Ensemble Transform Kalman Filter (ETKF) extensions for multiscale localization and near-bound variables

Craig H. Bishop

(Benefited and supported by discussions and work with Bureau DA team, Diego Carrio, Dillon Sherlock, Jianyu Liang, Yuka Muto, Takemasa Miyoshi, Shunji Kotsuki, Xuguang Wang)

ARC Centre of Excellence for Climate Extremes, Bureau of Meteorology, and the School of Geography, Earth and Atmospheric Sciences, The University of Melbourne, Parkville, Victoria, Australia

Background

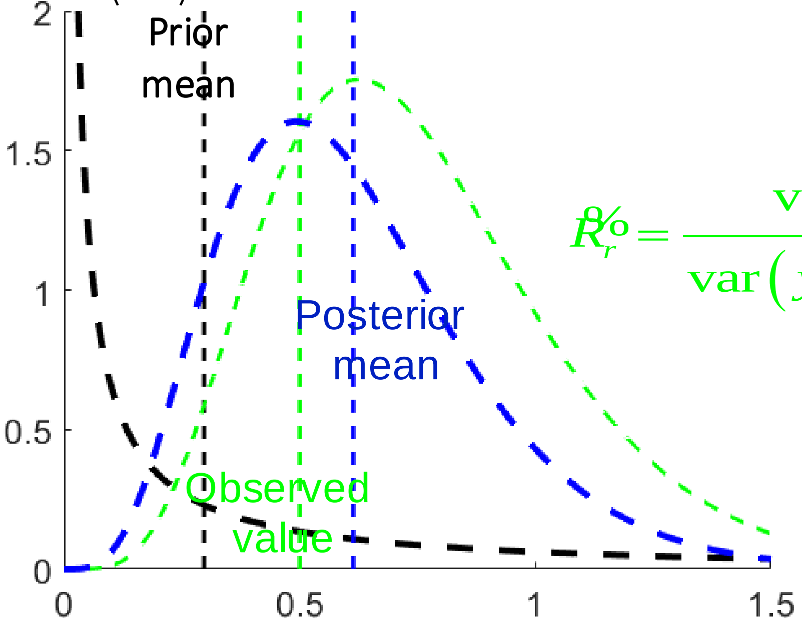
- July 2023: Met Office DA team considering using NOAA's EnKF/ETKF for convective scale DA via their participation in JEDI
- Sep 2023: NOAA's Sergey Frolov asks for help to improve NOAA's EnKF/ETKF for convective scale DA
- Foci
  - New method for multiscale fcst error covariance localisation
  - New method for near bound variables that:
    - Accounts for fact that ob-error variance is a strong function of the *unknown* true state; e.g. ob-error for rainfall small/large when true rainfall small/large
    - Account for highly skewed nature of uncertainty distributions

Skewness can make the posterior mean be bigger than both the prior mean and the observed value

Dashed black lines pertain to prior/forecast pdf with

$$P^r = \frac{\mathrm{var}\left(y^f\right)}{\left\langle y^f \right\rangle^2} = 11.3 \text{ and } \left\langle y^f \right\rangle = \left(P^r\right)^{-1/2}$$

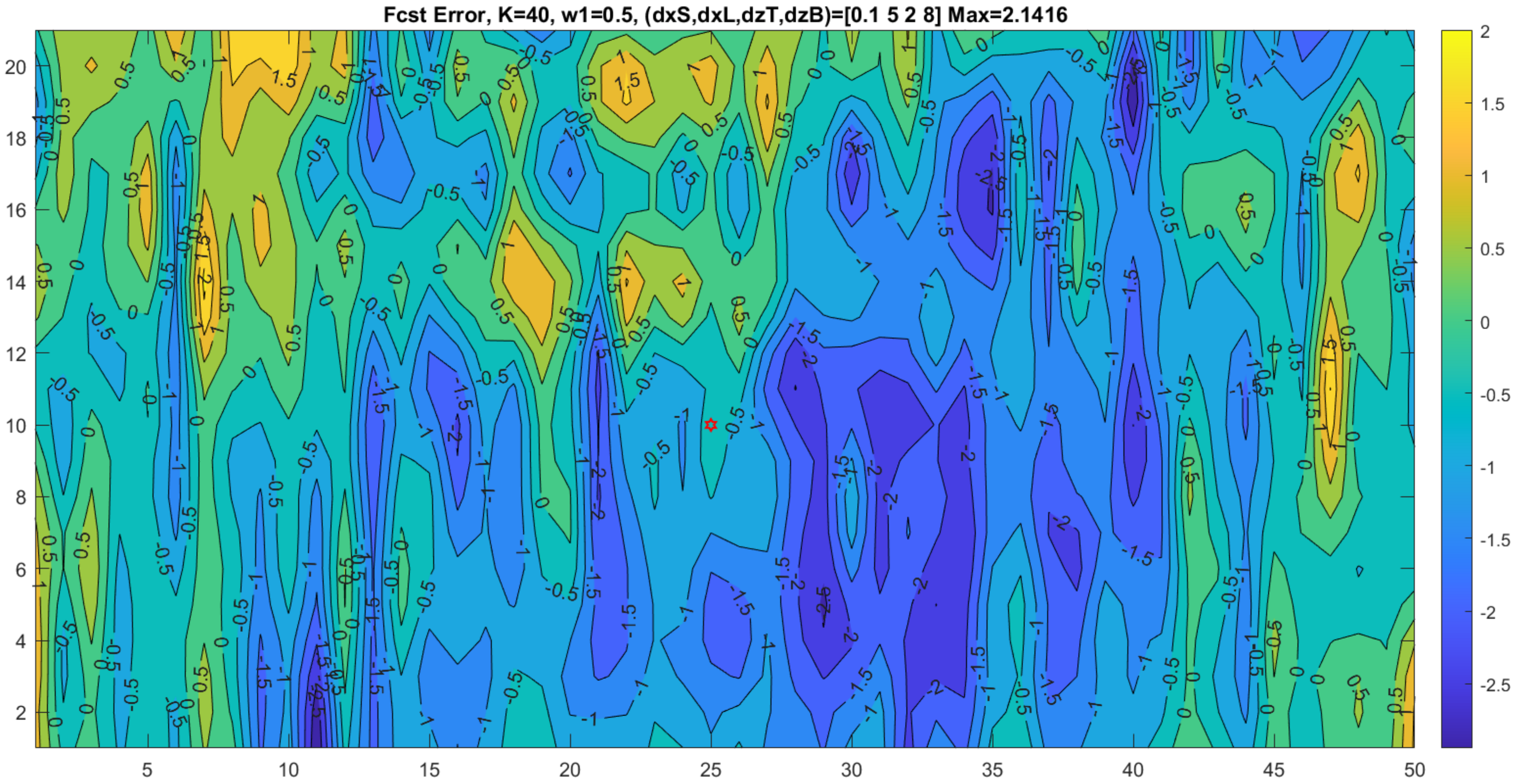
Dashed blue lines pertain to posterior/analysis pdf



