



# EnKF overview

Jeff Whitaker*
NOAA Earth System Research Lab
<jeffrey.s.whitaker@noaa.gov>

* GSI EnVar development team: Daryl Kleist, Dave Parrish, John Derber (NCEP) and Xuguang Wang (OU)

# GSI Var cost function

$$J_{3DVAR}(\mathbf{x}') = \frac{1}{2}(\mathbf{x}')^{T}\mathbf{B}_{\mathbf{f}}^{-1}(\mathbf{x}') + \frac{1}{2}(\mathbf{H}\mathbf{x}' - \mathbf{y}')^{T}\mathbf{R}^{-1}(\mathbf{H}\mathbf{x}' - \mathbf{y}')$$

*J* : Penalty (Fit to background + Fit to observations)

**x**’ : Analysis increment ($\mathbf{x}^a - \mathbf{x}^b$) ; where $\mathbf{x}^b$ is a background

$\mathbf{B}_f$ : Background error covariance

**H** : Observations (forward) operator

**R** : Observation error covariance (Instrument + representativeness)

$\mathbf{y}' = \mathbf{y}^o - \boldsymbol{H}\mathbf{x}^b$, where $\mathbf{y}^o$ are the observations

Cost function (*J*) is minimized to find solution, **x**’ [$\mathbf{x}^a = \mathbf{x}^b + \mathbf{x}'$]

# GSI hybrid ensemble Var cost function

$$\mathbf{J}_{hybrid}\left(x'\right)=\frac{\beta}{2}\left(\mathbf{x}'\right)^{\mathrm{T}}\mathbf{B}_{\mathrm{f}}^{-1}\left(\mathbf{x}'\right)+\frac{1-\beta}{2}\left(\mathbf{x}'\right)^{\mathrm{T}}\mathbf{B}_{ens}^{-1}\left(\mathbf{x}'\right)+\frac{1}{2}\left(\mathbf{Hx}'-\mathbf{y}'\right)^{\mathrm{T}}\mathbf{R}^{-1}\left(\mathbf{Hx}'-\mathbf{y}'\right)$$

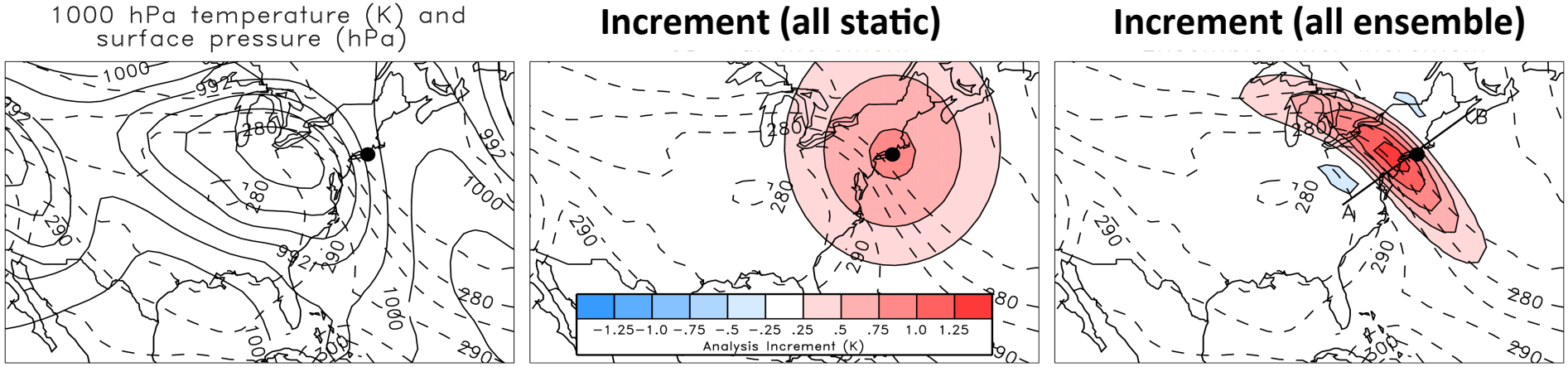
$\mathbf{B}_{\mathbf{f}}$ : (Fixed) background-error covariance (estimated offline)

$\mathbf{B}_{\mathbf{ens}}$ : (Flow-dependent) background-error covariance (estimated from ensemble)

$\beta$: Weighting factor (0.25 means total **B** is ¾ ensemble).

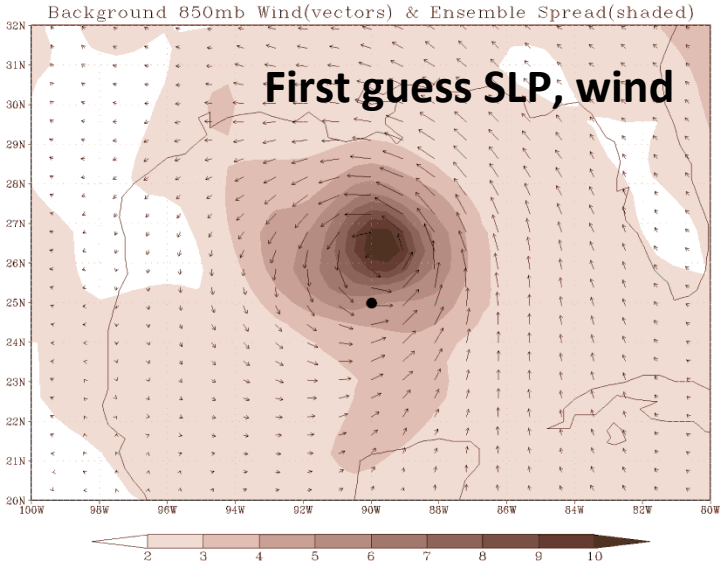
# What does $\mathbf{B_{ens}}$ do?

