DTC GSI Tutorial 2013

GSI Hybrid Data Assimilation

Daryl Kleist

Environmental Modeling Center

NOAA/NWS/NCEP

With acknowledgements to John Derber (EMC), Dave Parrish (EMC), Jeff Whitaker (NOAA/ESRL), and Kayo Ide (UMD)

Variational Data Assimilation

$$J_{\mathrm{Var}}\left(\mathbf{x}'\right)=\frac{1}{2}\left(\mathbf{x}'\right)^{\mathrm{T}}\mathbf{B}_{\mathrm{Var}}^{-1}\left(\mathbf{x}'\right)+\frac{1}{2}\left(\mathbf{H}\mathbf{x}'-\mathbf{y}_{\mathrm{o}}'\right)^{\mathrm{T}}\mathbf{R}^{-1}\left(\mathbf{H}\mathbf{x}'-\mathbf{y}_{\mathrm{o}}'\right)+J_{\mathrm{c}}$$

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

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

$\mathbf{B}_{\mathbf{var}}$: Background error covariance

H : Observations (forward) operator

R : Observation error covariance (Instrument + representativeness)

$\mathbf{y}_o'$: Observation innovations

J_c : Constraints (physical quantities, balance/noise, etc.)

B is typically static and estimated a-priori/offline

Kalman Filter in Var Setting

Forecast Step

$$\begin{cases} \mathbf{x}^{b} = M\left(\mathbf{x}^{a}\right) \\ \mathbf{B}_{KF} = \mathbf{M}\mathbf{A}_{KF}\mathbf{M}^{T} + \mathbf{Q} \end{cases}$$

Analysis

$$\begin{cases} \mathbf{x}^{a} = \mathbf{x}^{b} + \mathbf{K}\left(\mathbf{H}\mathbf{x}^{b} - \mathbf{y}\right) \\ \mathbf{K} = \mathbf{B}_{KF}\mathbf{H}^{T}\left(\mathbf{R} + \mathbf{H}\mathbf{B}_{KF}\mathbf{H}^{T}\right)^{-1} \\ \mathbf{A}_{KF} = \left(\mathbf{I} - \mathbf{K}\mathbf{H}\right)\mathbf{B}_{KF} \end{cases}$$

Extended Kalman Filter

- Analysis step in variational framework (cost function)

$$J_{KF}\left(\mathbf{x}'\right) = \frac{1}{2}\left(\mathbf{x}'\right)^{T}\mathbf{B}_{KF}^{-1}\left(\mathbf{x}'\right) + \frac{1}{2}\left(\mathbf{y}'_{o} - \mathbf{H}\mathbf{x}'\right)^{T}\mathbf{R}^{-1}\left(\mathbf{y}'_{o} - \mathbf{H}\mathbf{x}'\right)$$

- $\mathbf{B}_{KF}$: Time evolving background error covariance
- $\mathbf{A}_{KF}$: Inverse [Hessian of $J_{KF}(\mathbf{x}')$]

Motivation from KF

- ***Problem***: dimensions of $\mathbf{A}_{KF}$ and $\mathbf{B}_{KF}$ are huge, making this practically impossible for large systems (GFS for example).
- ***Solution***: sample and update using an ensemble instead of evolving $\mathbf{A}_{KF}/\mathbf{B}_{KF}$ explicitly

Forecast Step: $\mathbf{X}^a \rightarrow \mathbf{X}^b \quad \mathbf{B}_{KF} \approx \mathbf{B}_\mathbf{e} = \frac{1}{K-1}\mathbf{X}^b\left(\mathbf{X}^b\right)^T$

Analysis Step: $\mathbf{X}^b \rightarrow \mathbf{X}^a \quad \mathbf{A}_{KF} \approx \mathbf{A}_\mathbf{e} = \frac{1}{K-1}\mathbf{X}^a\left(\mathbf{X}^a\right)^T$

