GSI Hybrid Data Assimilation

Jeff Whitaker*
<jeffrey.s.whitaker@noaa.gov>

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

What is "hybrid DA"?

Ingredients of a ensemble-Var hybrid system

1. A forecast model.
2. An existing Var (3 or 4d) DA system (such as GSI).
3. A method of generating ensembles of first-guess forecasts that accurately represents forecast uncertainty (an EnKF DA system).

*The Var "cost function" is modified to use an ensemble estimate of the background-error covariance matrix **B** (in the " J_B term")*

GSI 3DVar cost function

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

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

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

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

H : Observations (forward) operator

R : Observation error covariance (Instrument + representativeness)

$\mathbf{y}' = \mathbf{y}^{\mathrm{o}} - \boldsymbol{H}\mathbf{x}^{\mathrm{b}}$, 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 ensemble 3DVar cost function

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

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

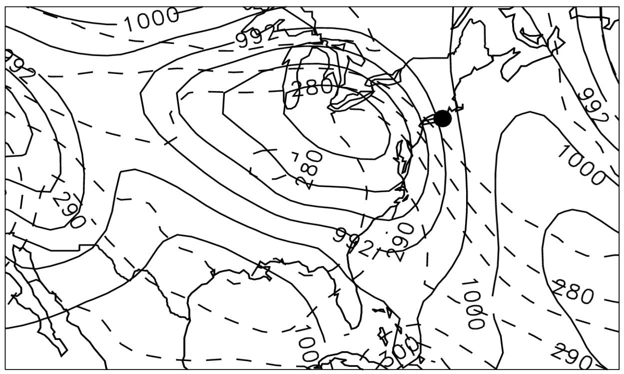
$\mathbf{B}_{ens}$: (Flow-dependent) background-error covariance (estimated from ensemble)

β: Weighting factor (0.25 means total **B** is ¾ ensemble).

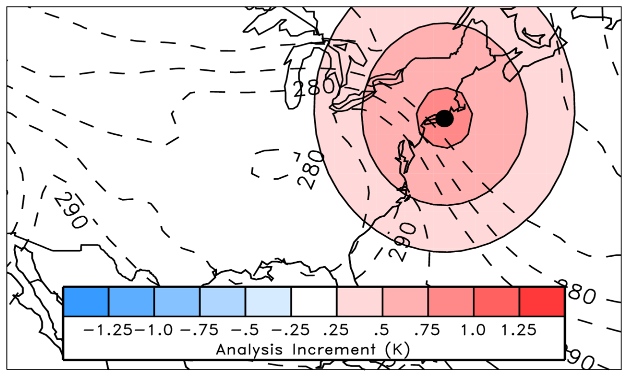
What does $\mathbf{B}_{\mathbf{ens}}$ do?

Temperature observation near a warm front

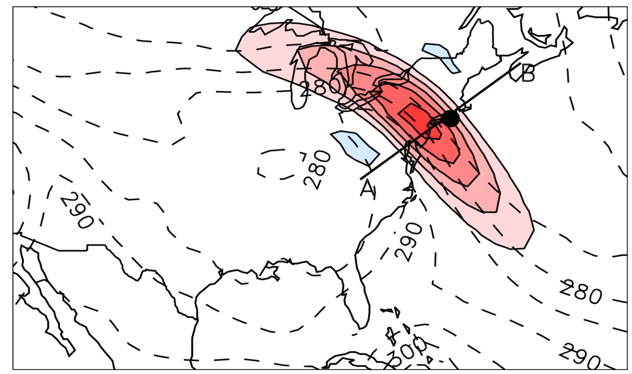
1000 hPa temperature (K) and surface pressure (hPa)



Increment (all static)