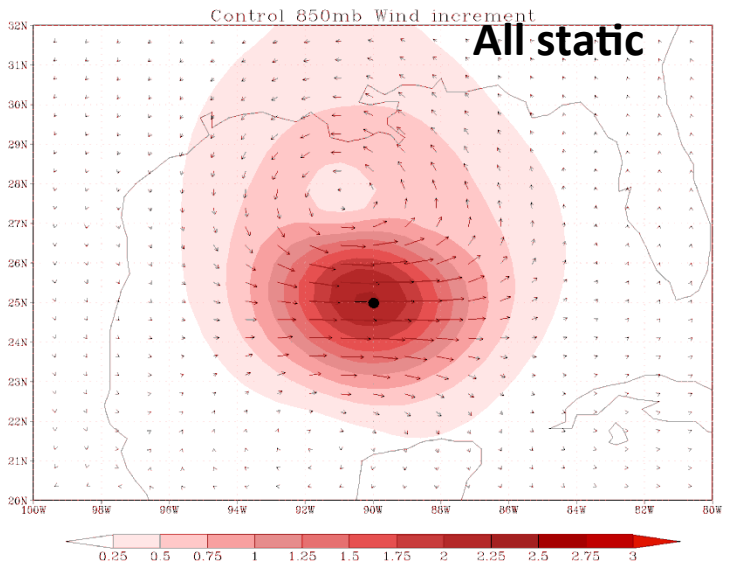
Temperature observation near a warm front

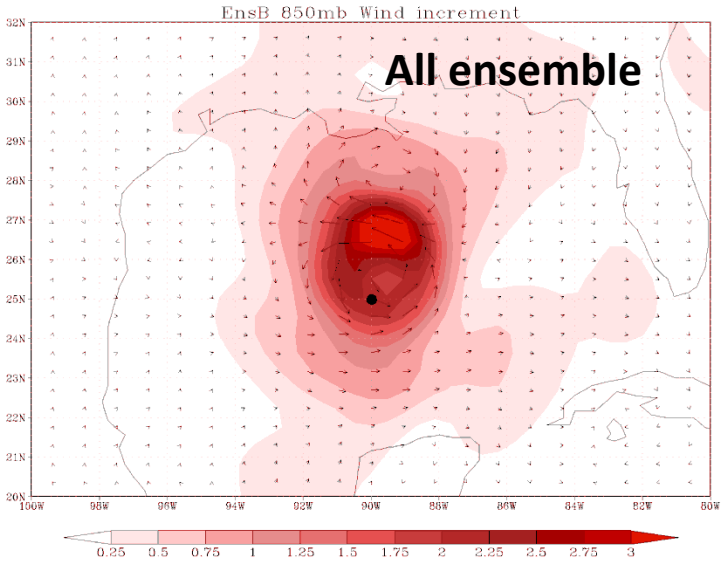


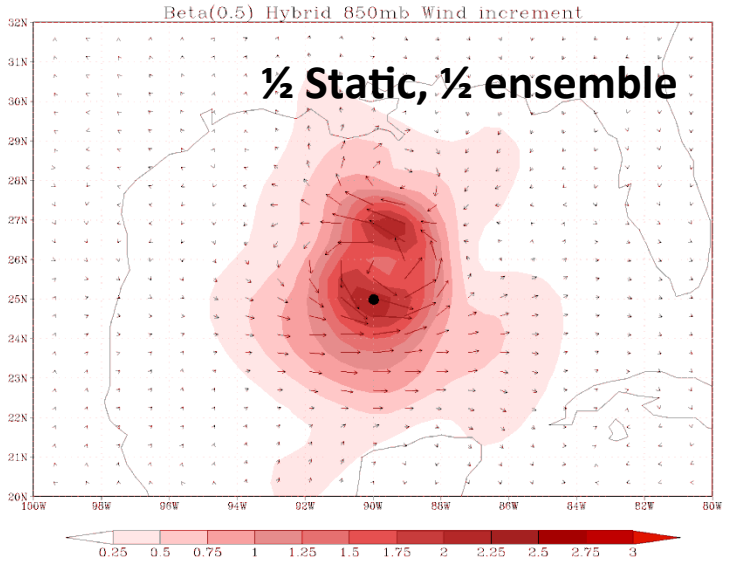
# What does $\mathbf{B_{ens}}$ do?

Zonal wind observation near a hurricane (Ike)



