE\left(u^a,v^a\right)=\frac{1}{2}\left(u^{a2}+v^{a2}\right)\neq\left(KE\right)^a$$

**Accuracy of direct and derived KE analyses are now the same**

# The observation to model space consistency iteration. 28 independent tests in 2D model




Ob-to-model space consistency iteration reduces mse in (u,v) field by 75%;
i.e. standard deviation of analysis error is halved
Chance of getting 28 wins (as above) by pure chance is 1 in $2.8 \times 10^8$ .