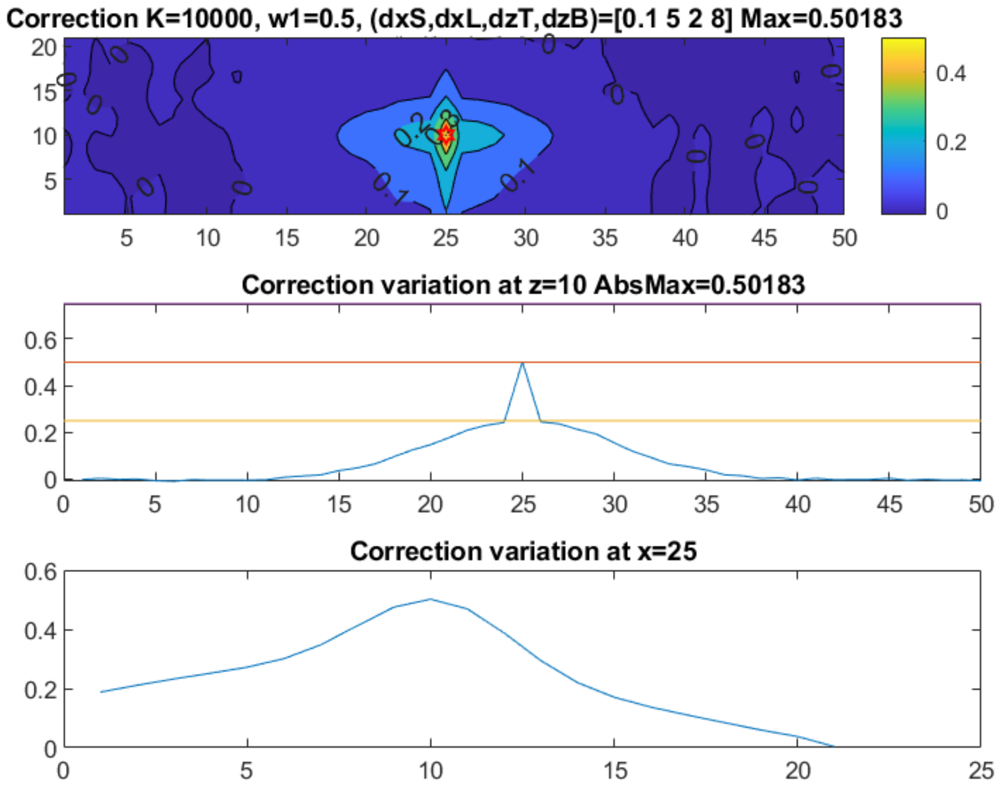
$$R_r^o = \frac{\mathrm{var}\left(y^o - y^t\right)}{\mathrm{var}\left(y^o - y^t\right) + \left(y^t\right)^2} = 1/4$$

Observation likelihood pdf

Days when x-z plots of forecast error look like this.

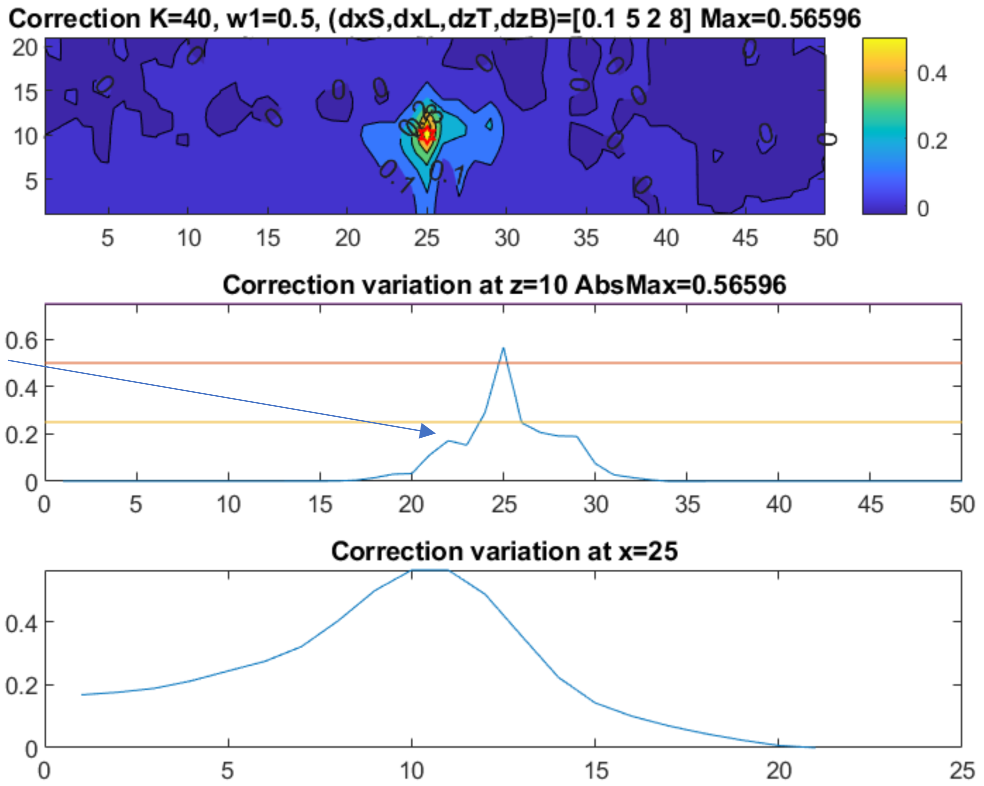


Optimal single-ob corrections look like this.

