2018 SIAM Annual Meeting

MS47: Advances in Data Assimilation for Geosciences

Portland, Oregon

July 9-13, 2018

A Formulation of Forecast Error Covariance in High-Dimensional Ensemble Data Assimilation Suitable for State-Space Localization

Milija Zupanski

Cooperative Institute for Research in the Atmosphere
Colorado State University
Fort Collins, Colorado

Motivation (1): Covariance localization

- *Covariance localization is necessary in high-dimensional ensemble and hybrid variational-ensemble data assimilation*
  - insufficient degrees of freedom from ensembles
  - state-space and observation-space localization

- *Observation-space localization:* increase observation error with distance from a central grid-point
  - Inconsistent for vertical localization when observations are vertically-integrated
  - Additional localization problem with observations impacting several components of a strongly-coupled modeling system

- *State-space localization:* apply Hadamard product between ensemble covariance and a pre-defined correlation matrix
  - Straightforward to apply vertical localization and to account for shared observations in a strongly-coupled systems

Motivation (1): Covariance localization

- *Covariance localization is necessary in high-dimensional ensemble and hybrid variational-ensemble data assimilation*
  - insufficient degrees of freedom from ensembles
  - state-space and observation-space localization

- *Observation-space localization:* increase observation error with distance from a central grid-point
  - Inconsistent for vertical localization when observations are vertically-integrated
  - Additional localization problem with observations impacting several components of a strongly-coupled modeling system

- *State-space localization:* apply Hadamard product between ensemble covariance and a pre-defined correlation matrix
  - Straightforward to apply vertical localization and to account for shared observations in a strongly-coupled systems

Motivation (2): Bayesian inference

- Data assimilation is a recursive application of Bayes formula over time
- The power of DA comes from Bayesian inference. In DA practice Bayesian inference typically reduces to first two moments of a PDF (e.g., Gaussian PDF assumption)
- ***VAR***: Bayesian inference in terms of the first moment of a PDF. No recursive estimation of forecast/analysis error covariance.
- ***ENS***: Bayesian inference in terms of the first and second moments of a PDF (mean, covariance).
- ***Hybrid ENS-VAR***: Bayesian inference in terms of the first moment is kept, but Bayesian inference in terms of the second moment is broken! Typically two separate DA systems, VAR and ENS. VAR produces the analysis, but the analysis error covariance estimate is from ENS and therefore corresponds to the ENS analysis.
- Therefore, ***hybrid ENS-VAR*** represents an improvement of VAR, but a degradation of ENS method

How to improve hybrid ENS-VAR?

- Estimate error covariance in hybrid ENS-VAR consistently with the analysis. This points to a need for a single system that does both the first and second moment estimation.

- Several benefits of hybrid ENS-VAR are generally noticed:

(1) Increased degrees of freedom result from combining static and ensemble error covariance
(2) Optimization/minimization is important for nonlinear processes and operators
(3) State-space covariance localization is advanageous for satellite observations.

- ENS has a formulation that allows Bayesian inference of the second PDF moment, but needs to address the above points.

- Advantages (1) and (2) have been addressed in ENS, but (3) is still a challenge.

- In this presentation we investigate a possibility for state-space covariance localization in ENS framework.

State-space square-root covariance localization used in current hybrid methods

Ensemble covariance is an outer product of ensemble perturbations

$$P_E = \sum_{i=1}^{N_e} p_i p_i^T \qquad p_i = m(x_{t-1}^a + [p_i^a]_{t-1}) - m(x_{t-1}^a)$$

Localized error covariance is obtained as a Hadamard product with correlation ***L***

$$L \circ P_E = L \circ \sum_{i=1}^{N_e} p_i p_i^T = \sum_{i=1}^{N_e} (L \circ p_i p_i^T)$$

Use Hadamard product identity …

$$L \circ ab^T = diag(a) \cdot L \cdot diag(b)$$

… to obtain localized error covariance

$$L \circ P_E = \sum_{i=1}^{N_e} diag(p_i) \cdot L \cdot diag(p_i)$$

The square root localized forecast error covariance is

$$(L \circ P_E)^{1/2} = \begin{pmatrix} D_1 L^{1/2} & \cdots & D_{N_e} L^{1/2} \end{pmatrix} \qquad D_i = diag(p_i)$$

New state-space localized covariance

Consider cost-function with localized forecast error covariance

$$J(x) = \frac{1}{2}(x - x^f)^T (L \circ P_E)^{-1} (x - x^f) + \frac{1}{2}[y - h(x)]^T R^{-1} [y - h(x)]$$

Embed identity matrix using random sample $N(0,1)$

$$(L \circ P_E) = (L \circ P_E)^{1/2} (L \circ P_E)^{T/2} = (L \circ P_E)^{1/2} I (L \circ P_E)^{T/2}$$

$$I \approx E\left[\varphi\varphi^T\right] = \frac{1}{N_r - 1} \sum_{j=1}^{N_r} \varphi_j \varphi_j^T$$

$$P_f = (L \circ P_E)^{1/2} \left[\frac{1}{N_r - 1} \sum_{j=1}^{N_r} \varphi_j \varphi_j^T \right] (L \circ P_E)^{T/2} = \frac{1}{N_r - 1} \sum_{j=1}^{N_r} \left[(L \circ P_E)^{1/2} \varphi_j \right] \left[(L \circ P_E)^{1/2} \varphi_j \right]^T$$

Use Hadamard product identity $L \circ ab^T = diag(a) \cdot L \cdot diag(b)$

to obtain new forecast error covariance

$$P_f = \frac{1}{N_r - 1} \sum_{i=1}^{N_e} \sum_{j=1}^{N_r} \left[D_i L^{1/2} \varphi_j \right] \left[D_i L^{1/2} \varphi_j \right]^T \qquad D_i = diag(p_i)$$

New square-root localized error covariance

$$F = (\; f_1 \;\; \cdots \;\; f_{N_e \times N_r} \;) \qquad f_k = \frac{1}{\sqrt{N_r - 1}} D_i L^{1/2} \varphi_j \qquad (k = 1, \ldots, N_e \times N_r)$$

State-space square-root covariance localization used in current hybrid methods

Ensemble covariance is an outer product of ensemble perturbations

$$P_E = \sum_{i=1}^{N_e} p_i p_i^T \qquad p_i = m(x_{i-1}^a + [p_i^a]_{i-1}) - m(x_{i-1}^a)$$

Localized error covariance is obtained as a Hadamard product with correlation ***L***

$$L \circ P_E = L \circ \sum_{i=1}^{N_e} p_i p_i^T = \sum_{i=1}^{N_e} (L \circ p_i p_i^T)$$

Use Hadamard product identity …

$$L \circ ab^T = diag(a) \cdot L \cdot diag(b)$$

… to obtain localized error covariance

$$L \circ P_E = \sum_{i=1}^{N_e} diag(p_i) \cdot L \cdot diag(p_i)$$

The square root localized forecast error covariance is

$$(L \circ P_E)^{1/2} = \begin{pmatrix} D_1 L^{1/2} & \cdots & D_{N_e} L^{1/2} \end{pmatrix} \qquad D_i = diag(p_i)$$

New state-space localized covariance

Consider cost-function with localized forecast error covariance

$$J(x) = \frac{1}{2}(x - x^f)^T (L \circ P_E)^{-1}(x - x^f) + \frac{1}{2}[y - h(x)]^T R^{-1}[y - h(x)]$$

Embed identity matrix using random sample *N*(0,1)

$$(L \circ P_E) = (L \circ P_E)^{1/2}(L \circ P_E)^{T/2} = (L \circ P_E)^{1/2} I (L \circ P_E)^{T/2}$$

$$I \approx E\left[\varphi\varphi^T\right] = \frac{1}{N_r - 1}\sum_{j=1}^{N_r}\varphi_j\varphi_j^T$$

$$P_f = (L \circ P_E)^{1/2}\left[\frac{1}{N_r - 1}\sum_{j=1}^{N_r}\varphi_j\varphi_j^T\right](L \circ P_E)^{T/2} = \frac{1}{N_r - 1}\sum_{j=1}^{N_r}\left[(L \circ P_E)^{1/2}\varphi_j\right]\left[(L \circ P_E)^{1/2}\varphi_j\right]^T$$

Use Hadamard product identity $L \circ ab^T = diag(a) \cdot L \cdot diag(b)$
to obtain new forecast error covariance

$$P_f = \frac{1}{N_r - 1}\sum_{i=1}^{N_e}\sum_{j=1}^{N_r}\left[D_i L^{1/2}\varphi_j\right]\left[D_i L^{1/2}\varphi_j\right]^T$$

$$D_i = diag(p_i)$$

New square-root localized error covariance

$$F = (\; f_1 \quad \cdots \quad f_{N_e \times N_r} \;) \qquad f_k = \frac{1}{\sqrt{N_r - 1}} D_i L^{1/2}\varphi_j \qquad (k = 1, \ldots, N_e \times N_r)$$