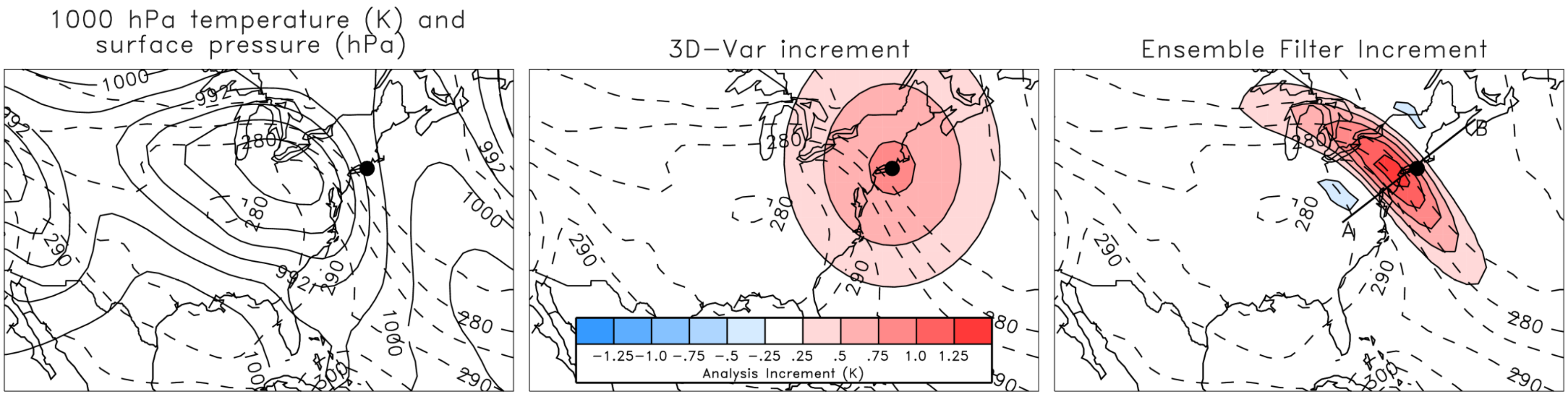
} Ensemble Perturbations

What does $\mathbf{B}_e$ gain us?

- Allows for flow-dependence/errors of the day
- Multivariate correlations from dynamic model
  - Quite difficult to incorporate into fixed error covariance models
- Evolves with system, can capture changes in the observing network
- More information extracted from the observations => better analysis => better forecasts

What does $\mathbf{B}_e$ gain us?

Temperature observation near warm front

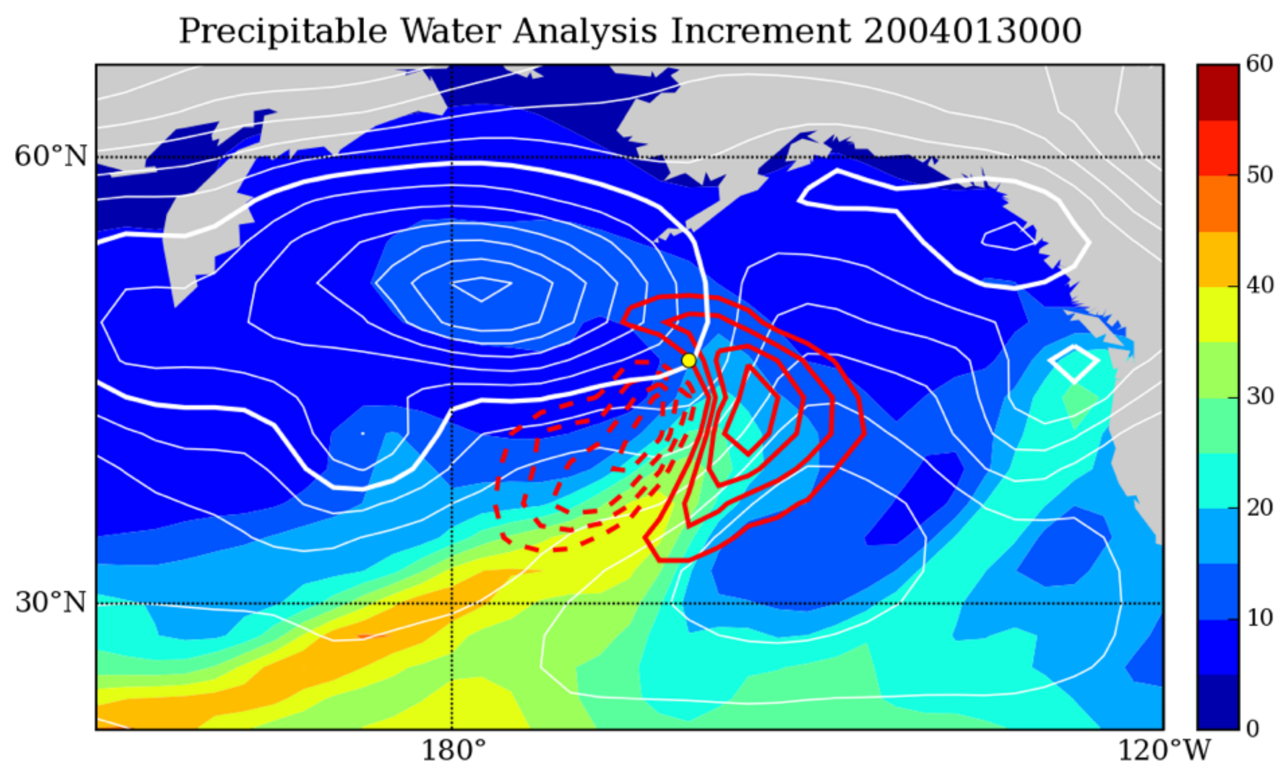


$\mathbf{B}_f$

$\mathbf{B}_e$

What does $\mathbf{B}_e$ gain us?