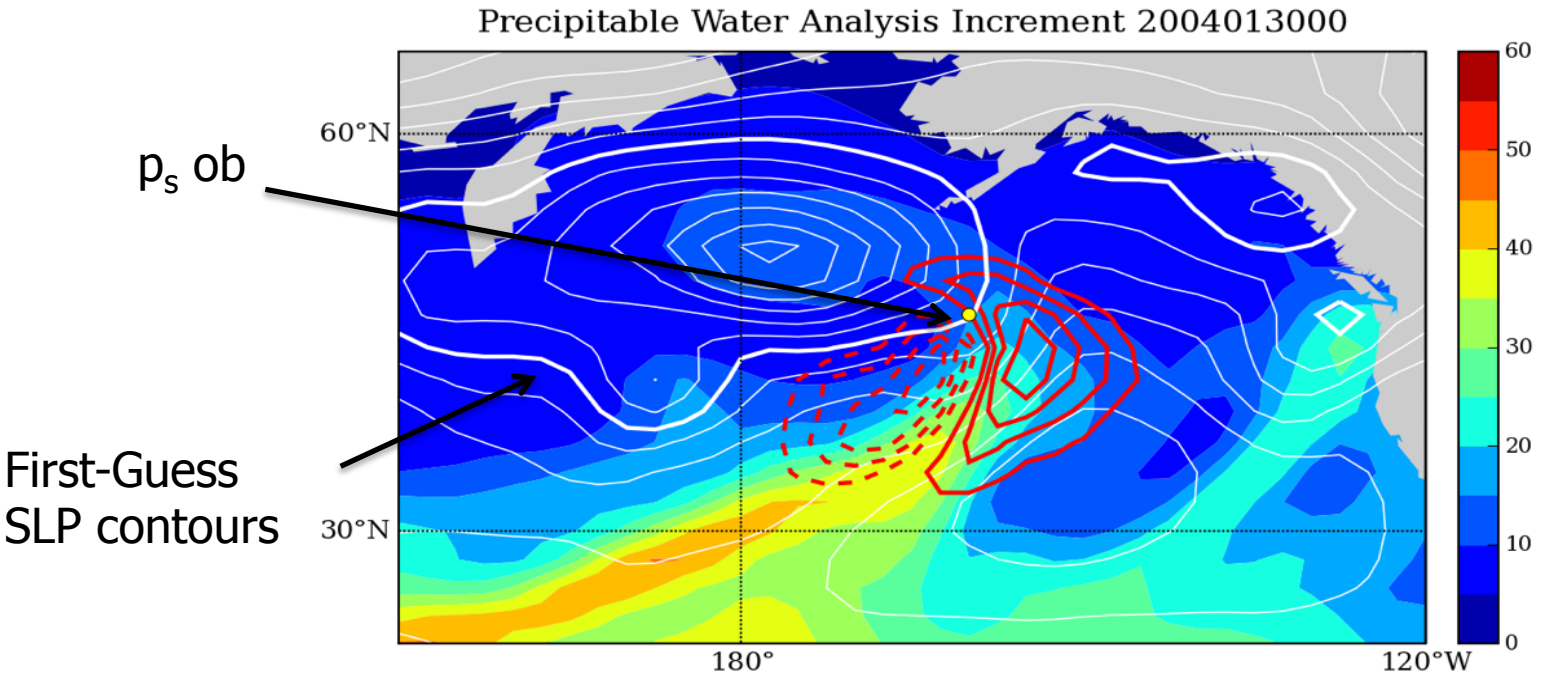



# What does $\mathbf{B_{ens}}$ do?

Surface pressure observation near an "atmospheric river"

PWAT increment




***3Dvar increment would be zero!***

***(cross-variable covariances hard to model with static*** $\mathbf{B_f}$***)***

# What does $\mathbf{B}_{\mathbf{ens}}$ do?

- Adds flow-dependence to analysis increments.
- Sparse observations near coherent dynamical features used more effectively.
- Changes in the observing network can be captured in background-error variance.
- ***More information extracted from observations => More skillful forecasts***

# So what's the catch?

- Need an ensemble (fairly large) that accurately represents the uncertainty in the first-guess forecast.
- "Fairly large" means O(50-100) -- smaller ensembles with have large sampling errors (and more weight will have to be given to $\mathbf{B}_f$). Expensive to run.
- The GSI variational system does not provide the ensemble – it provides an analysis that can be interpreted as the ensemble mean, given an ensemble that represents forecast uncertainty.
- In NCEP operations, an "Ensemble Kalman Filter" (EnKF) is used to generate the background ensemble.

# Data assimilation terminology

- **y** : Observation vector (weather balloons, satellite radiances, etc.)
- **x** : the state of the atmosphere as represented by the model
- $\mathbf{x}^b$ : Background state vector ("prior")
- $\mathbf{x}^a$ : Analysis state vector ("posterior")
- **H** : (hopefully linear) operator to convert model state → observation location & type
- **R** : Observation - error covariance matrix
- $\mathbf{P}^b$ : Background - error covariance matrix
- $\mathbf{P}^a$ : Analysis - error covariance matrix

# From Bayes theorem to 4DVar and the (Ensemble) Kalman Filter

$$p(\mathbf{x}|\mathbf{y}) \propto \exp\left(-(\mathbf{x}-\mathbf{x}_\mathrm{b})^\mathrm{T}\mathbf{P}^{\mathrm{b}\,-1}(\mathbf{x}-\mathbf{x}_\mathrm{b}) - (\mathbf{y}-\mathbf{H}\mathbf{x})^\mathrm{T}\mathbf{R}^{-1}(\mathbf{y}-\mathbf{H}\mathbf{x})\right)$$

Variational methods maximize the posterior PDF to find the state trajectory **x** that best fits the obs **y** in a least-squares sense. In practice, this is done by minimizing a cost function, which is what's inside the *exp*:

$$J(\mathbf{x}) \propto (\mathbf{x}-\mathbf{x}_\mathrm{b})^\mathrm{T}\mathbf{P}^{\mathrm{b}\,-1}(\mathbf{x}-\mathbf{x}_\mathrm{b}) + (\mathbf{y}-\mathbf{H}\mathbf{x})^\mathrm{T}\mathbf{R}^{-1}(\mathbf{y}-\mathbf{H}\mathbf{x})$$

The minimum can be found analytically if **H** is linear (see Lorenc 1986 *QJRMS* for the algebra). This gives the equations for the Kalman Filter

$$\mathbf{x}_\mathrm{a} = \mathbf{x}_\mathrm{b} + \mathbf{K}\left(\mathbf{y}-\mathbf{H}\mathbf{x}_\mathrm{b}\right),\ \mathbf{P}^\mathrm{a} = \left(\mathbf{I}-\mathbf{K}\mathbf{H}\right)\mathbf{P}^\mathrm{b}$$
$$\mathbf{K} = \mathbf{P}^\mathrm{b}\mathbf{H}^\mathrm{T}\left(\mathbf{H}\mathbf{P}^\mathrm{b}\mathbf{H}^\mathrm{T}+\mathbf{R}\right)^{-1}$$

- *Matrix* $\mathbf{P}^\mathrm{b}$ *is too big for any computer, covariance update step impractical.*
- *Instead, represent PDFs of* **x** *and* **y** *by an ensemble, compute sample estimate of* $\mathbf{P}^\mathrm{b}$ *and* $\mathbf{x}_\mathrm{b}$*. Evolve the sample, not the full covariance.* ***EnKF*** *gives same result as full KF if ensemble size becomes infinite.*