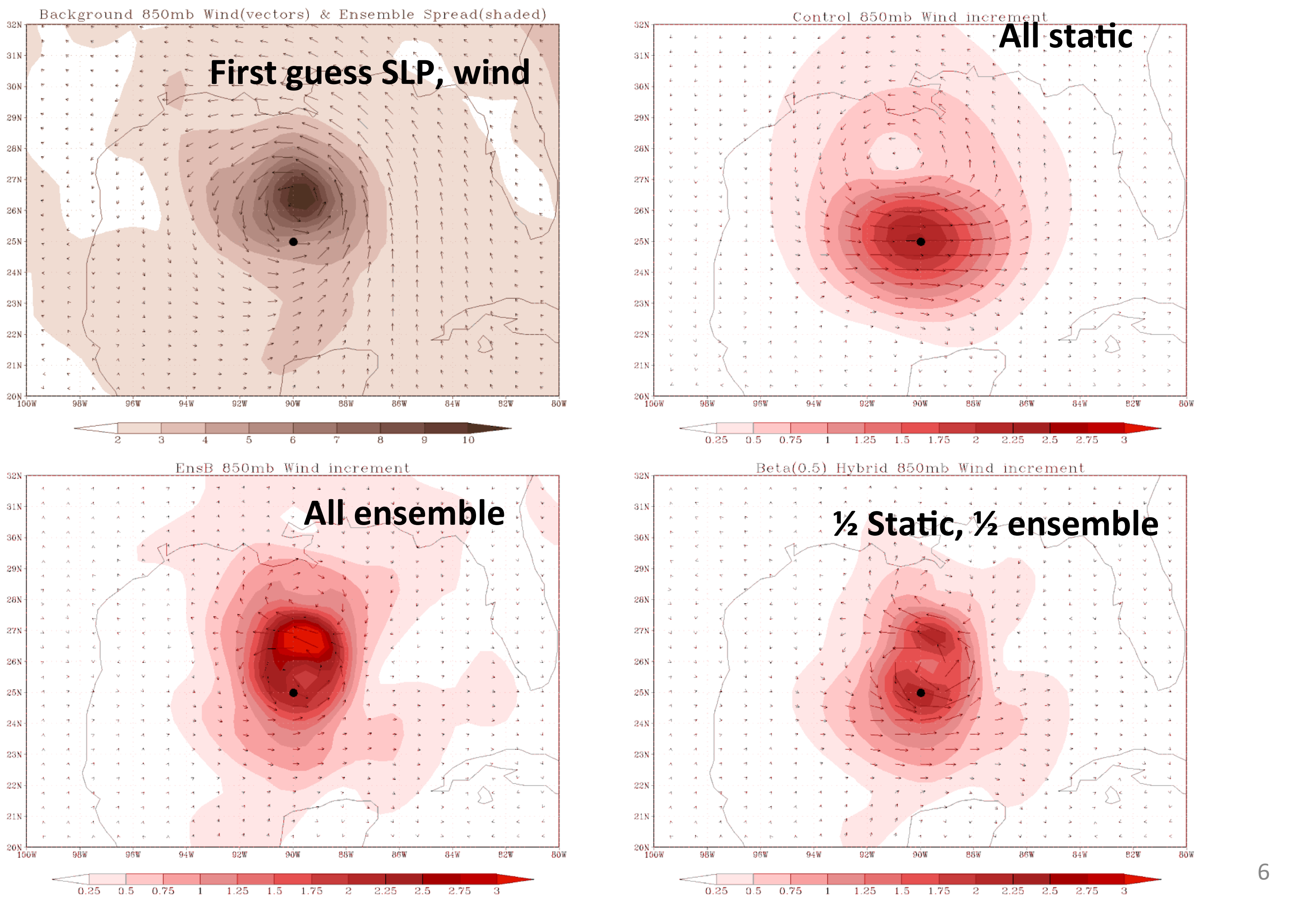


Increment (all ensemble)