Surface pressure observation near "atmospheric river"



First guess surface pressure (white contours) and precipitable water increment (A-G, red contours) after assimilating a single surface pressure observation (yellow dot) using $\mathbf{B}_e$.

What is "hybrid DA"?

$$J(\mathbf{x}') = \beta_f \frac{1}{2}(\mathbf{x}')^T \mathbf{B}_f^{-1}(\mathbf{x}') + \beta_e \frac{1}{2}(\mathbf{x}')^T \mathbf{B}_e^{-1}(\mathbf{x}') + \frac{1}{2}(\mathbf{H}\mathbf{x}' - \mathbf{y}')^T \mathbf{R}^{-1}(\mathbf{H}\mathbf{x}' - \mathbf{y}')$$

Simply put, linear combination of fixed and ensemble based **B:**

$\mathbf{B}_f$: Fixed background error covariance

$\mathbf{B}_e$: Ensemble estimated background error covariance

β_f: Weighting factor for fixed contribution (0.25 means 25% fixed)

β_e: Weighting factor for ensemble contribution (typically 1- β_f)

GSI Hybrid [3D] EnVar

(ignoring preconditioning for simplicity)

- Incorporate ensemble perturbations ***directly*** into variational cost function through extended control variable
  - Lorenc (2003), Buehner (2005), Wang et. al. (2007), etc.

$$J(\mathbf{x}_f', \boldsymbol{\alpha}) = \beta_f \frac{1}{2}(\mathbf{x}_f')^T \mathbf{B}_f^{-1}(\mathbf{x}_f') + \beta_e \frac{1}{2}\sum_{n=1}^{N}(\boldsymbol{\alpha}^n)^T \mathbf{L}^{-1}(\boldsymbol{\alpha}^n) + \frac{1}{2}(\mathbf{H}\mathbf{x}_t' - \mathbf{y}')^T \mathbf{R}^{-1}(\mathbf{H}\mathbf{x}_t' - \mathbf{y}')$$

$$\mathbf{x}_t' = \mathbf{x}_f' + \sum_{n=1}^{N}(\boldsymbol{\alpha}^n \circ \mathbf{x}_e{}^n)$$

β_f & β_e: weighting coefficients for fixed and ensemble covariance respectively

$\mathbf{x}_t'$: (total increment) sum of increment from fixed/static **B** ($\mathbf{x}_f'$) and ensemble **B**

$\boldsymbol{a}_k$: extended control variable; $\mathbf{x}_k^e$:ensemble perturbations

- analogous to the weights in the LETKF formulation

L: correlation matrix [effectively the localization of ensemble perturbations]

Preconditioning Sidebar

$$\boldsymbol{v}^n = \beta_\mathrm{e}\mathbf{L}^{-1}\boldsymbol{\alpha}^n \qquad \mathbf{z} = \beta_\mathrm{f}\mathbf{B}^{-1}\mathbf{x}'_\mathrm{f}$$

$$J(\mathbf{z},\boldsymbol{v}) = \frac{1}{2}(\mathbf{x}'_\mathrm{f})^\mathrm{T}\mathbf{z} + \frac{1}{2}\sum_{n=1}^{N}(\boldsymbol{\alpha}^n)^\mathrm{T}(\boldsymbol{v}^n) + J_\mathrm{o}$$