# Computational shortcuts in EnKF:
## (1) Simplifying Kalman gain calculation

$$\mathbf{K} = \mathbf{P}^{\mathrm{b}} H^{\mathrm{T}} \left( H \mathbf{P}^{\mathrm{b}} H^{\mathrm{T}} + \mathbf{R} \right)^{-1}$$

$$define \quad \overline{H\mathbf{x}^{\mathrm{b}}} = \frac{1}{m} \sum_{i=1}^{m} H\mathbf{x}_{\mathrm{i}}^{\mathrm{b}}$$

$$\mathbf{P}^{\mathrm{b}} H^{\mathrm{T}} = \frac{1}{m-1} \sum_{i=1}^{m} \left( \mathbf{x}_{\mathrm{i}}^{\mathrm{b}} - \overline{\mathbf{x}^{\mathrm{b}}} \right) \left( H\mathbf{x}_{\mathrm{i}}^{\mathrm{b}} - \overline{H\mathbf{x}^{\mathrm{b}}} \right)^{\mathrm{T}}$$

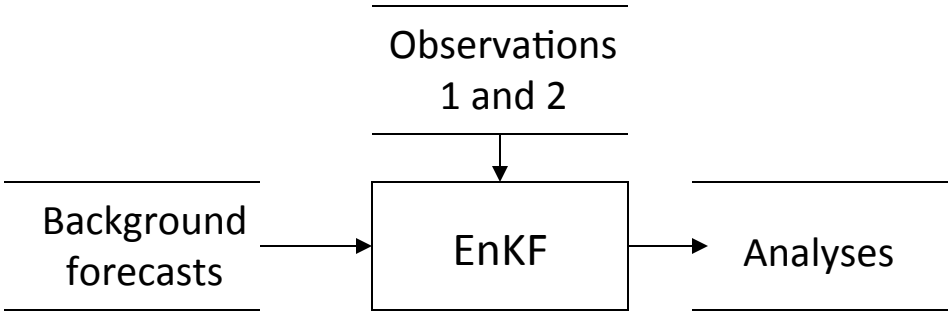
$$H\mathbf{P}^{\mathrm{b}} H^{\mathrm{T}} = \frac{1}{m-1} \sum_{i=1}^{m} \left( H\mathbf{x}_{\mathrm{i}}^{\mathrm{b}} - \overline{H\mathbf{x}^{\mathrm{b}}} \right) \left( H\mathbf{x}_{\mathrm{i}}^{\mathrm{b}} - \overline{H\mathbf{x}^{\mathrm{b}}} \right)^{\mathrm{T}}$$

The key here is that the huge matrix $\mathbf{P}^{\mathrm{b}}$ is never explicitly formed

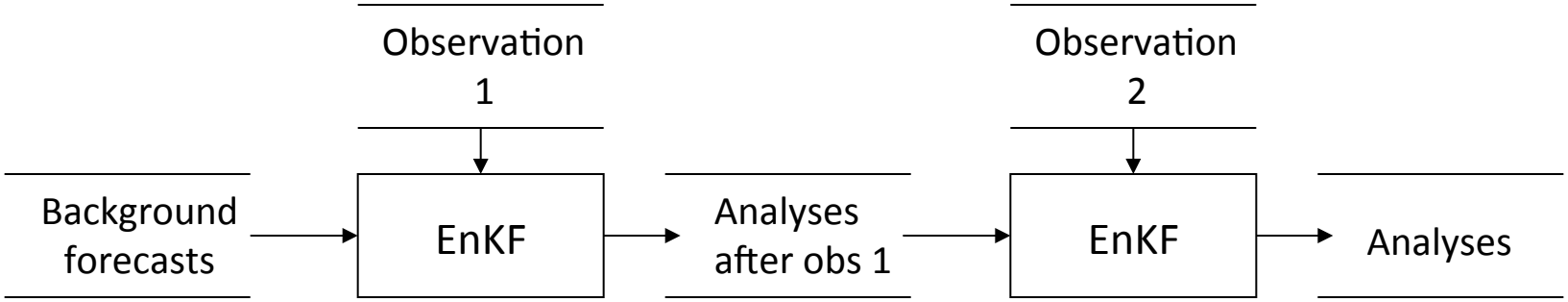
# Computational shortcuts in EnKF:

## (2) serial processing of observations (requires observation error covariance **R** to be diagonal)

Method 1




Method 2



# The serial EnKF – a recipe

Given a single ob $y^o$ with expected error variance R, an ensemble of model forecasts $\mathbf{x}^b$ (model priors), and an ensemble of predicted observations $\mathbf{y}^b = \mathbf{H}\mathbf{x}^b$ (observation priors):

***Step 1***: Update observation priors.

(1a) $\overline{\mathrm{y}}_\mathrm{a} = (1-\mathrm{K})\overline{\mathrm{y}}_\mathrm{b} + \mathrm{K}\mathrm{y}^\mathrm{o}$ *update for ob prior means*

(1b) $\mathbf{y}'_\mathrm{a} = \sqrt{(1-\mathrm{K})}\mathbf{y}'_\mathrm{b}$ *rescaling of ob prior perturbations*

where the scalar K = var($\mathbf{y}^b$)/(var($\mathbf{y}^b$)+ R), overbar denotes means, prime denotes perturbations, superscript *b* denotes prior, *a* denotes analysis.

***Linear interpolation between observation and observation prior mean*** *with weight K (0<=K<=1), rescaling of observation prior ensemble so posterior variance is consistent with Kalman filter, i.e.* var($\mathbf{y}^a$)=(1-K) var($\mathbf{y}^b$)= var($\mathbf{y}^b$)R/(var($\mathbf{y}^b$)+R).

when var($\mathbf{y}^b$) << R, all weight given to prior.

when var($\mathbf{y}^b$) >> R, all weight given to observation.