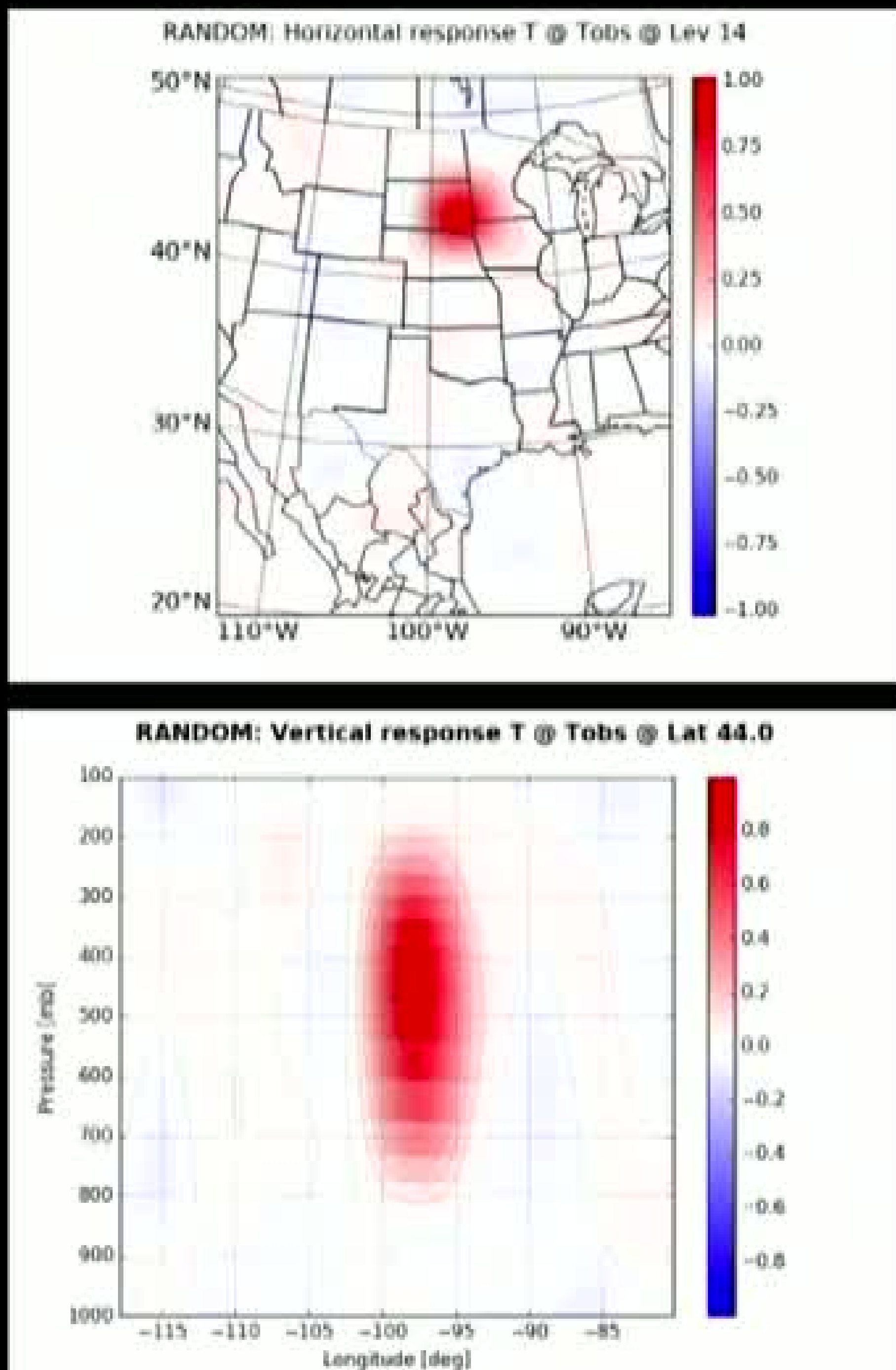
Experimental setup

- Weather Research and Forecasting (WRF) model
- 27 km /31 layer
- 32 dynamical ensembles
- (1) 1024 and (2) 4096 random ensembles
- 6-hour forecast error covariance
- ***Random:*** use $D_i=I$ in localized covariance formulation $P_f = \frac{1}{N_r - 1}\sum_{i=1}^{N_r}(L^{1/2}\varphi_i)(L^{1/2}\varphi_i)^T$
- ***Total:*** use complete localized covariance formulation $P_f = \frac{1}{N_r - 1}\sum_{i=1}^{N_e}\sum_{j=1}^{N_r}\left[D_i L^{1/2}\varphi_j\right]\left[D_i L^{1/2}\varphi_j\right]^T$
- Single observation experiment

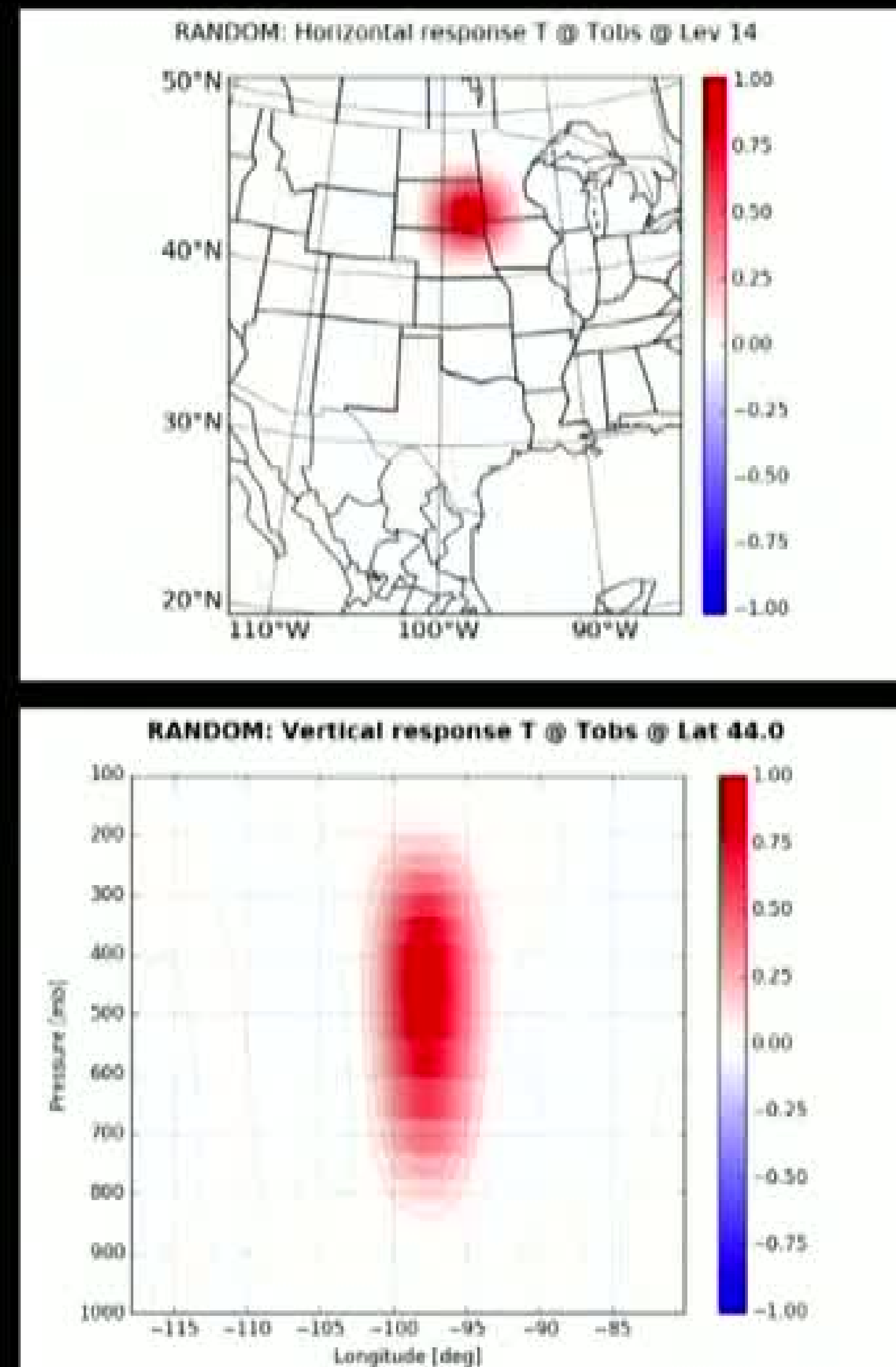
$$x^a - x^f = (L \circ P_E)^{1/2}(L \circ P_E)^{T/2}\frac{\Delta x}{\left(\sigma_f^2 + \sigma_R^2\right)}(0 \quad \cdots \quad 1_k \quad \cdots \quad 0)^T$$