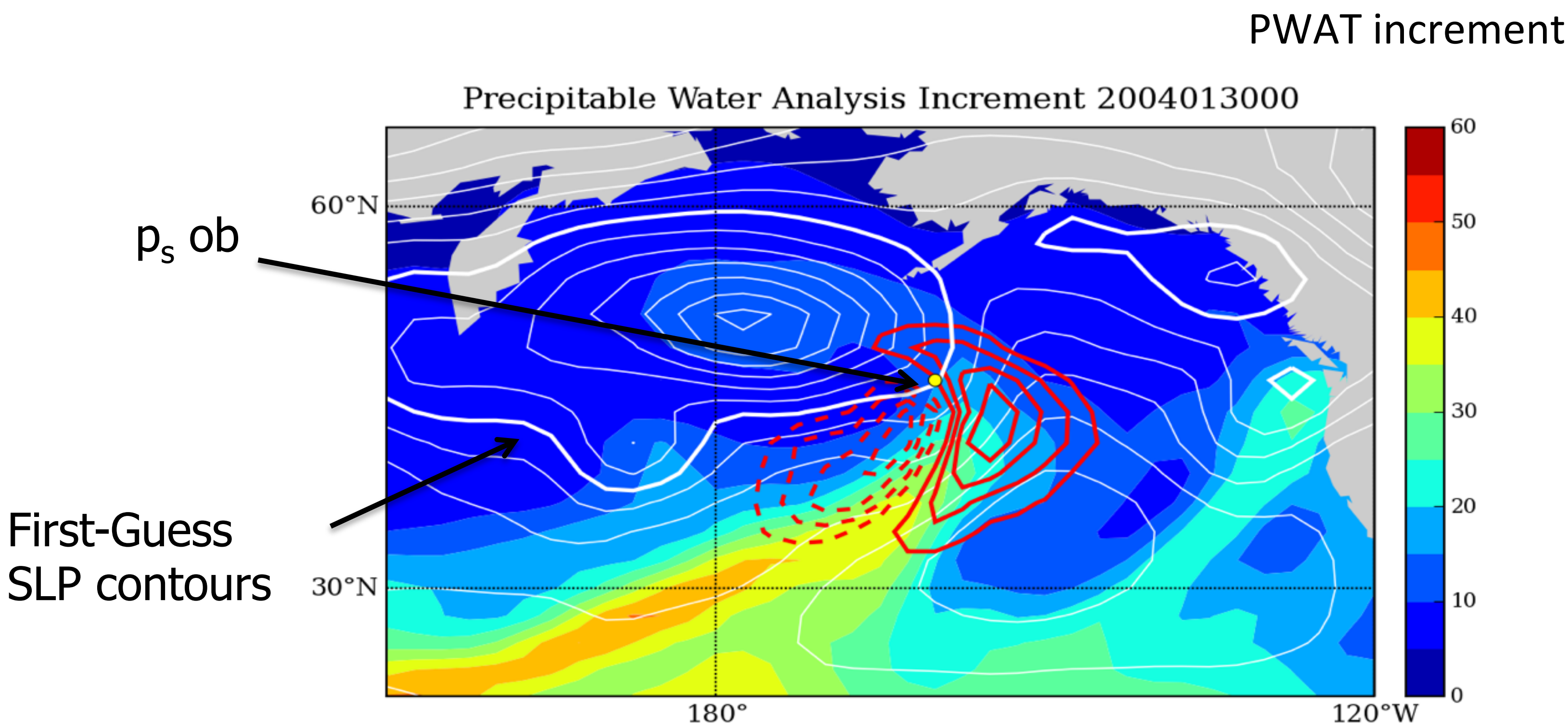
What does $\mathbf{B}_{\mathbf{ens}}$ do?

Zonal wind observation near a hurricane (Ike)



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

Surface pressure observation near an “atmospheric river”



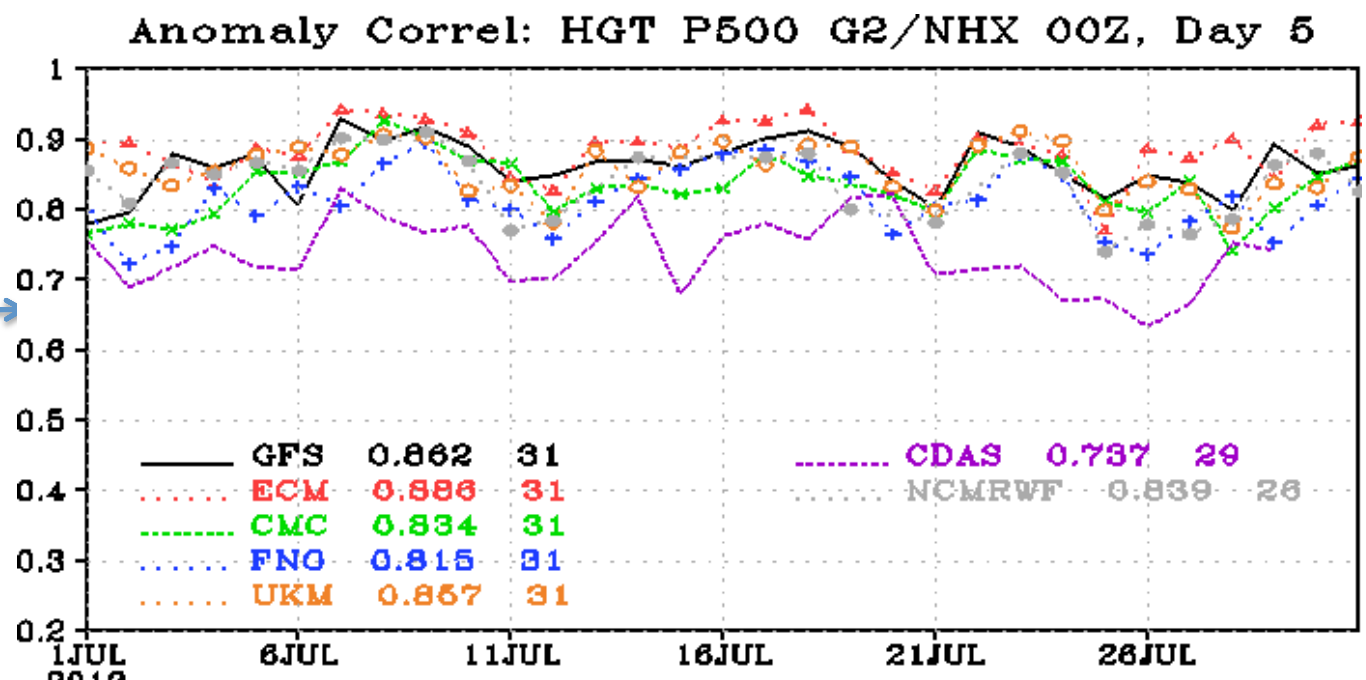
3Dvar increment would be zero!