# The serial EnKF – a recipe (2)

***Step 2***: Update model priors.

Let $\Delta\mathbf{x} = \mathbf{x}^a - \mathbf{x}^b$ be analysis increment for model priors, $\Delta\mathbf{y} = \mathbf{y}^a - \mathbf{y}^b$ is analysis increment for observation priors.

(2) $\Delta\mathbf{x} = \mathbf{G}\Delta\mathbf{y}$ *computation of increments to model prior*

where $\mathbf{G} = \text{cov}(\mathbf{x}^b, \mathbf{y}^{bT})/\text{var}(\mathbf{y}^b)$

***Linear regression of model priors on observation priors.***

Only changes model priors when $\mathbf{x}^b$ and $\mathbf{y}^b$ are correlated within the ensemble.

If there is more than one ob to be assimilated, the observation priors for other (not yet assimilated) obs ($\mathbf{Y}^b$) should be also be updated using (2) with $\Delta\mathbf{x}$ replaced by $\Delta\mathbf{Y}$. Next iteration, replace $\mathbf{y}^b$ with next column of $\mathbf{Y}^b$, removing that column from $\mathbf{Y}^b$. After each iteration the model priors and observation priors are set to the latest analysis values ($\mathbf{x}^a$ replaces $\mathbf{x}^b$, $\mathbf{Y}^a$ replaces $\mathbf{Y}^b$). Continue iterating until $\mathbf{Y}^b$ is empty.

# Factors limiting EnKF performance

## *1) Treatment of model error*

Must account for the background error covariance associated with "model error" (any difference between simulated and true environment). Methods used so far:

1) multiplicative inflation (mult. ens perts by a factor > 1).
2) additive inflation (random perts added to each member – e.g. differences between 24 and 48-h forecasts valid at the same time).
3) model-based schemes (e.g. stochastic kinetic energy backscatter for representing unresolved processes, stochastically perturbed physics tendecies for representing parameterization uncertainty).

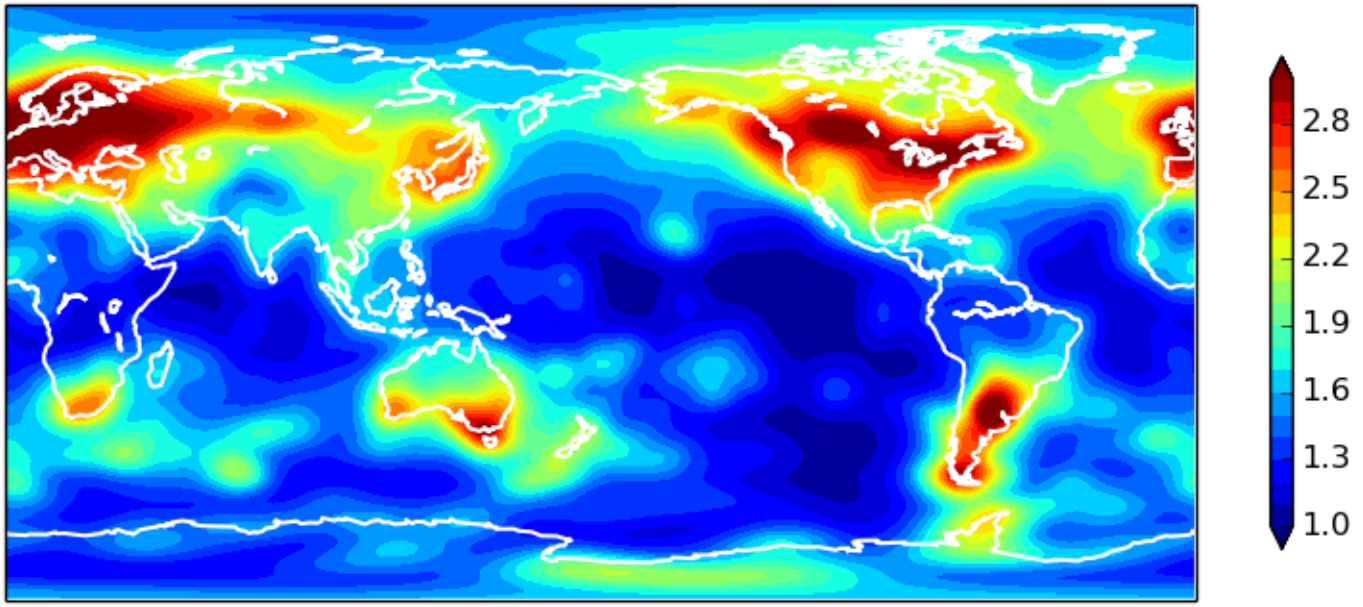
***Opnl NCEP system uses a combination all three.***

# Relaxation To Prior Spread (RTPS) Inflation

*Described in DOI: 10.1175/MWR-D-11-00276.1*

Inflate posterior spread (std. dev) $\sigma^a$ back toward prior spread $\sigma^b$ : $\sigma^a \leftarrow \alpha\sigma^b + (1-\alpha)\sigma^a$