Random autocovariance ($COV_{T\text{-}T}$)

1024 ensembles



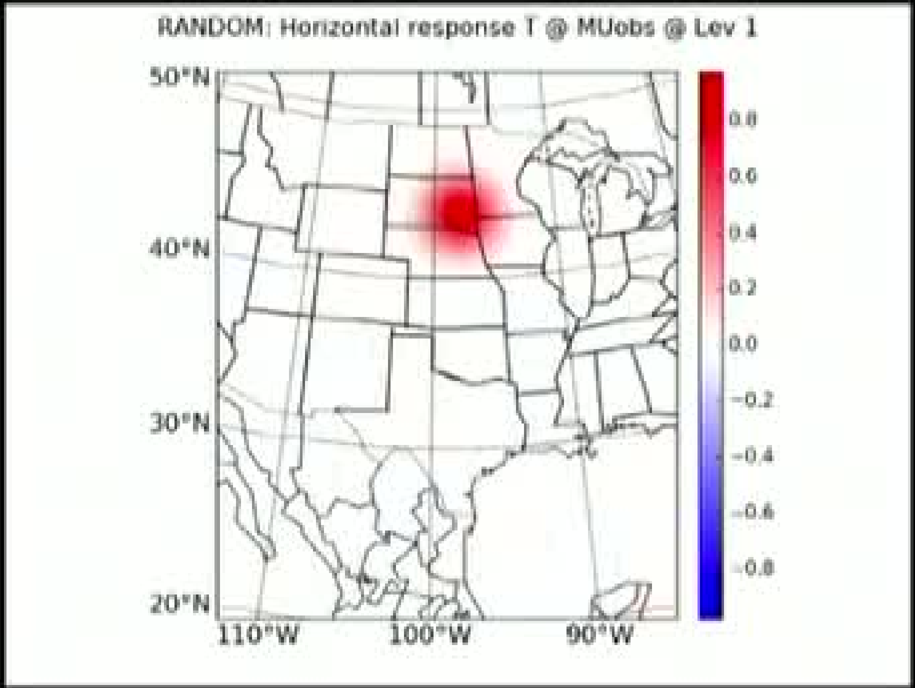
4096 ensembles

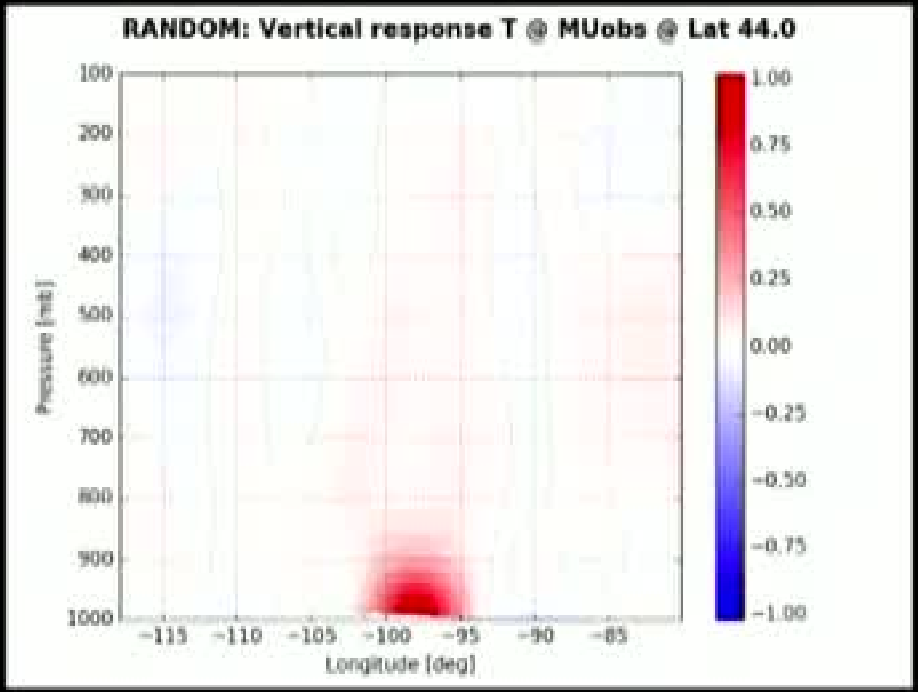


More random ensembles produce less noise

Results: random cross-covariance ($COV_{Ps\text{-}T}$)