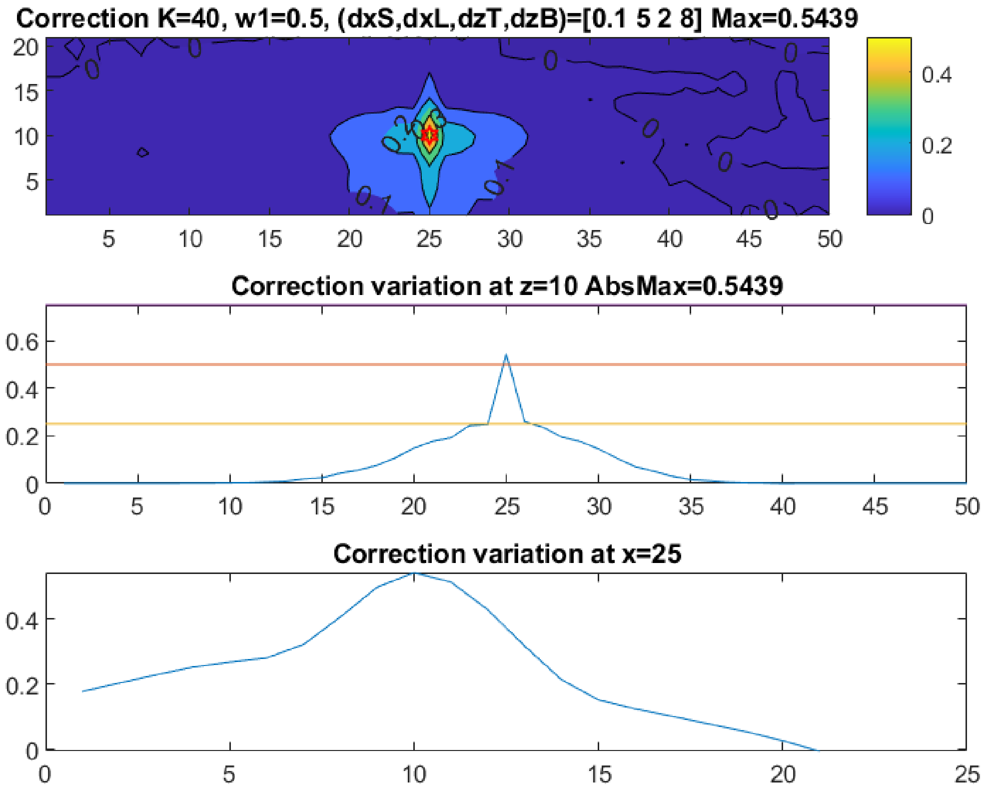


NOAA’s LETKF correction with nens=40

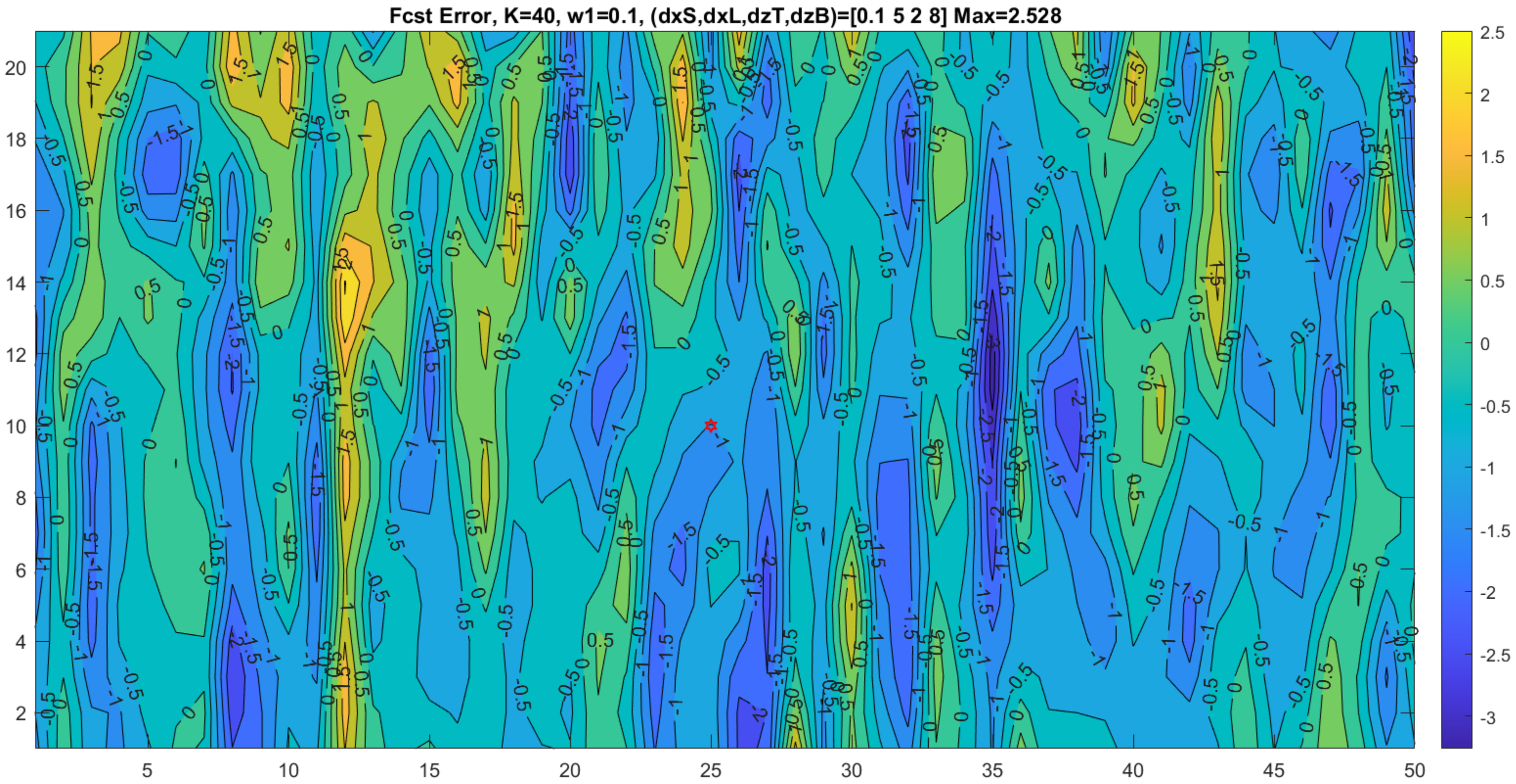
Spurious covariances



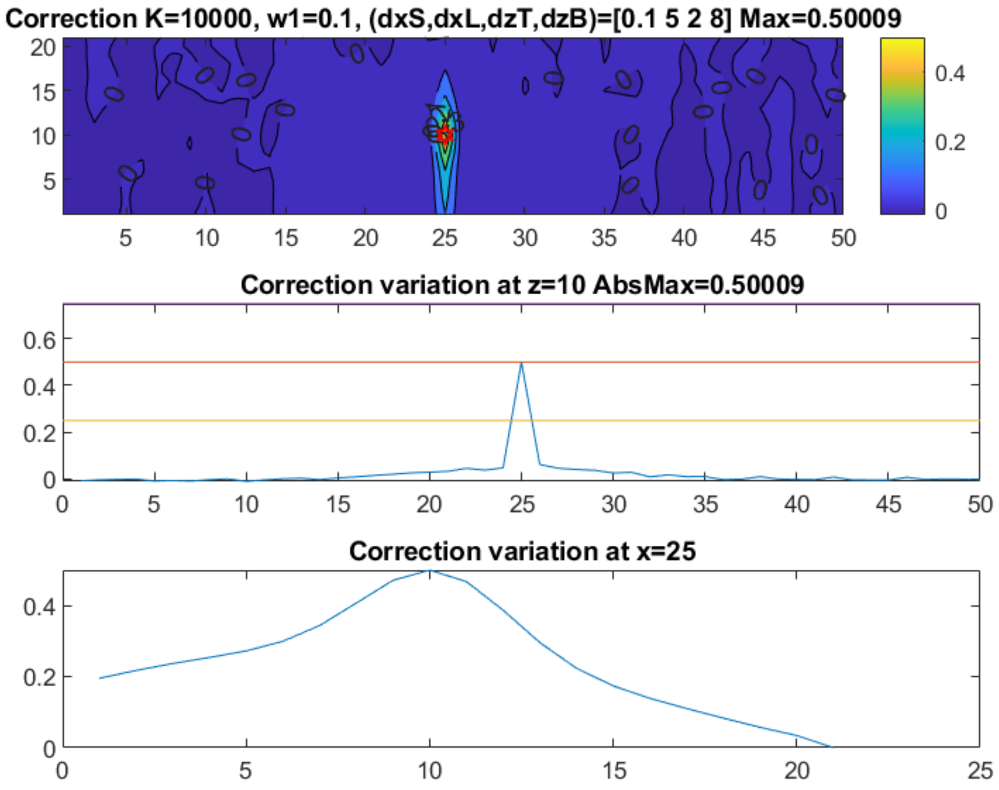
Whereas, new multiscale R-localization provides the improved correction