Equivalent to: $x'^a \leftarrow x'^a [1+\alpha(\sigma^b-\sigma^a)/\sigma^a]$

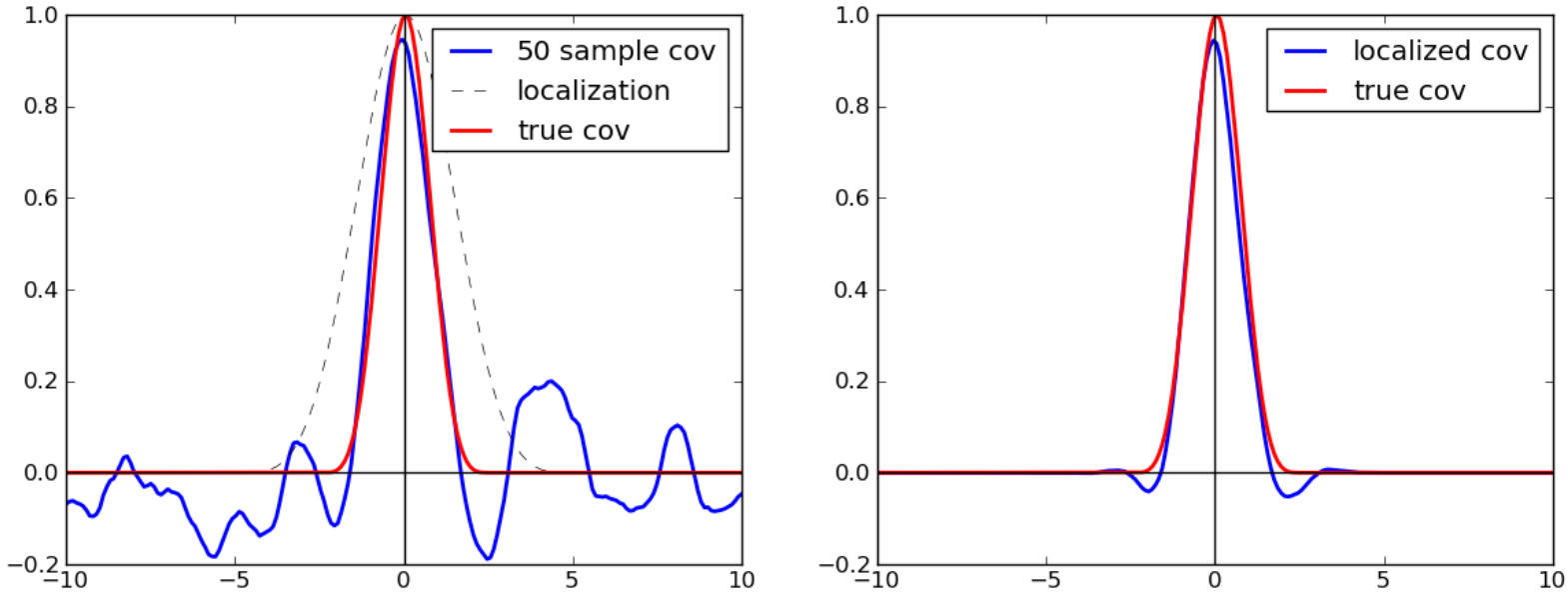


# Factors limiting EnKF performance

## *2) Treatment of sampling error (localization)*

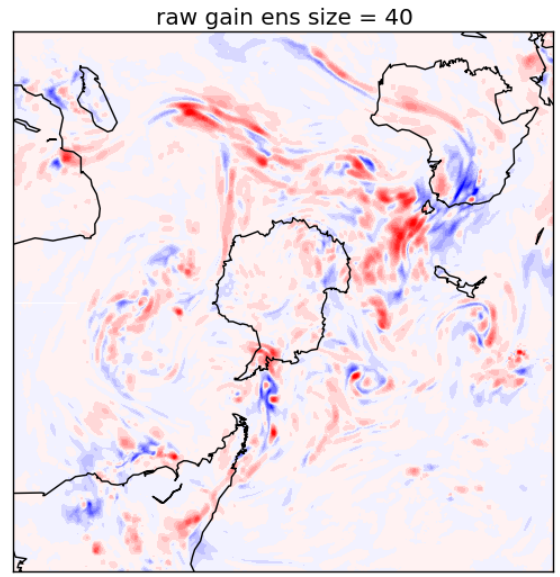
- All EnKF implementations localize the spatial impact of observations on the model state.
- Done by spatially modulating covariance between obs. prior and model state, or by only using observations 'close' to a model state variable to update that variable.
- Needed to account for low rank of ensemble (compared to model state).
- Methods used currently are not flow dependent, and assume there is no sampling error at ob location.

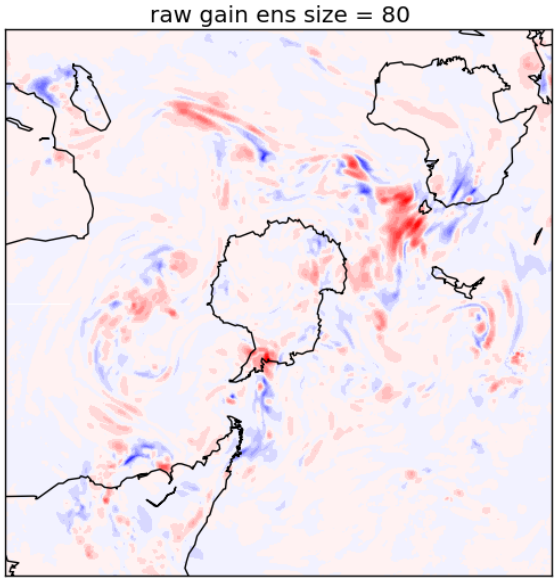
# A simple example of covariance localization