4096 ensembles

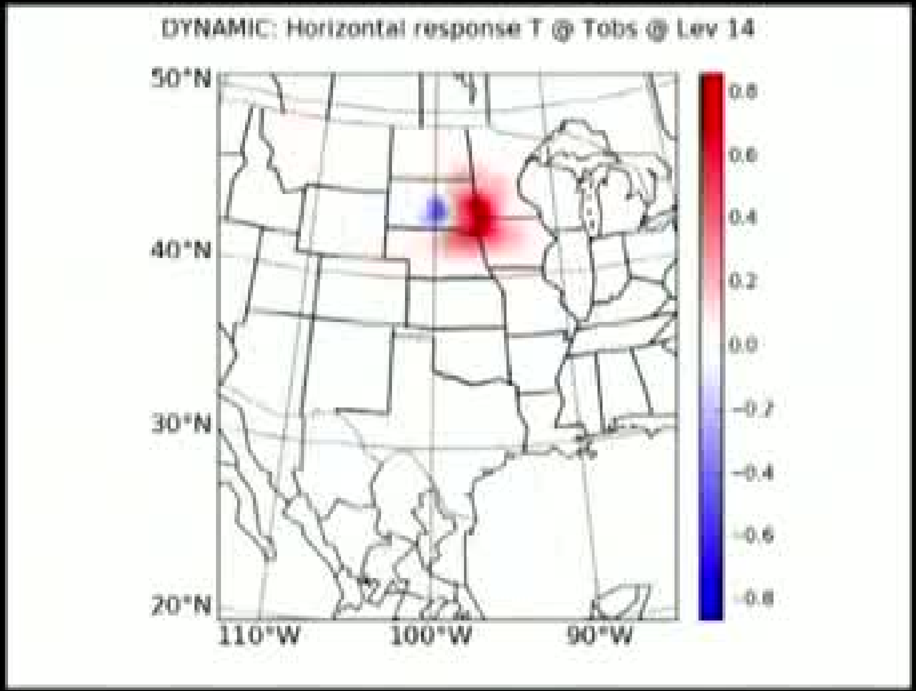


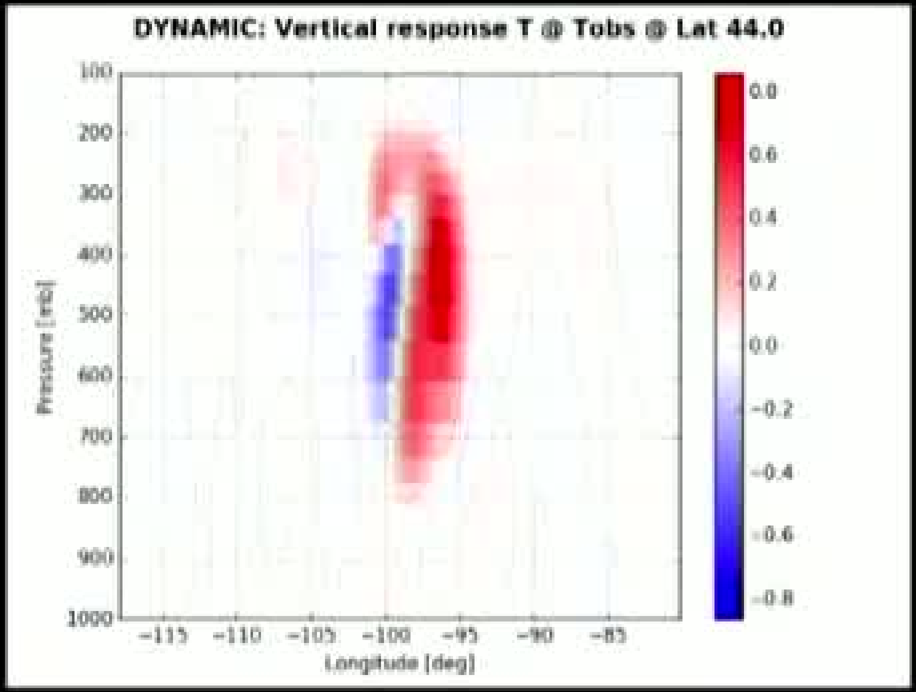


Acceptable noise in cross-covariances

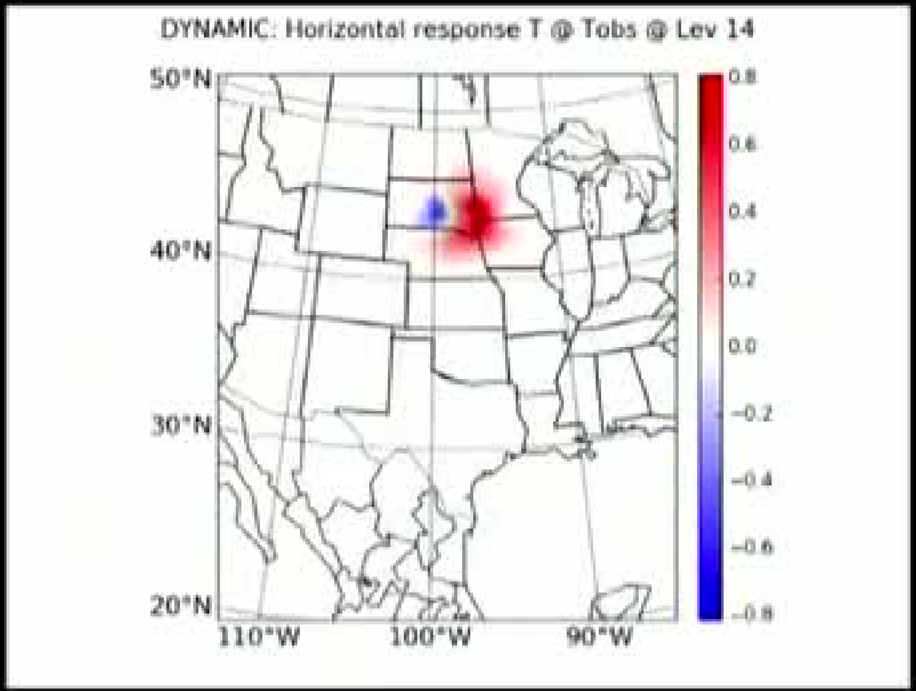
Results: total autocovariance - COV_{TT} (4096 random x 32 dynamic)