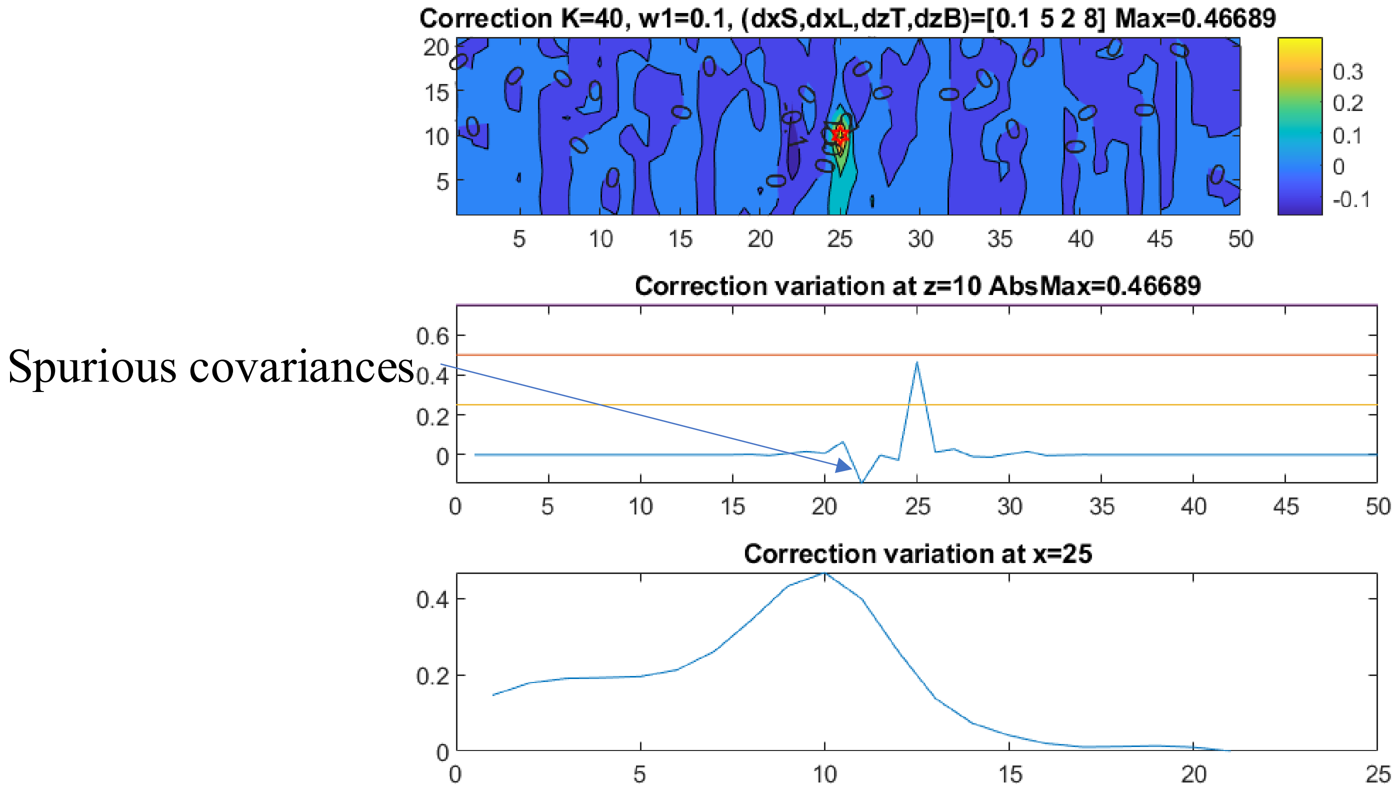
Days when x-z plots of forecast error look like this.



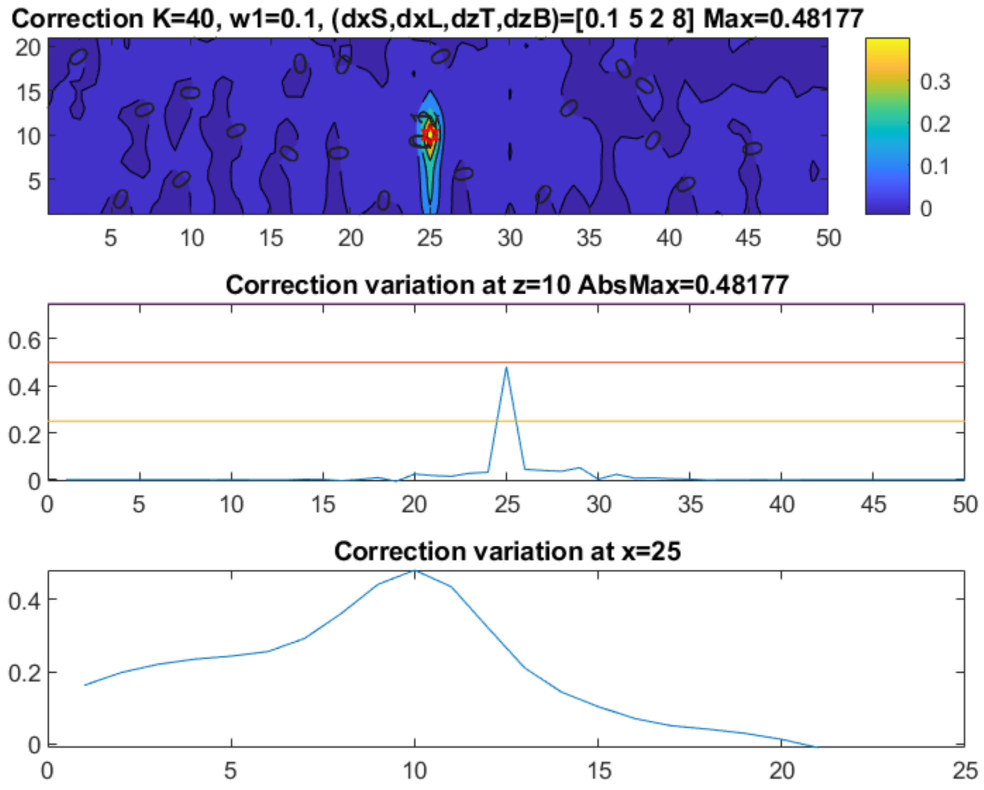
Optimal single-ob corrections look like this.