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

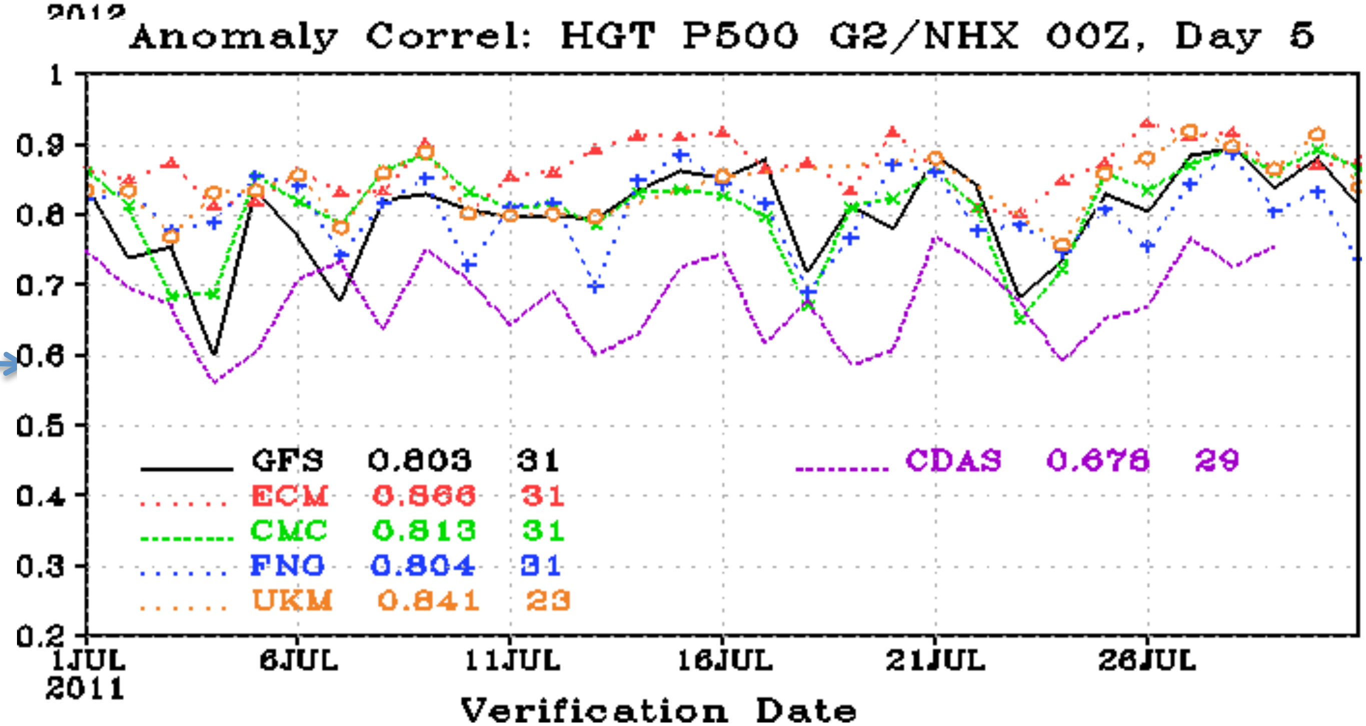
What does $\mathbf{B}_{\mathbf{ens}}$ do?

- 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***

Stats for July 2012 (hybrid implemented May)



Stats for July 2011 (GFS used all static B)