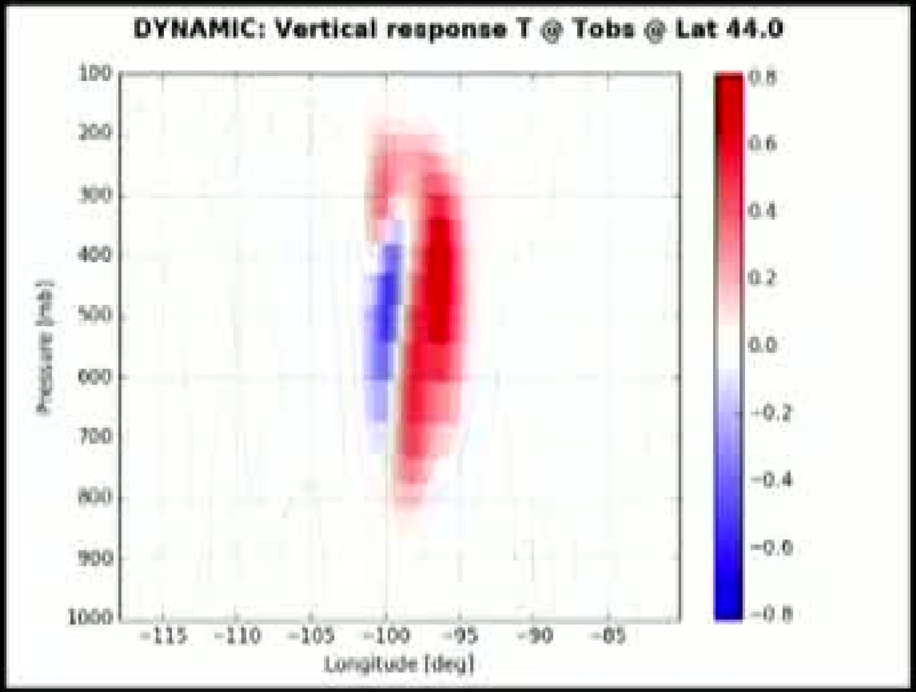
1024 ensembles





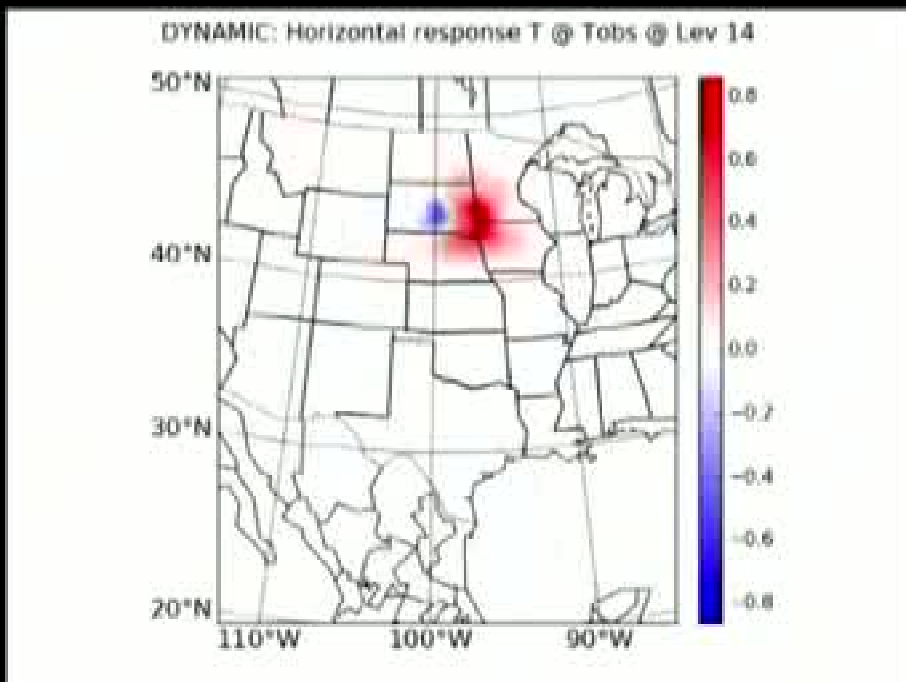
4096 ensembles

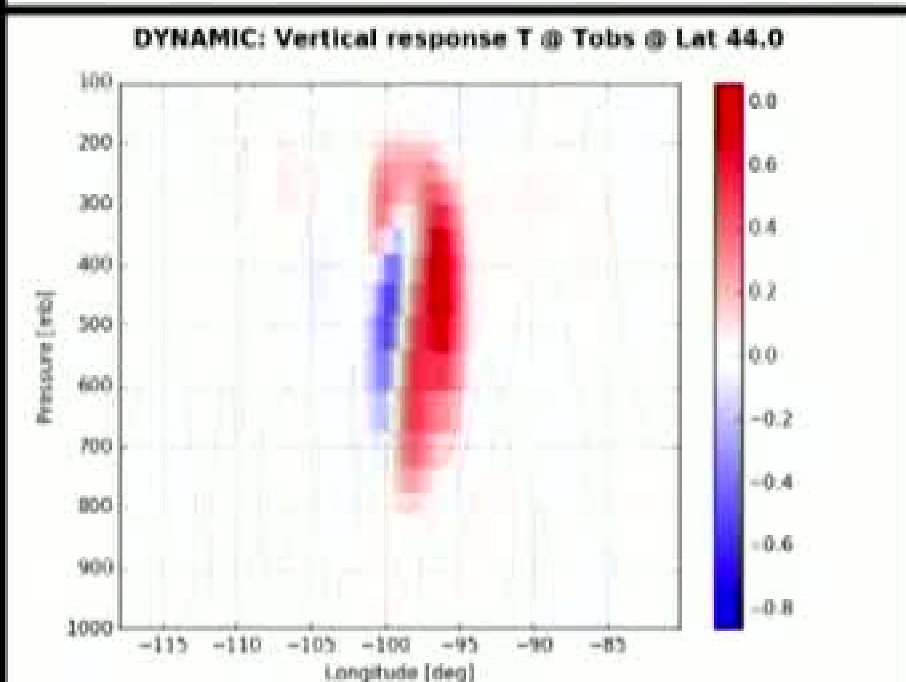




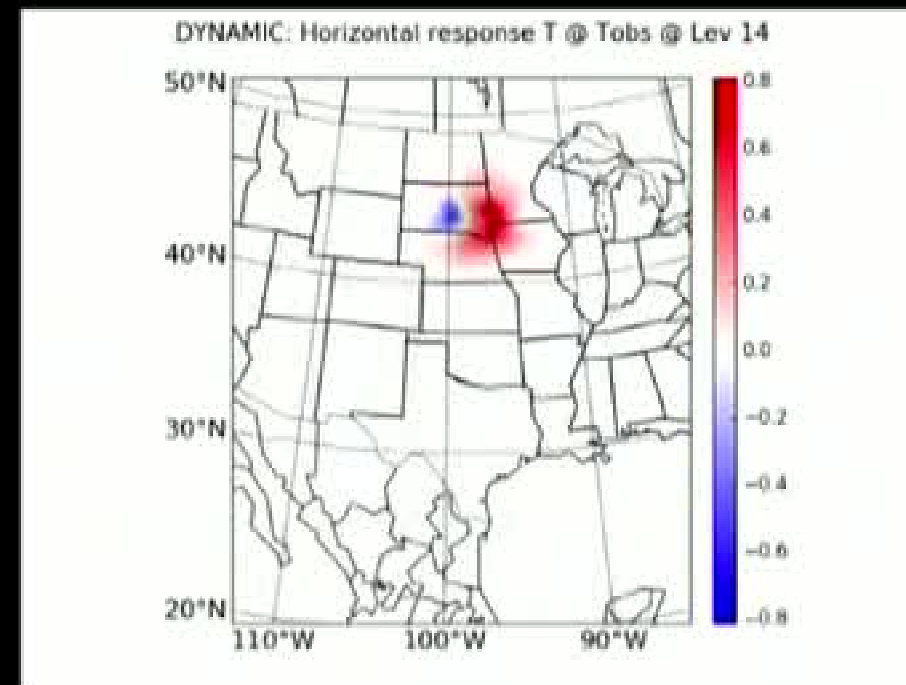
Results: total autocovariance - COV_{TT} (4096 random x 32 dynamic)