NOAA's LGETKF correction with nens=40



New multiscale R-localization provides the improved correction (nens=40).

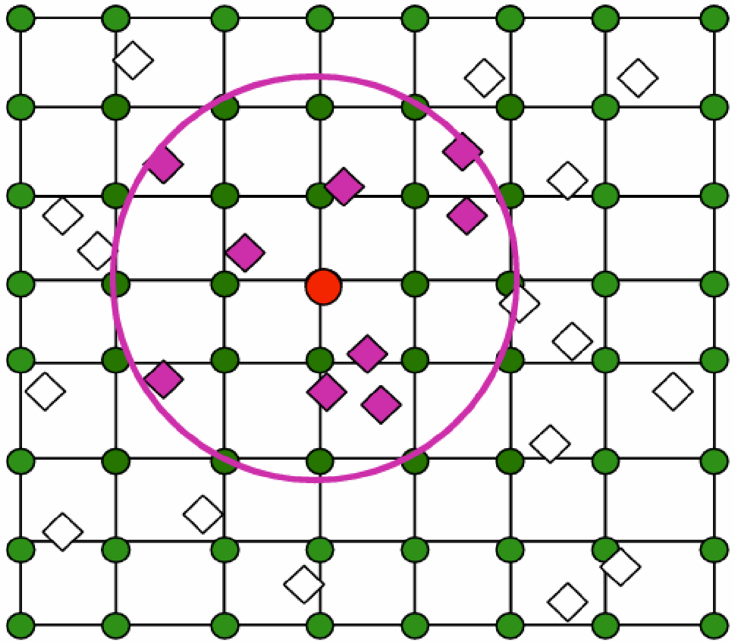


Background: NOAA's Gain form of LETKF (Bishop et al. 2017, Lei et al. 2018)

Vertical column of variables at red dot is updated using local obs (purple triangles).

Vertical localisation is achieved by inexpensively expanding the ensemble size from 80 to *~1200* member ensemble using modulation. (Bishop and Hodyss, 2009, et al.)

Horizontal localisation is achieved by artificially increasing ob-error variances with horizontal distance from red dot



MultiScale localization in Ens/Hybrid-Var versus LETKF

$$J(\mathbf{x}) = \frac{1}{2}\left\{\left(\mathbf{x} - \langle \mathbf{x}^f \rangle\right)^T \mathbf{P}^{-1}\left(\mathbf{x} - \langle \mathbf{x}^f \rangle\right) + \left(\mathbf{y} - H(\mathbf{x})\right)^T \mathbf{R}^{-1}\left(\mathbf{y} - H(\mathbf{x})\right)\right\}$$

In Var, multiscale error covariance localisation done through **P**

In LETKF, we must do it through **R.**

How?

MultiScale R localisation for LGETKF: new aspects in blue

Letting $\mathbf{x}' = \left(\mathbf{x} - \mathbf{x}^g\right) = \left[\mathbf{N}_1 \ \mathbf{N}_2\right]\mathbf{b}$, and $\mathbf{y}' = \mathbf{y} - H\left(\mathbf{x}^g\right)$ and $\mathbf{y}' - \mathbf{H}\mathbf{x}' \approx \mathbf{y} - H\left(\mathbf{x}\right)$.

$$J\left(\mathbf{b}\right) = \frac{1}{2}\left\{\mathbf{b}^T\left(\mathbf{I} + \begin{bmatrix}\mathbf{N}_1^T \\ \mathbf{N}_2^T\end{bmatrix}\mathbf{H}^T\mathbf{R}^{-1}\mathbf{H}\left[\mathbf{N}_1 \ \mathbf{N}_2\right]\right)\mathbf{b}\right\} - \mathbf{b}^T\begin{bmatrix}\mathbf{N}_1^T \\ \mathbf{N}_2^T\end{bmatrix}\mathbf{H}^T\mathbf{R}_{MS}^{-1}\left(\mathbf{y}'\right) - \mathbf{b}^T\mathbf{d} + c,$$

$$= \frac{1}{2}\left\{\mathbf{b}^T\left(\mathbf{I} + \left[\begin{bmatrix}\mathbf{N}_1^T\mathbf{H}^T \\ \mathbf{N}_2^T\mathbf{H}^T\end{bmatrix}\left(\mathbf{R}_{MS}^{-1/2}\right)^T\right]\left[\left(\mathbf{R}_{MS}^{-1/2}\right)\left[\mathbf{H}\ \mathbf{N}_1 \ \mathbf{H}\ \mathbf{N}_2\right]\right]\right)\mathbf{b}\right\} - \mathbf{b}^T\left[\begin{bmatrix}\mathbf{N}_1^T\mathbf{H}^T \\ \mathbf{N}_2^T\mathbf{H}^T\end{bmatrix}\left(\mathbf{R}_{MS}^{-1/2}\right)^T\right]\mathbf{R}_D^{-1/2}\left(\mathbf{y}'\right) - \mathbf{b}^T\mathbf{d} + c,$$

$$\mathbf{R}_{MS}^{-1/2} = \left[\mathbf{R}_1^{-1/2}\mathbf{H}\ \mathbf{N}_1 \ \mathbf{R}_2^{-1/2}\mathbf{H}\ \mathbf{N}_2\right]\left\{\begin{bmatrix}\mathbf{N}_1^T\mathbf{H}^T \\ \mathbf{N}_2^T\mathbf{H}^T\end{bmatrix}\left[\mathbf{H}\ \mathbf{N}_1 \ \mathbf{H}\ \mathbf{N}_2\right]\right\}^{-1}\begin{bmatrix}\mathbf{N}_1^T\mathbf{H}^T \\ \mathbf{N}_2^T\mathbf{H}^T\end{bmatrix}$$

$\mathbf{R}_D^{-1/2}$ is a diagonal matrix with $\mathbf{R}_D^{-1/2}$ where

$$\mathbf{R}_D^{-1/2}\left(j,j\right) = \left(\frac{P_{j1}R_{j1}^{-1/2} + P_{j2}R_{j2}^{-1/2}}{P_j}\right), \ P_{j1} = \mathbf{H}\left(j,:\right)\mathbf{N}_1\mathbf{N}_1^T\mathbf{H}\left(j,:\right)^T, P_{j2} = \mathbf{H}\left(j,:\right)\mathbf{N}_2\mathbf{N}_2^T\mathbf{H}\left(j,:\right)^T$$

$P_j = P_{j1} + P_{j2}$, $R_{j1}^{-1/2}$ and $R_{j2}^{-1/2}$ have large and small scale attenuation rates, respectively.