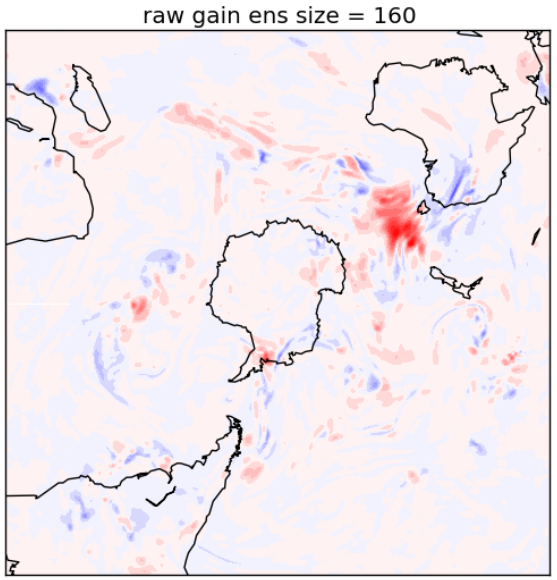


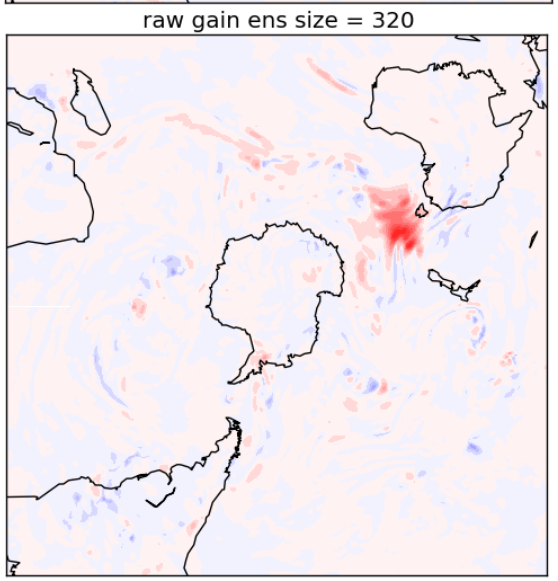

Estimates of covariances from a small ensemble will be noisy, with signal-to-noise small especially when covariance is small

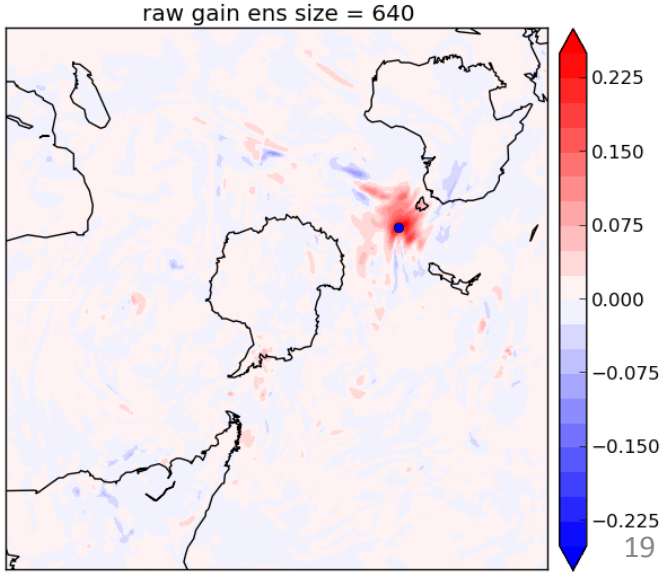
# Covariance localization



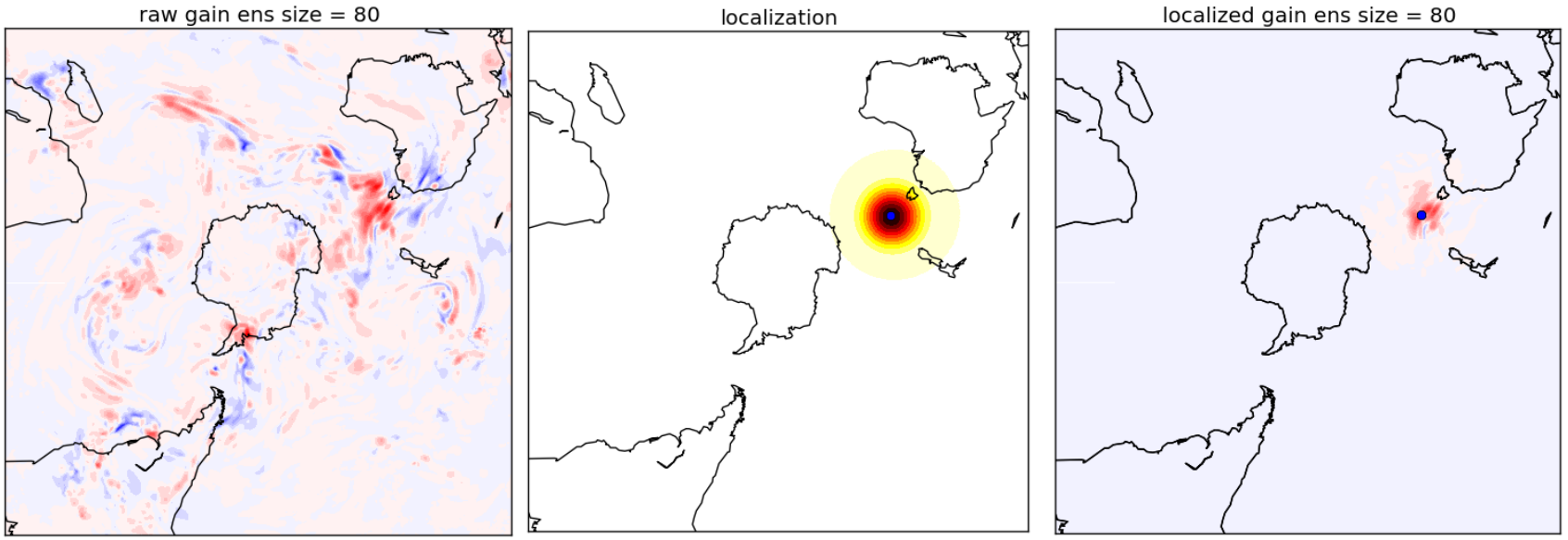




- AMSUA_n15 channel 6 radiance at 150E,-50S.
- Increment to level 30 (~310mb) temperature for a 1K O-F for 40,80,160,320 and 640 ens members with no localization.
- Ens generated by running with stochastic physics from operational analysis.