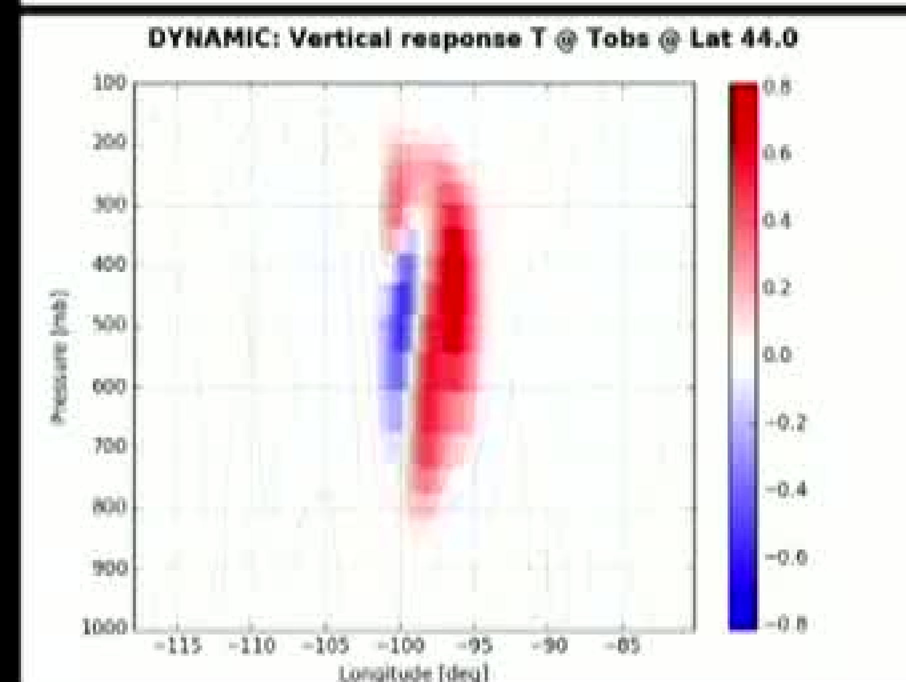
1024 ensembles





4096 ensembles

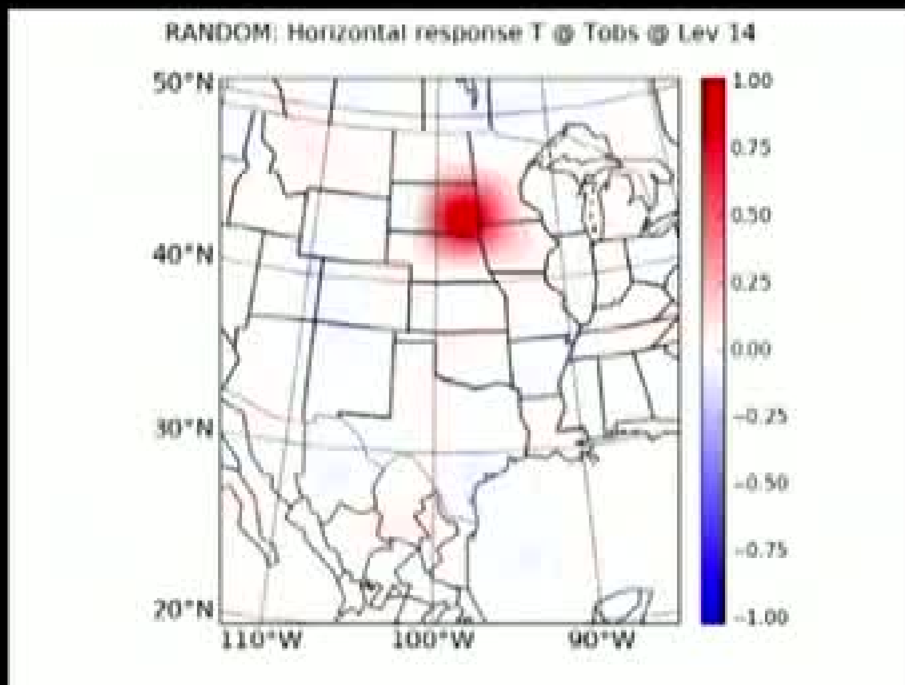


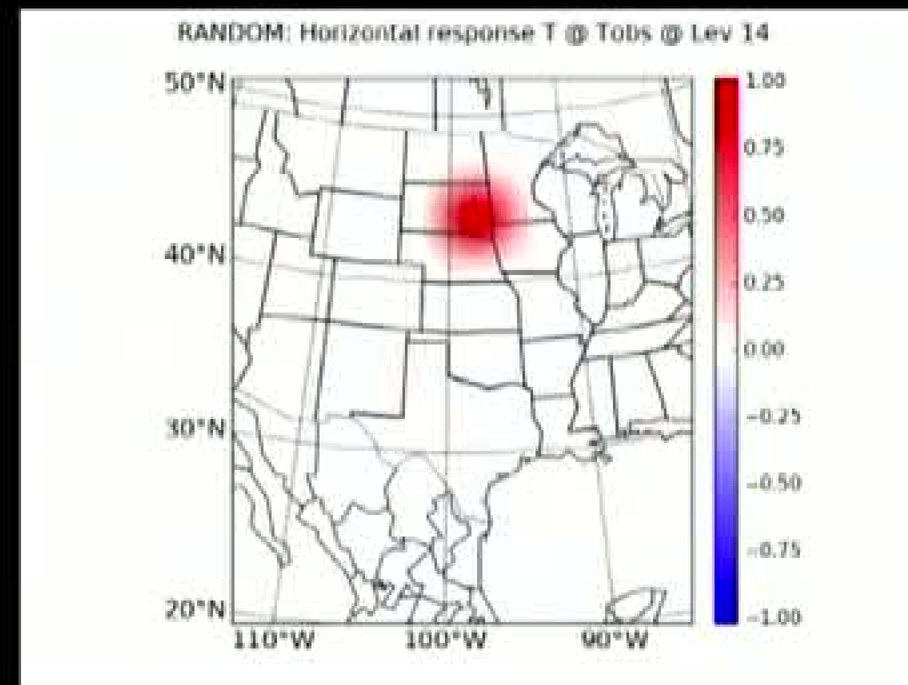


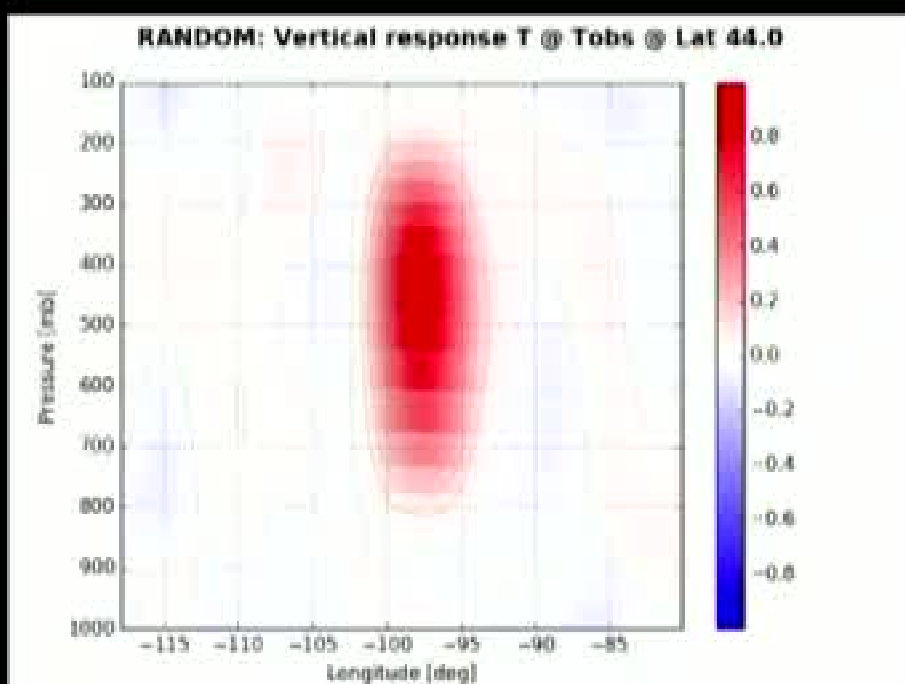
Random autocovariance ($COV_{T\text{-}T}$)

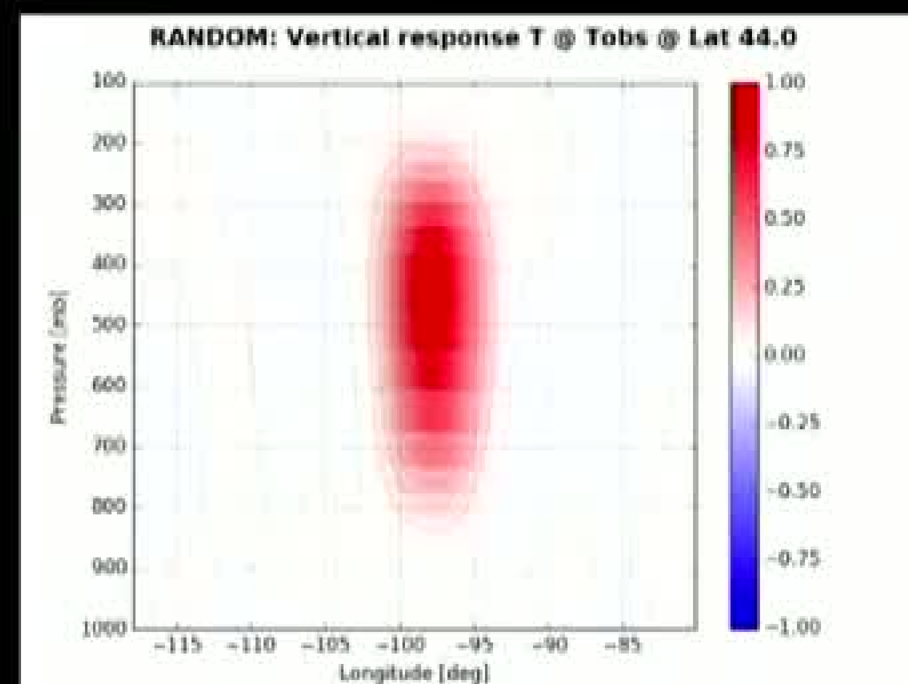
1024 ensembles

4096 ensembles









More random ensembles produce less noise

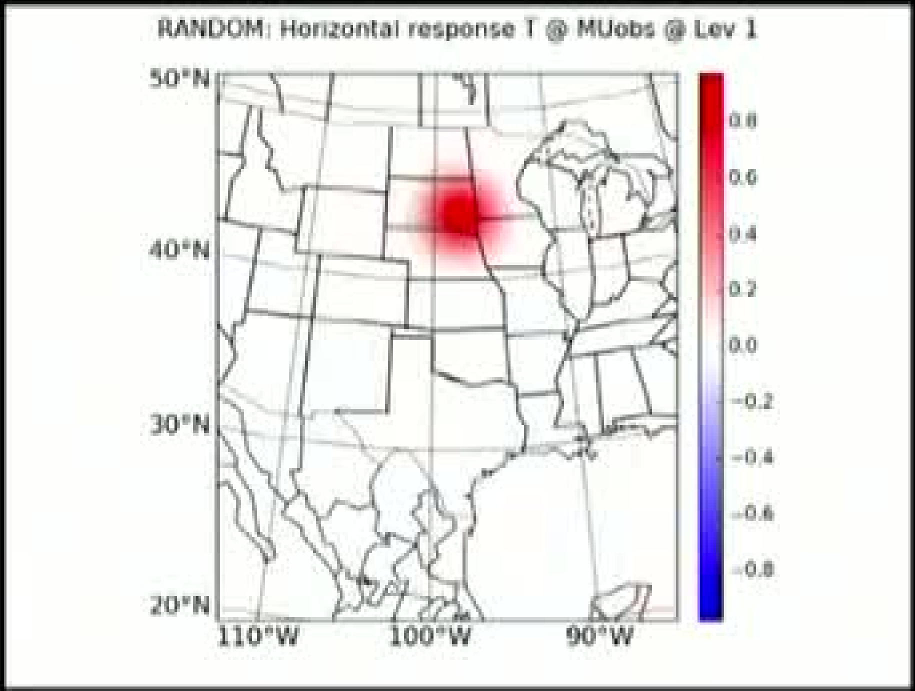
Experimental setup

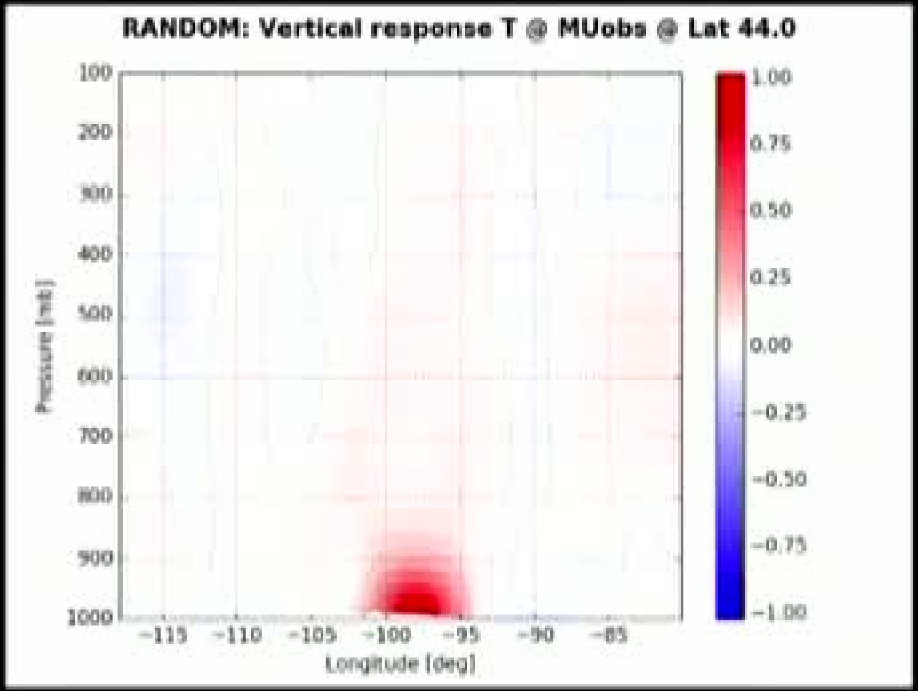
- Weather Research and Forecasting (WRF) model
- 27 km /31 layer
- 32 dynamical ensembles
- (1) 1024 and (2) 4096 random ensembles
- 6-hour forecast error covariance
- ***Random:*** use D_i=*I* in localized covariance formulation $$P_f = \frac{1}{N_r - 1}\sum_{i=1}^{N_r}(L^{1/2}\varphi_i)(L^{1/2}\varphi_i)^T$$
- ***Total:*** use complete localized covariance formulation $$P_f = \frac{1}{N_r - 1}\sum_{i=1}^{N_e}\sum_{j=1}^{N_r}\left[D_i L^{1/2}\varphi_j\right]\left[D_i L^{1/2}\varphi_j\right]^T$$
- Single observation experiment

$$x^a - x^f = (L \circ P_E)^{1/2}(L \circ P_E)^{T/2}\frac{\Delta x}{\left(\sigma_f^2 + \sigma_R^2\right)}(0 \quad \cdots \quad 1_k \quad \cdots \quad 0)^T$$