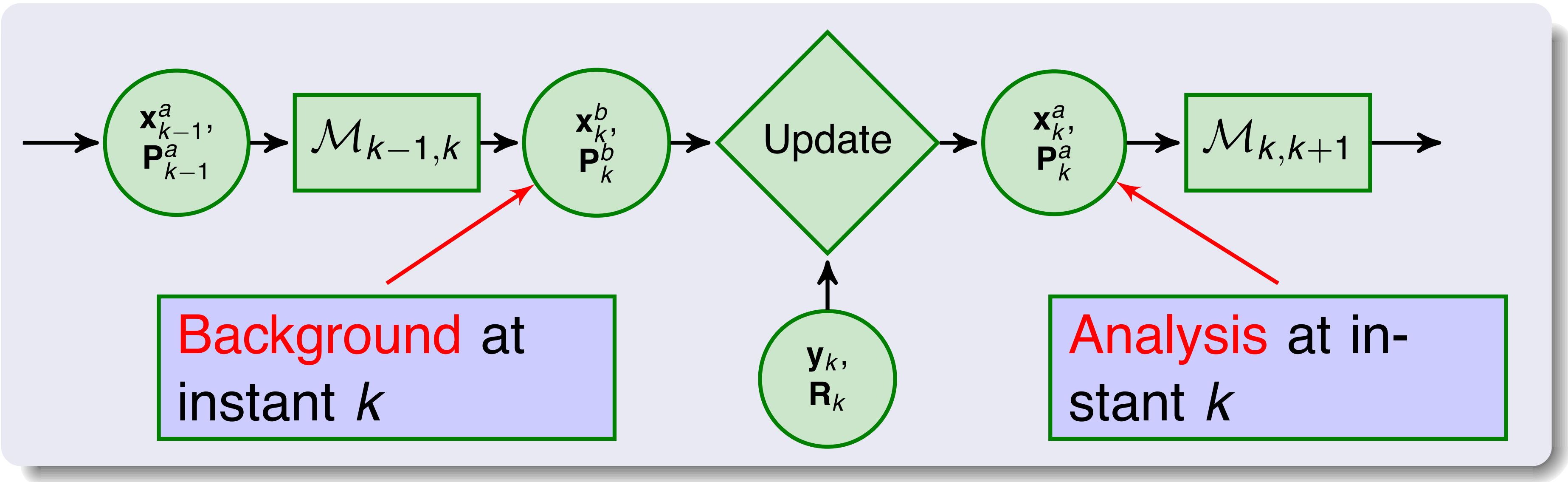


So what's the catch?

- 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.
- In NCEP operations, an "Ensemble Kalman Filter" (EnKF)* is used to generate the background ensemble.

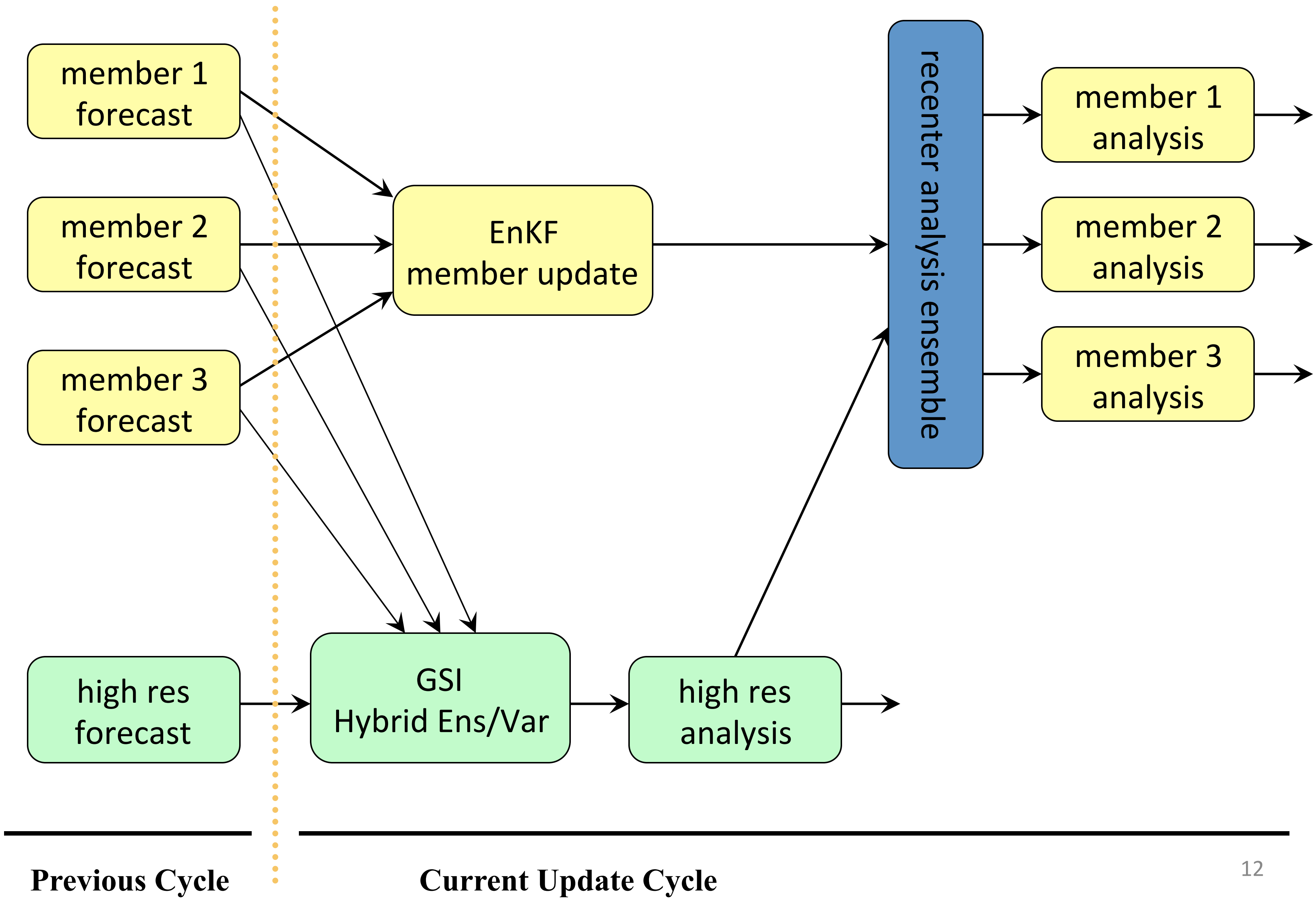
*EnKF: A standalone DA system that updates every ensemble member with new observations every analysis time using the ensemble to estimate the background-error covariance (no static part). Google "ensemble-based atmospheric data assimilation" for a review article by Tom Hamill.