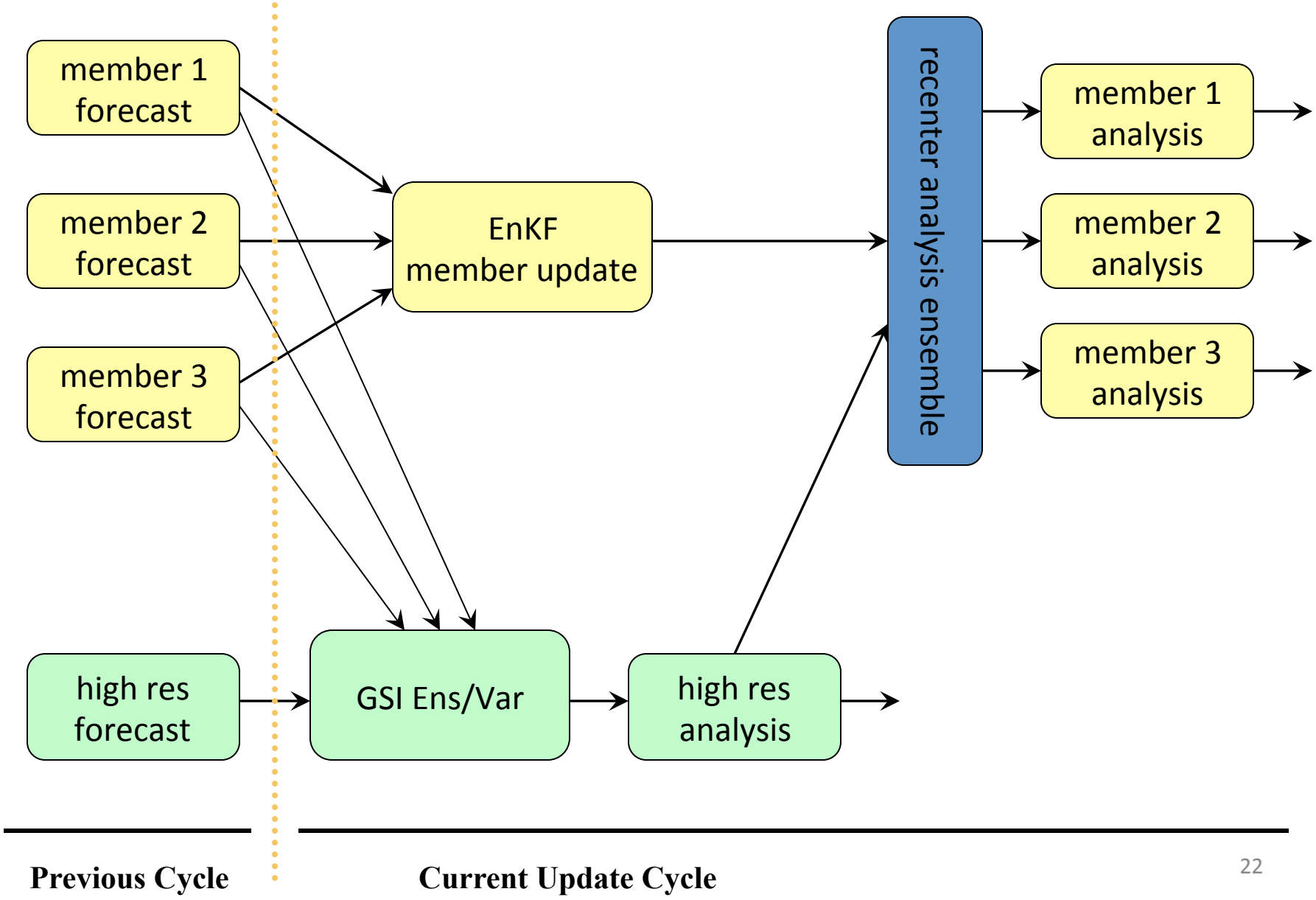
# Covariance Localization



# Why combine EnKF and Var?

| Features from EnKF | Features from Var |
|---|---|
| Can propagate $\mathbf{P}^b$ from across assimilation windows | Treatment of sampling error in ensemble $\mathbf{P}^b$ estimate does not depend on **H**. |
| More flexible treatment of model error (can be treated in ensemble) | Dual-resolution capability – can produce a high-res "control" analysis. |
| Automatic initialization of ensemble forecasts. | Ease of adding extra constraints to cost function, including a static $\mathbf{P}^b$ component. |

# Ensemble-Var workflow



# Summary

- The EnKF uses an ensemble of first-guess forecasts to estimate the background-error covariance. Every ensemble member is updated at each analysis time.
  - Parallel code, scalable out to O(1000's) of processors as long as number of obs << number of state vars.
  - Requires state vector in model and ob space, plus obs, as input.
  - GSI used to compute forward observation operator (separate step run before EnKF).
- Need to carefully tune localization length scales (depends on model resolution, observing network).
- Ensemble (co)variances must be representative of control forecast error. Treatment of model and sampling error is crucial.

# GSI ensemble 3DVar cost function (with localization)

$$\mathbf{J}_{\mathbf{hybrid}}\left(\mathbf{x}'\right)=\frac{\beta}{2}\left(\mathbf{x}'\right)^{\mathrm{T}}\mathbf{B}_{\mathrm{f}}^{-1}\left(\mathbf{x}'\right)+\frac{1-\beta}{2}\left(\mathbf{x}'\right)^{\mathrm{T}}\left(\mathbf{B}\circ\mathbf{S}\right)_{ens}^{-1}\left(\mathbf{x}'\right)+\frac{1}{2}\left(\mathbf{H}\mathbf{x}'-\mathbf{y}'\right)^{\mathrm{T}}\mathbf{R}^{-1}\left(\mathbf{H}\mathbf{x}'-\mathbf{y}'\right)$$

$\mathbf{B_f}$ : (Fixed) background-error covariance (estimated offline)

$\mathbf{B_{ens}}$ : (Flow-dependent) background-error covariance (estimated from ensemble). ***Schur product with correlation matrix S implies localization.***

$\beta$: Weighting factor (0.25 means total **B** is ¾ ensemble).

*Extra parameters control horizontal and vertical scales in* **S**.