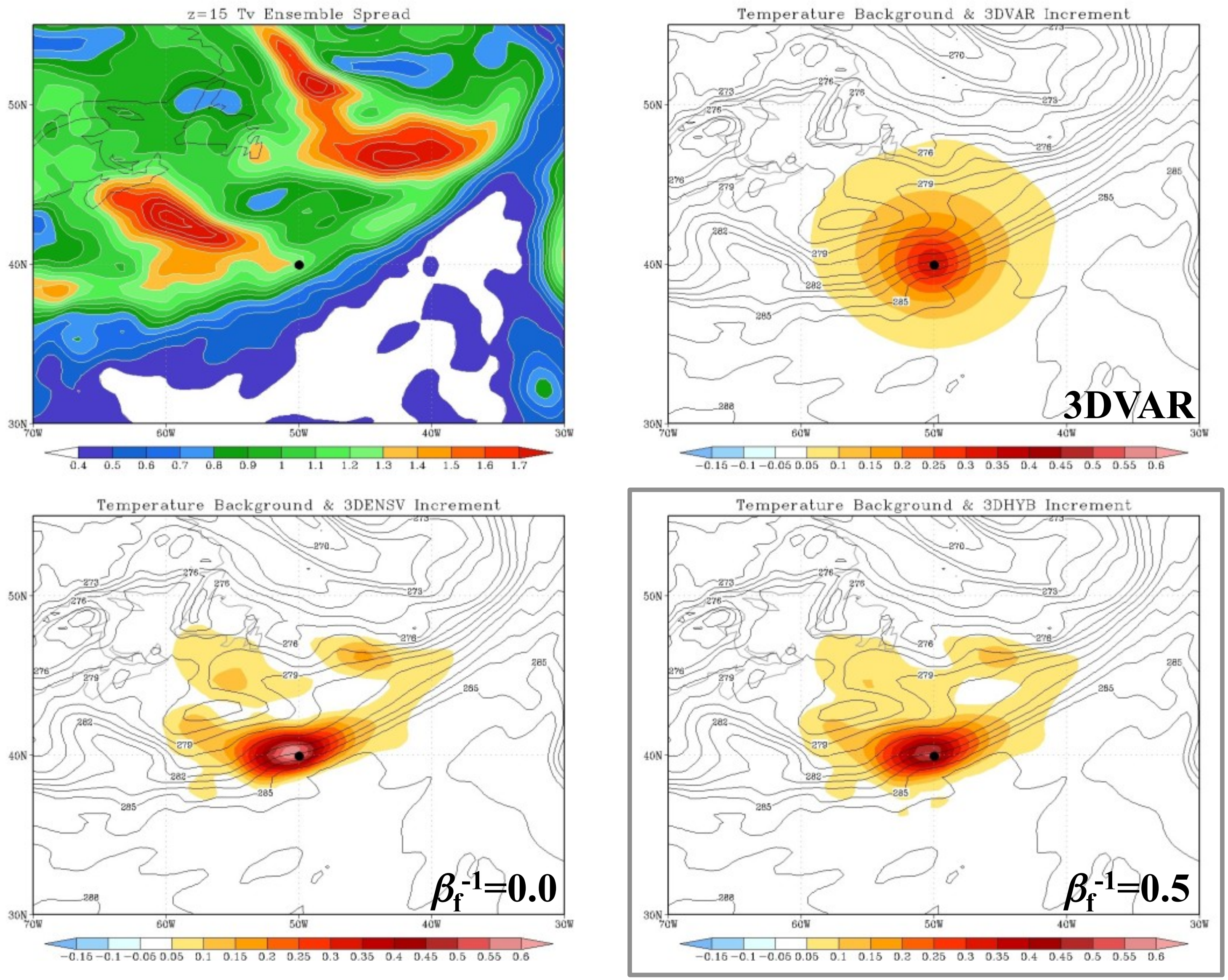
$$\mathbf{x}'_\mathrm{f} = (\beta_\mathrm{f})^{-1}\mathbf{B}\mathbf{z} \qquad \boldsymbol{\alpha} = (\beta_\mathrm{e})^{-1}\mathbf{L}\boldsymbol{v}$$

In default GSI minimization, double CG, inverses of **B** and **L** not needed and the solution is pre-conditioned by full **B**. Discussed later, this is why the GSI specifies the inverses of beta via the hybrid namelist.

This formulation differs from the UKMO and Canadians, who use a square root formulation and apply the weights directly to the increment:

$$\mathbf{x}'_\mathrm{t} = \beta_\mathrm{f}\mathbf{x}'_\mathrm{f} + \beta_\mathrm{e}\sum_{n=1}^{N}(\boldsymbol{\alpha}^n \circ \mathbf{x}_\mathrm{e}{}^n)$$