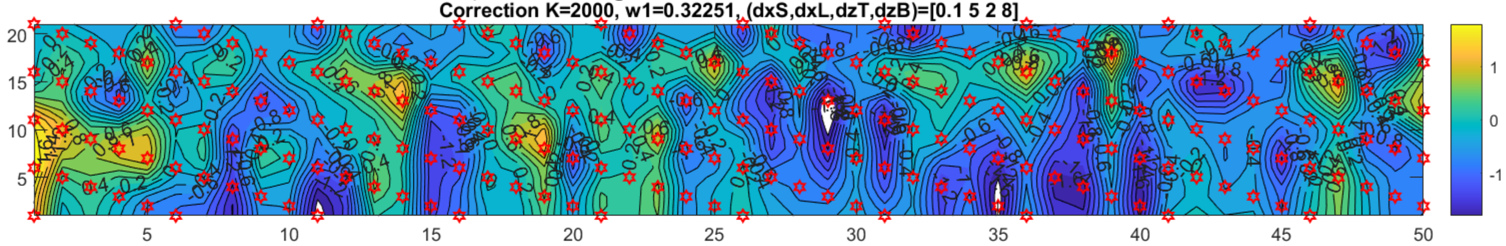
$$R_{j1}^{-1/2} = R_j^{-1/2}\sqrt{\frac{a_{j1}}{r_{ij}\left[\frac{P_j}{R_j}\left(1 - a_{j1}\right) + 1\right]}} \text{ and } r_{ij} = \frac{P_{j1} + a_{j2}P_{j2}}{P_j}$$

MultiScale R localisation for ETKF

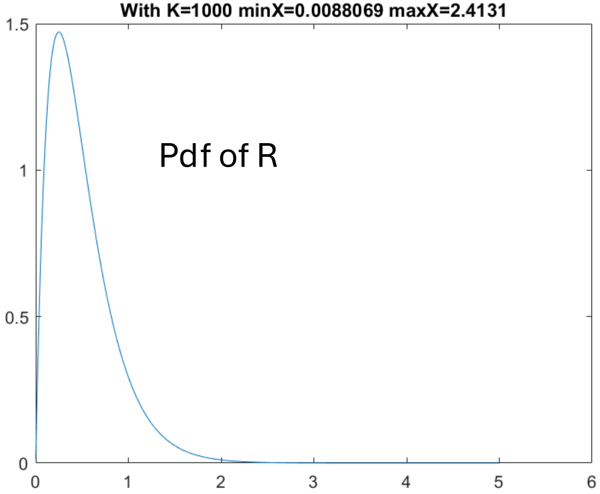
Remarkably, the equations simplify greatly when the above formulae are substituted, and one can implement multiscale R localisation by only altering the ob-operator **H** and the innovation **y-Hxf**. Otherwise, the ETKF algorithm is unchanged.

…so simpler than it looks at first glance.

Obs network as shown by red hexagrams (every 5th grid point is observed). w1 is a uniform random number between 0.01 and 1. w1 =1 means large scales have weight of 1. We'll do 140 trials of the 3 methods with 7 possible localisation lengths for each method. The field below is just a single correction from the ETKF with 2000 ensemble members.



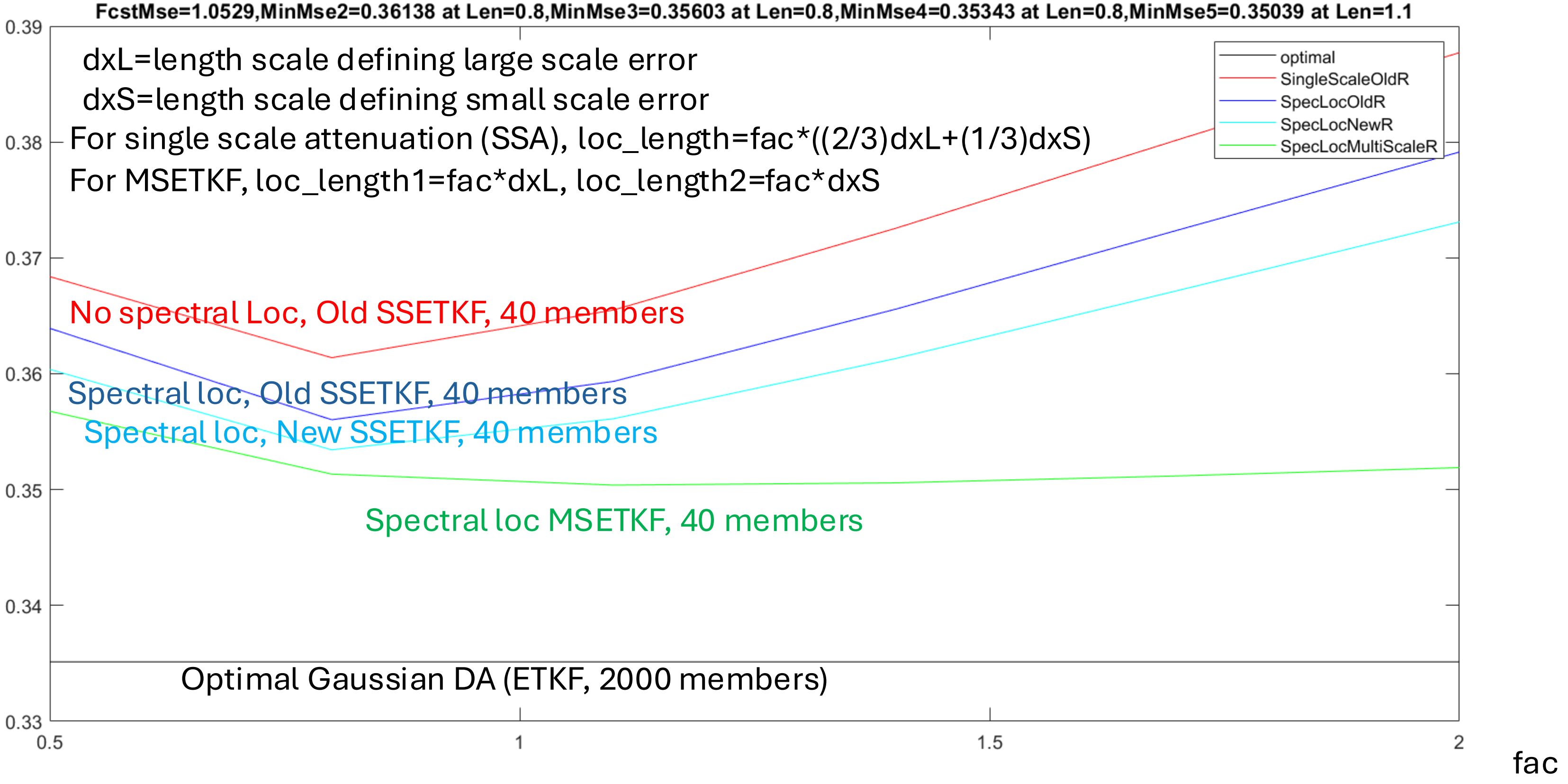
The ob-error variance was chosen to be a random draw from a gamma pdf with a mean of 0.5 and a relative variance of 0.5 (k=2,theta=0.25). **Pdf of R values given below.** Recall that the forecast error variance at all grid points has been set to 1. These small to medium R values make it difficult for the old R localisation to work properly. However, the new R localisation should have no problem in this case. Note that for univariate DA and R=0.5 and Pf=1, then Pa=Pf-Pf^2/(P+R)=1-1/1.5=1/3.