Results: random cross-covariance (COV_{Ps-T})

4096 ensembles

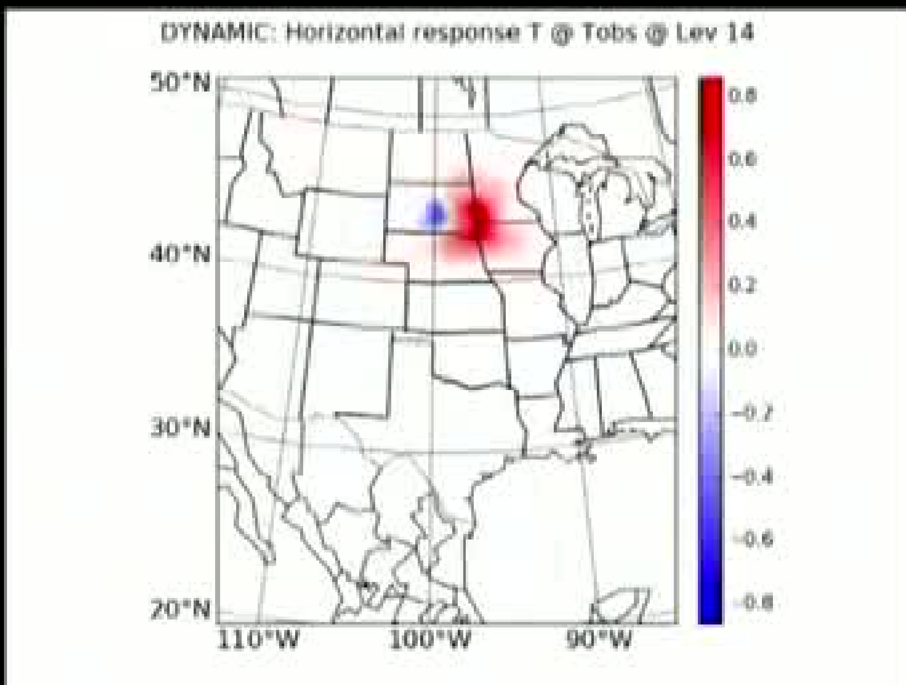


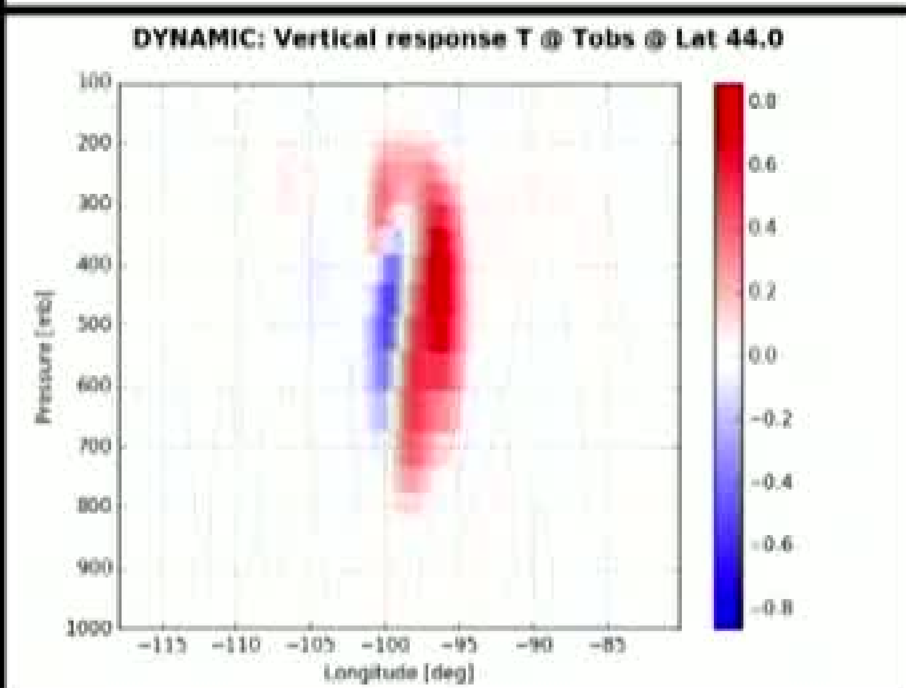


Acceptable noise in cross-covariances

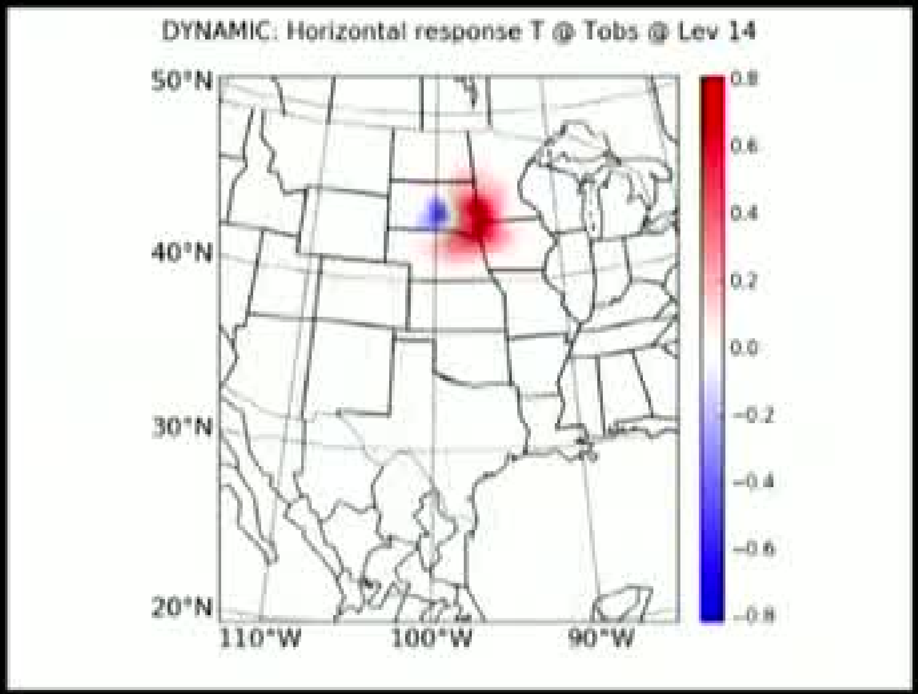
Results: total autocovariance - COV_{TT} (4096 random x 32 dynamic)