The Ensemble Kalman Filter (EnKF)



- Update step uses background-error covariances (**$\mathbf{B}=\mathbf{B}_{ens}=\mathbf{P}^b$**) estimated from ensemble to update ensemble state variables directly (no variational minimization).
- Ad-hoc techniques needed to account for un-represented sources of error (sampling, model) – *covariance inflation and localization*.

Dual-Res Coupled Ensemble 3DVar



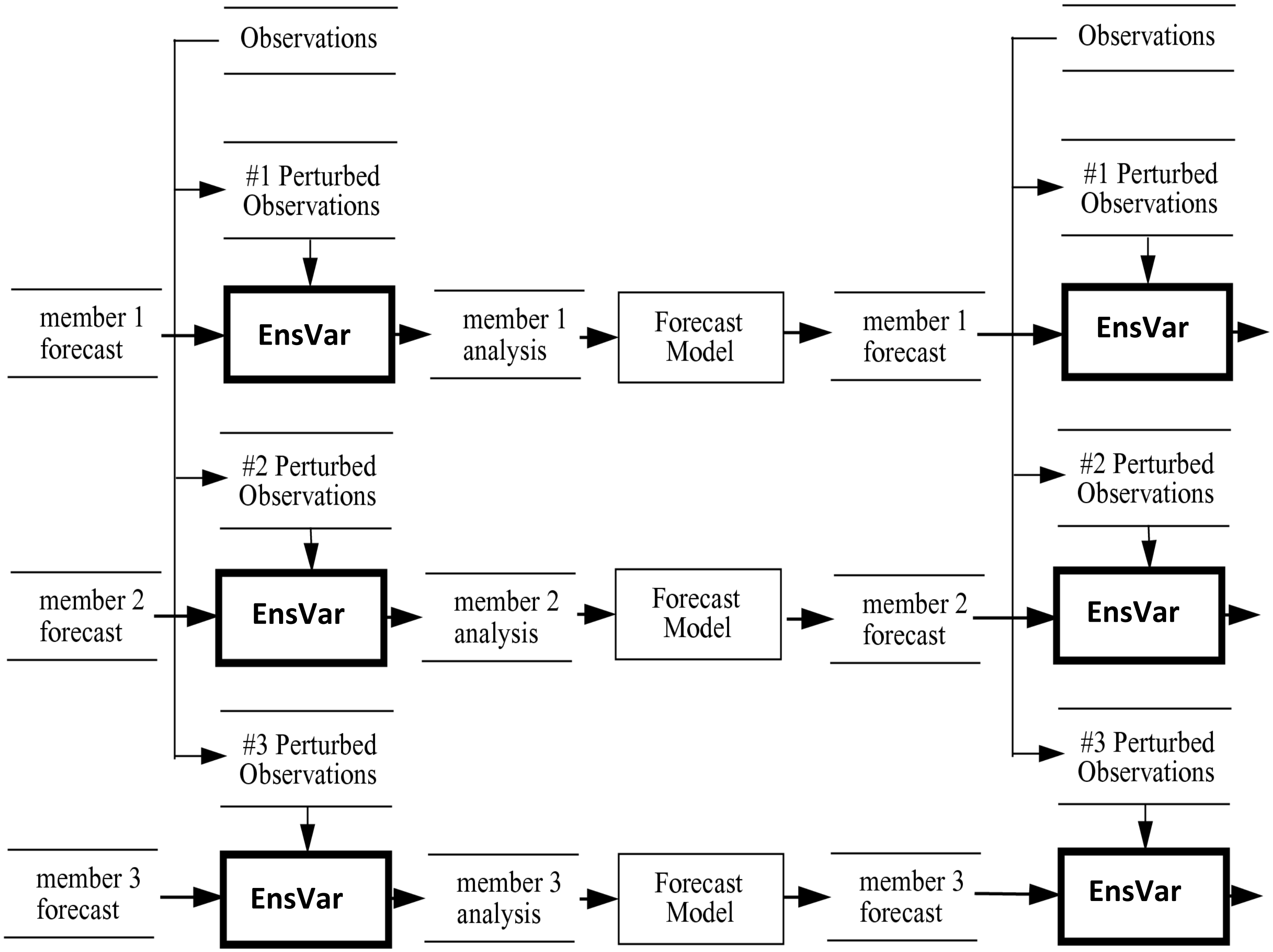
Advantages of the hybrid approach

Features from EnKF	Features from VAR
Extra flow-dependence in **B**	Localization done better for non-local obs (radiances).
More flexible treatment of model error (can be treated in ensemble)	Dual-resolution capability – can produce a high-res "control" (deterministic) analysis.