For the runs without Spectral Localisation, a tighter vertical localisation was required to remove spurious correlations in the vertical. This meant that 8 Vertical modes were required to represent vertical localisation and the Modulated ensemble size became 40x8=320 members. With spectral Localisation, a much broader vertical localisation could be used so only 4 Vertical modes were required yielding a 2x40x4=320 member ensemble;ie The modulated ensemble sizes with and without spectral localisation are the same.

MSE over 240 independent trials for 6 different localisation length scales

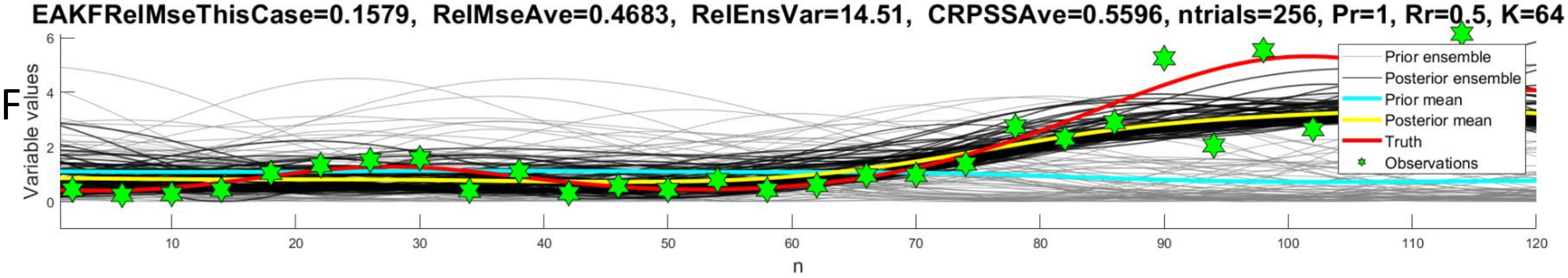
$$MSE = \frac{\left(\langle \mathbf{x}^a \rangle - \mathbf{x}^t\right)^T \left(\langle \mathbf{x}^a \rangle - \mathbf{x}^t\right)}{n_x n_z}$$



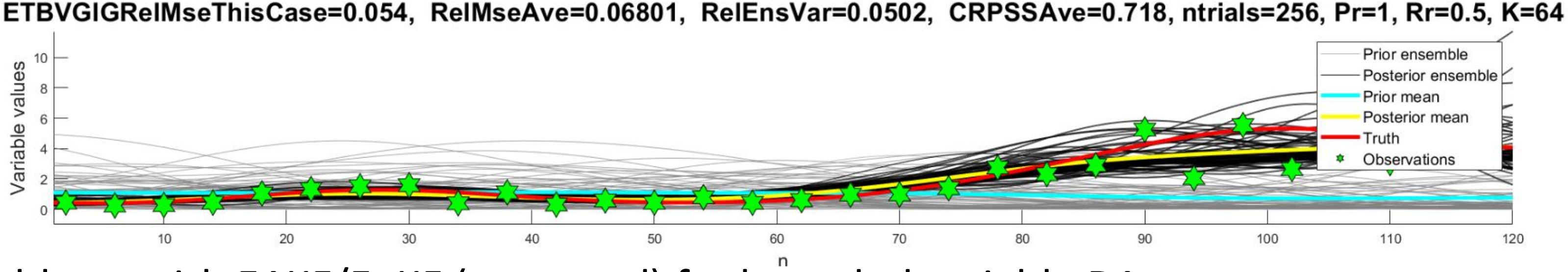
Finally, we need to add a capability for near bound variables like aerosols and H20 in all its phases

Results for Relative Error Variance (large scales): 85% Rel MSE reduction relative to Gaussian

EAKF/EnKF



New Method



Three problems with EAKF/EnKF (top panel) for bounded variable DA

1. Ob error variance is not a function of the unknown true state
2. The analysis mean does not get close enough to the truth when the observed values are bigger than the prior mean. (Associated with not accounting for skewness in ob-uncertainty)
3. Often produces unrealistic negative values in ensemble

New formulation based on generalisation of derivative of log of Gaussian posterior.

In the multivariate Gaussian case,

$$=> -\frac{\partial \ln\left(\rho\left(\mathbf{x}\,|\,\mathbf{y}\right)\right)}{\partial \mathbf{x}} = \mathbf{P}_a^{-1}\left(\left\langle \mathbf{x}^a \right\rangle - \mathbf{x}\right) = \mathbf{P}^{-1}\left(\left\langle \mathbf{x}^f \right\rangle - \mathbf{x}\right) + \frac{\partial\left[H\left(\mathbf{x}\right)\right]^T}{\partial \mathbf{x}^T}\mathbf{R}^{-1}\left(\mathbf{y} - H\left(\mathbf{x}\right)\right)$$

=> Ensemble of anals $\mathbf{x}_i^a, i = 1, 2, ..., K$ obtained from ensemble of fcsts $\mathbf{x}_i^f, i = 1, 2, ..., K$

$$\mathbf{0} = \mathbf{P}_f^{-1}\left(\mathbf{x}_i^f - \mathbf{x}_i^a\right) + \mathbf{H}^T\mathbf{R}^{-1}\left[\mathbf{y}_i - H\left(\mathbf{x}_i^a\right)\right], \quad \text{where } \mathbf{y}_i \sim N\left(\mathbf{y}, \mathbf{R}\right)$$

$$=> \mathbf{P}_a^{-1}\varepsilon^a = \mathbf{P}_f^{-1}\varepsilon^f + \frac{\partial\left[H\left(\mathbf{x}\right)\right]^T}{\partial \mathbf{x}^T}\mathbf{R}^{-1}\varepsilon^o \qquad (1)$$

Generalization hypothesis 1: Find normalized forms of ε^f and ε^o that are insensitive to unknowns such as the true prior/posterior mean (for $\varepsilon^f / \varepsilon^a$) and the true value of the observed variable (for ε^o). Let $\mathbf{P}_a$, $\mathbf{P}_f$ and $\mathbf{R}$ instead give the covariance of these normalised variables.

Generalization hypothesis 2: Choose normalized forms of ε^f and ε^o that are likely to have a similar distribution type.