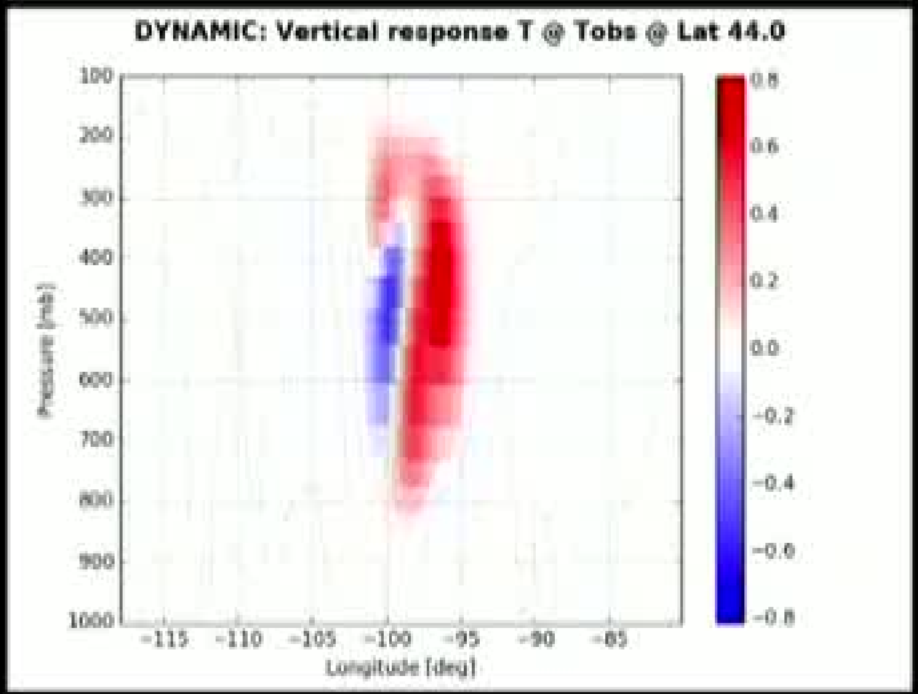
1024 ensembles





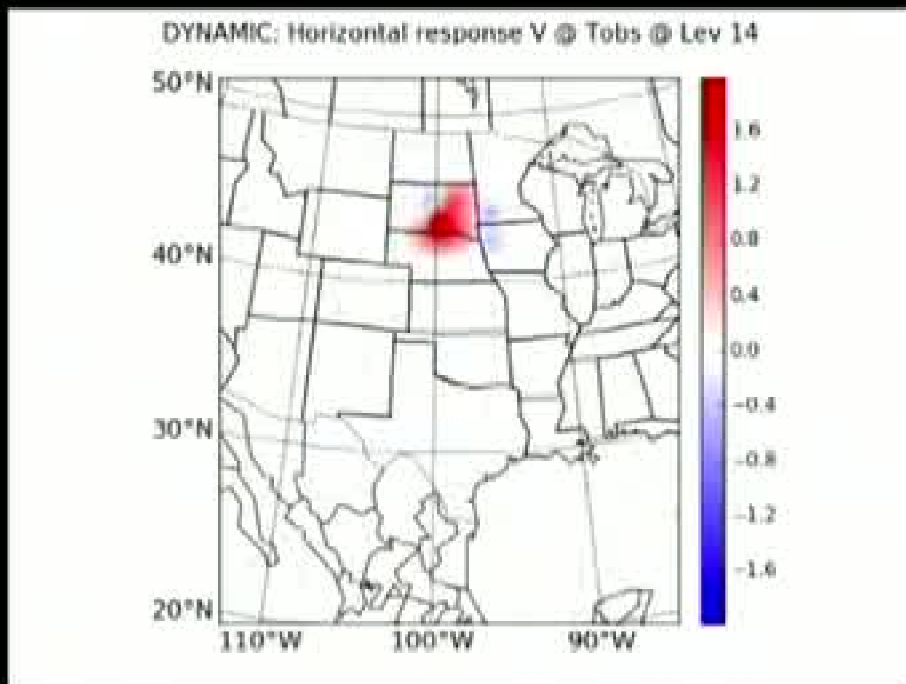
4096 ensembles

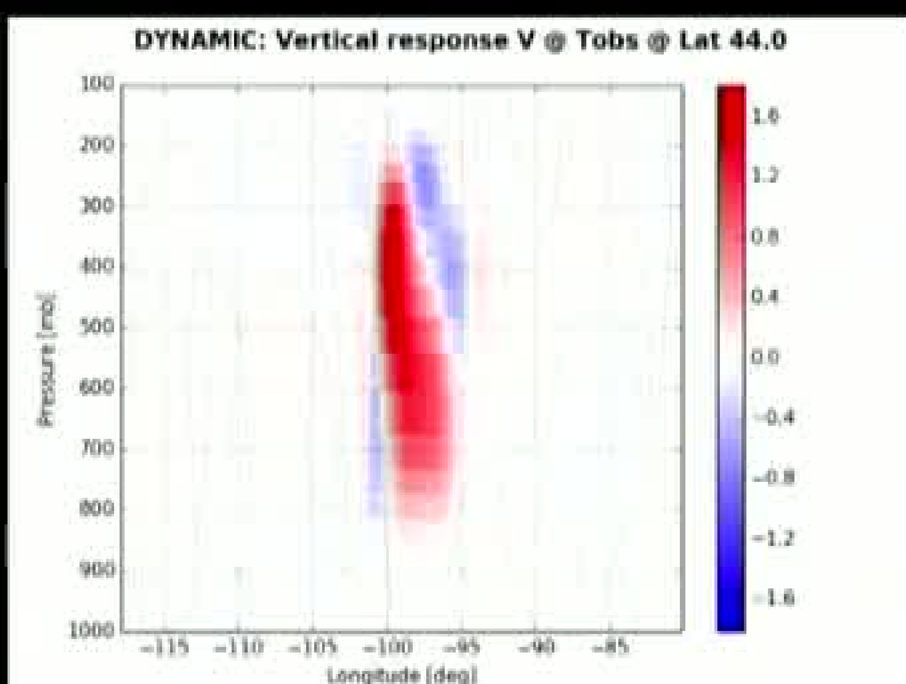




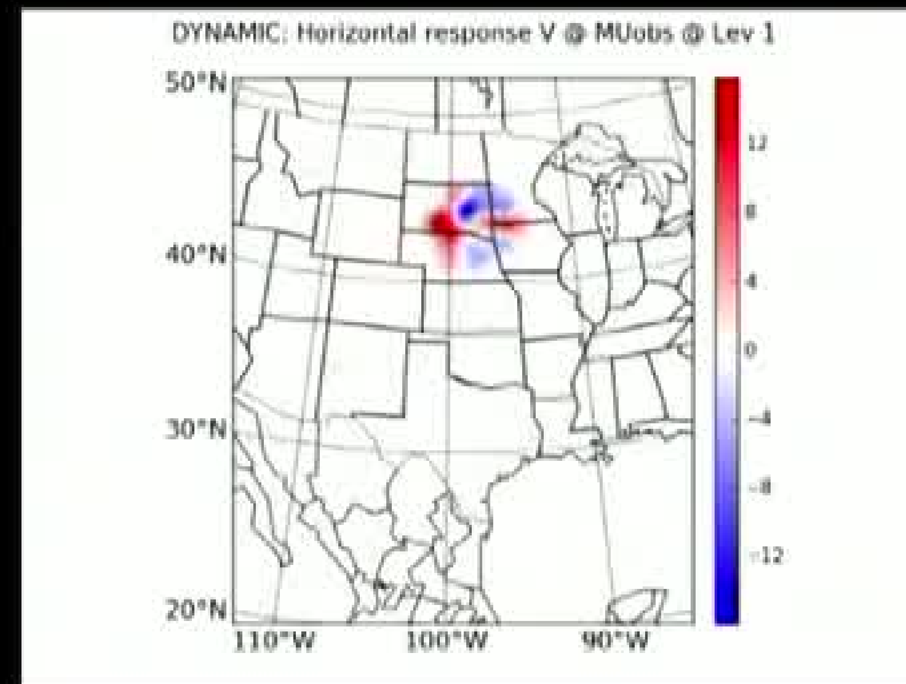
Results: total cross-covariance (4096 random x 32 dynamic)

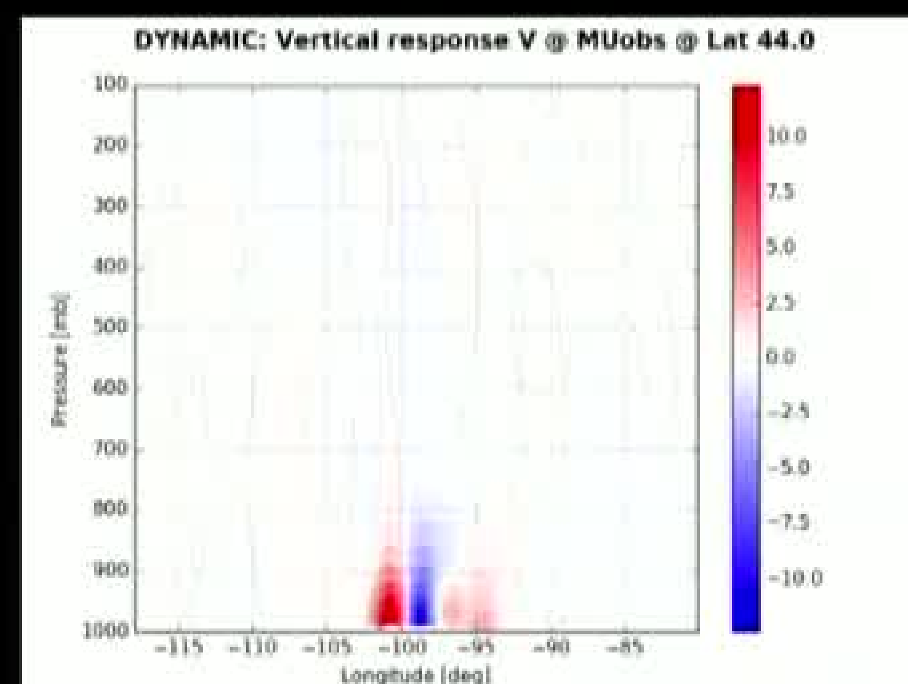
T_{obs} – V response





MU_{obs} – V response





Complex structure a consequence of dynamical covariance

Practical considerations

- *Hadamard product is expensive to calculate*
  - Localized random vectors are calculated off-line $g_j = L^{1/2}\varphi_j, (j = 1, N_r)$
  - Calculation depends on state vector specification
- *Analysis space dimension is large (random x dynamic)*
  - Number of random ensembles does not depend on state dimensions (identity random matrix only)
  - need parallel programs to process ~ $O(10^5)$ ensembles
  - optional reduced Hessian preconditioning for even faster code
- *Need only 32 + 1024 files to produce 32 x 1024 columns of covariance*
  - high efficiency
  - possibility to introduce another orthogonal basis and substitute random sample