Automatic initialization of ensemble forecasts, propagation of covariance info from one cycle to the next.	Ease of adding extra constraints to cost function

What if I’m not running an EnKF?

- In principle, any ensemble can be used (but analysis won’t be better than 3DVar unless the ensemble represents the forecast errors well).
- GSI can ingest GFS global ensemble to update regional models (WRF ARW/NMM).
- 80-member GFS/EnKF 6-h ensemble forecasts are archived at NCEP since May 2012 – but not publicly available right now.

Ensembles of EnsVar – no EnKF needed
(in the future – much too expensive now)

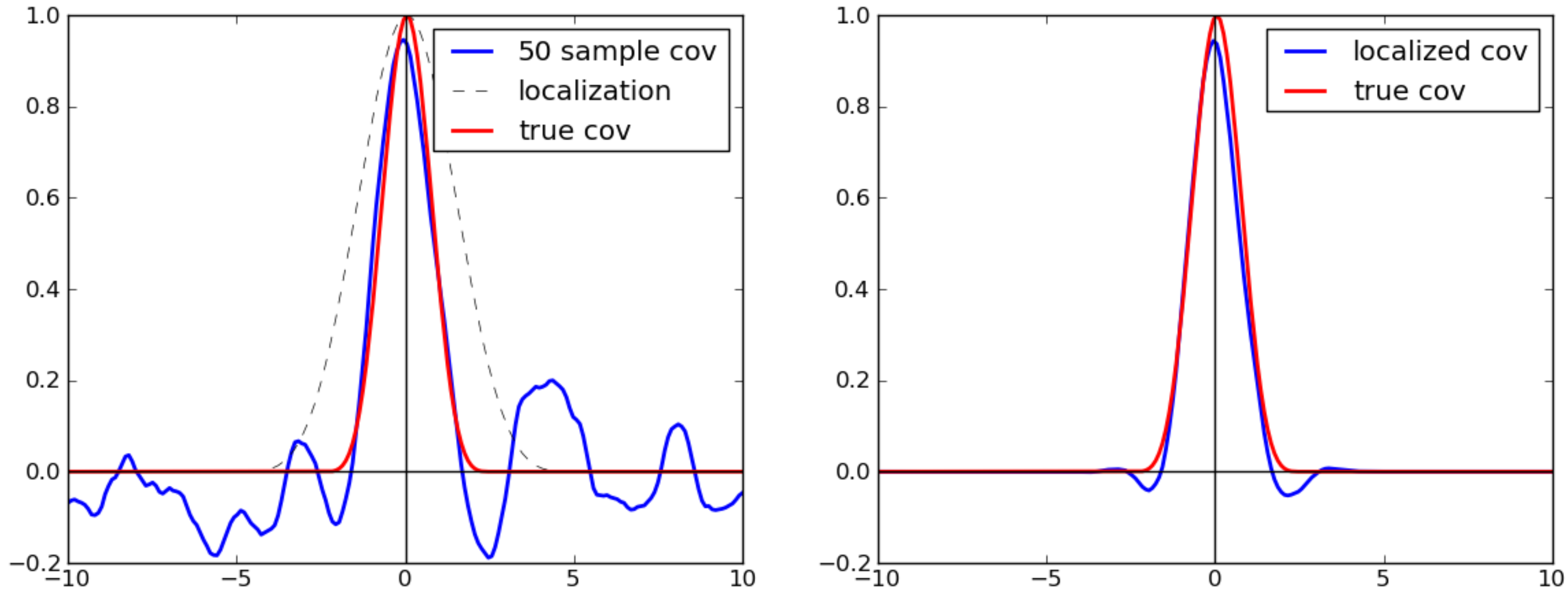


- This schematic is a simplification, since EnsVar uses every member to estimate background-error covariances