For bounded variables (gamma prior, inverse-gamma obs pdf) generalization leads to …

Ensemble of anals $\mathbf{x}_i^a, i = 1, 2, \ldots, K$ obtained from ensemble of fcsts $\mathbf{x}_i^f, i = 1, 2, \ldots, K$

$$\mathbf{0} = \mathbf{P}_f^{-1}\left(\mathbf{x}_i^f - \mathbf{x}_i^a\right) + \mathbf{H}^T \vec{\mathbf{R}}^{-1}\left[\frac{\mathbf{y} \circ \mathbf{y}}{\mathbf{y}_i} - H\left(\mathbf{x}_i^a\right)\right] \quad \text{where } \mathbf{y}_i \sim \Gamma^{-1}\left(\mathbf{y}, \mathbf{R}_r\right), \;\; \mathbf{R}_r\left(j, j\right) = \frac{\left\langle \left(\mathbf{y}(j) - \mathbf{H}(j,:)\mathbf{x}\right)^2 \right\rangle}{\left(\mathbf{H}(j,:)\mathbf{x}\right)^2}$$

and $\vec{\mathbf{R}}^{-1} = \mathbf{D}^{-1/2}\left(\mathbf{I} + \mathbf{R}_r^{-1}\right)\mathbf{D}^{-1/2}$, where $\mathbf{D} = \left(diag\left(\mathbf{y} \circ \overline{\mathbf{H}\mathbf{x}^f}\right)\right)$

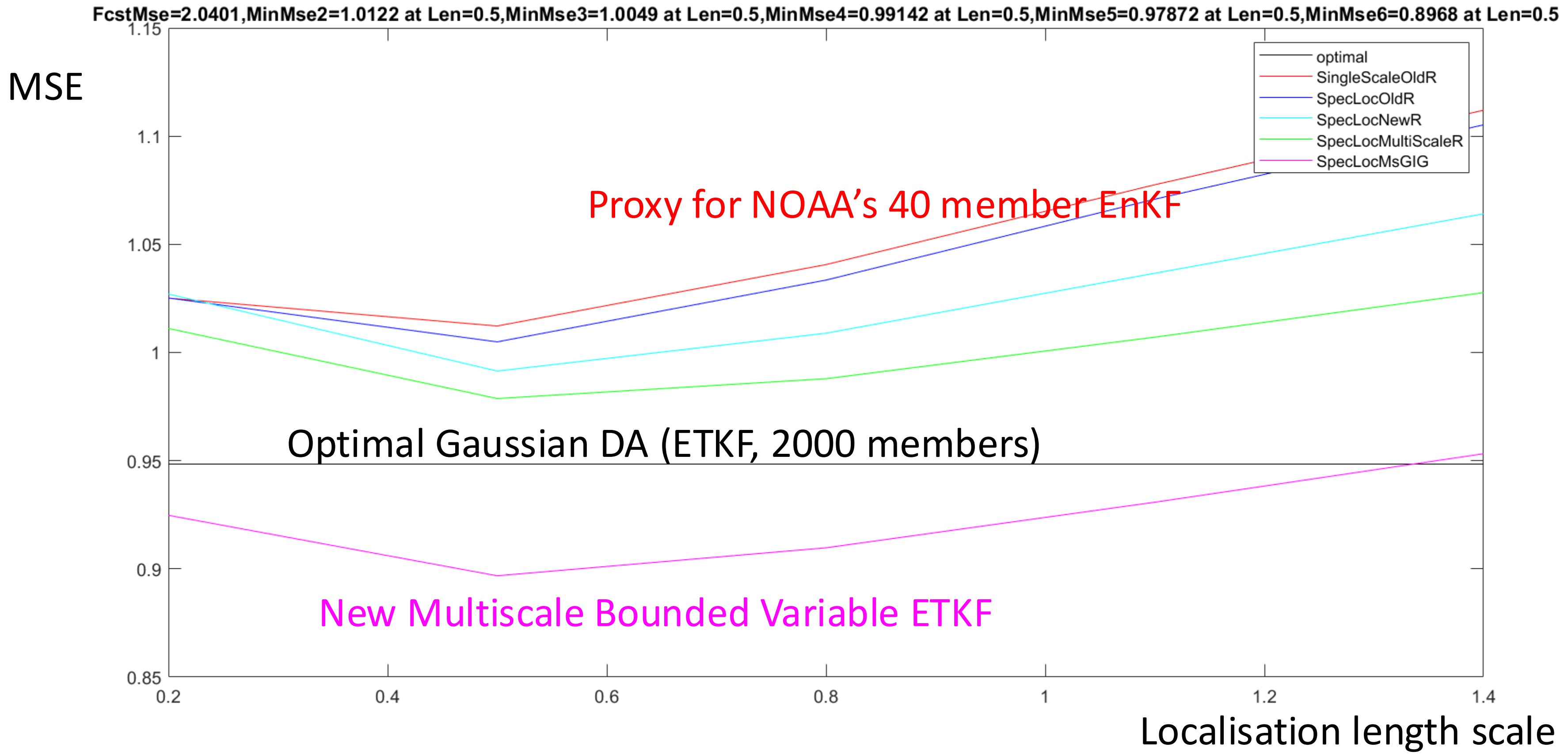
Remarkably, this equation is isomorphic to the perturbed obs Gaussian equation! (Ens of Vars)

Hence, relatively easy to implement in existing 3D/4DVar or EnKF DA schemes.

1. The fact that ob-error variance is a function of the unknown true state is accounted for by the inclusion of $\mathbf{D} = \left(diag\left(\mathbf{y} \circ \overline{\mathbf{H}\mathbf{x}^f}\right)\right)$ in the definition of $\vec{\mathbf{R}}^{-1}$.

2. Because $\left\langle \frac{\mathbf{y} \circ \mathbf{y}}{\mathbf{y}_i} \right\rangle = \left(2\mathbf{I} + \mathbf{R}_r^{-1}\right)\left(\mathbf{I} + \mathbf{R}_r^{-1}\right)^{-1}\mathbf{y} > \mathbf{y}$, the analysis mean is larger than that of the Gaussian case - other things being equal. This accounts for Likelihood skewness.

Over 210 Trials, with a fixed Rr=0.25 and the distribution of truth given by squaring each element of the vectors generated for our Gaussian experiment, we can beat optimal Gaussian performance with 40 members



Summary

- A new multiscale localisation technique for ETKF significantly outperforms NOAA's current single-scale localisation method in multiscale error regimes for a nens=40 member ensemble.
- An extension for 4DVar/ETKF type DA schemes has been developed to handle the assimilation of "near-bound" variables. Extension simultaneously accounts for variations of ob-error std with the truth and the skewness of near-bound uncertainty distributions.
- **These improvements enable the bounded variable ETKF with multiscale R-localisation to outperform the optimal Gaussian scheme (2000 ensemble members) with just 40 ensemble members.**
- With help from BoM's data assimilation team, we hope to implement aspects of these improvements in December 2024 during a planned trip to Boulder.