How to configure the GSI hybrid

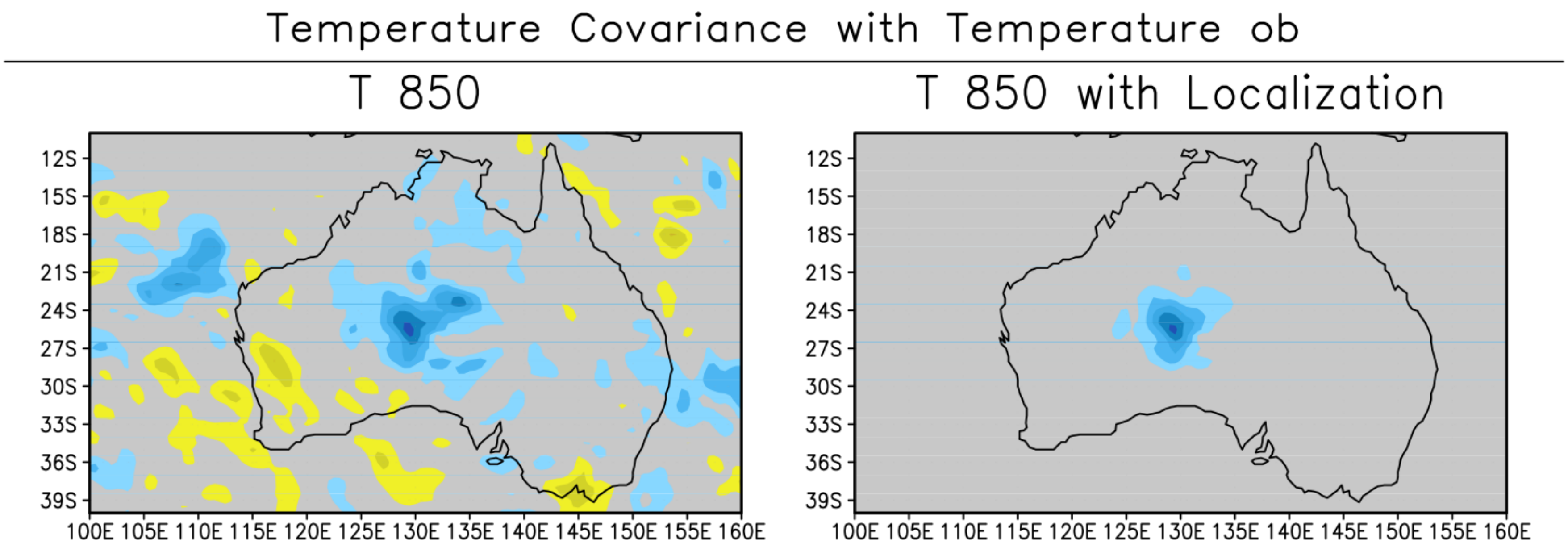
- Namelist parameters in **&hybrid_ensemble_parameters** control
  - ensemble size and horizontal resolution.
  - Source of ensemble (from GFS or host model).
  - Weighting factor for static covariance (1 means all static, 0 means all ensemble).
  - Whether to neglect cross-variable covariances in ozone update.
  - *Horizontal and vertical "covariance localization" distances*.
- Also need to setup symlinks in driver script so GSI can find ensemble files.
- Practical designed to illustrate sensitivity to static covariance weighting factor (BETA1_INV), ensemble size (N_ENS), and localization length scales (S_ENS_H, S_ENS_V).

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

A real-world example of covariance localization



GSI ensemble 3DVar cost function (with localization)

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

$\mathbf{B}_f$: (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.***

β: Weighting factor (0.25 means total **B** is ¾ ensemble).

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

Summary

- The "hybrid" ensemble 3DVar GSI system uses an ensemble of first-guess forecasts to better estimate the background-error covariance term in the cost function.
  - More information can be extracted from obs.
  - Added expense (and complexity) of running (and updating) an ensemble.
- Ensemble (co)variances must be representative of control forecast error – should be informed by observations.
- Need to carefully tune localization length scales (depends on model resolution, observing network).