Single Temperature Observation



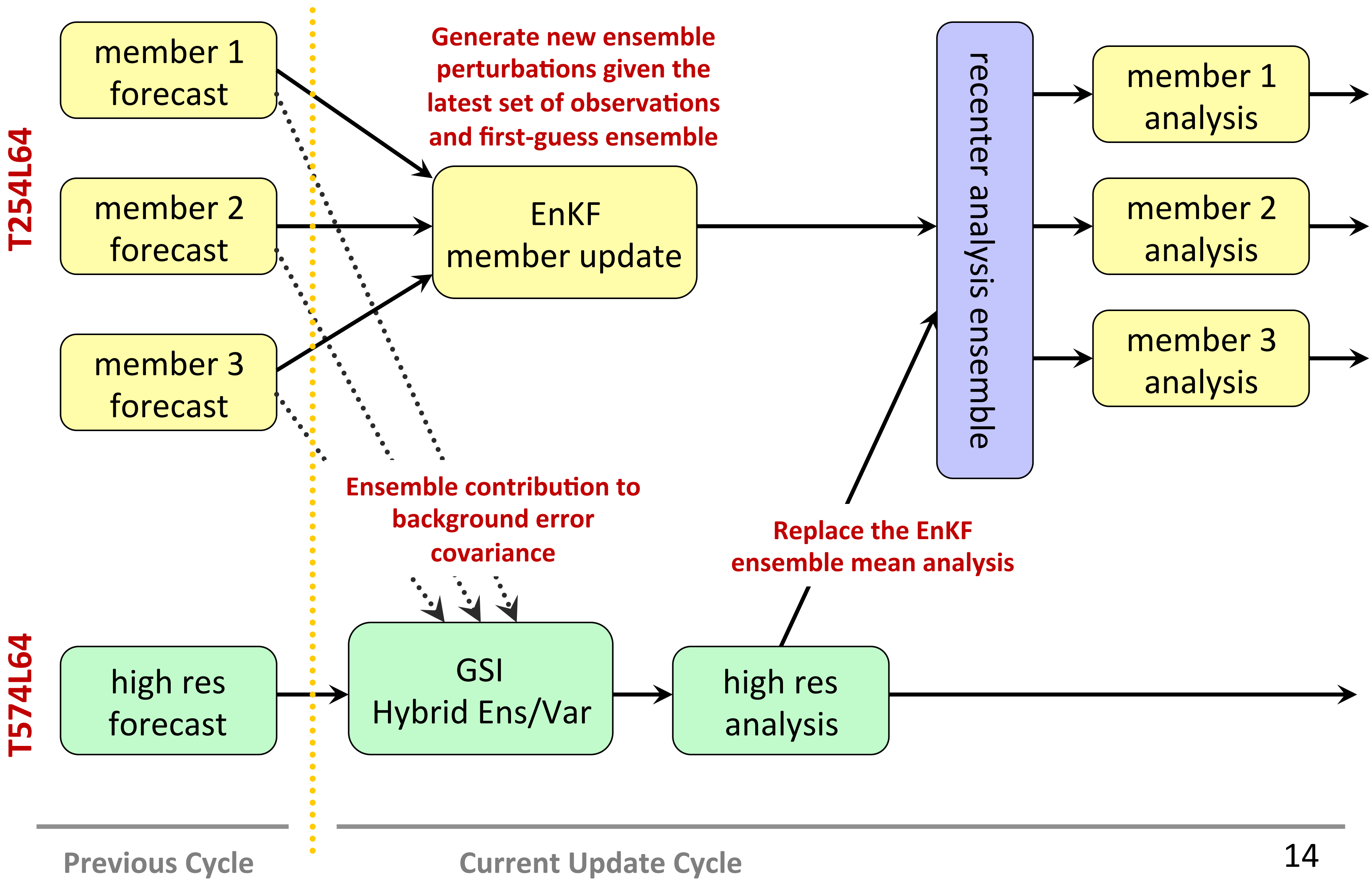
Why Hybrid?

	VAR (3D, 4D)	EnKF	Hybrid	References
Benefit from use of flow dependent ensemble covariance instead of static B		x	x	Hamill and Snyder 2000; Wang et al. 2007b,2008ab, 2009b, Wang 2011; Buehner et al. 2010ab
Robust for small ensemble			x	Wang et al. 2007b, 2009b; Buehner et al. 2010b
Better localization (physical space) for integrated measure, e.g. satellite radiance			x	Campbell et al. 2009
Easy framework to add various constraints	x		x	Kleist 2012
Framework to treat non-Gaussianity	x		x	
Use of various existing capabilities in VAR	x		x	Kleist 2012

So what's the catch?

- Need an ensemble that represents first guess uncertainty (background error)
- This can mean O(50-100+) for NWP applications
  - Smaller ensembles have larger sampling error (rely more heavily on $\mathbf{B}_f$)
  - Larger ensembles have increased computational expense
- Updating the ensemble: In NCEP operations, we currently utilize an Ensemble Kalman Filter
  - EnKF is a standalone (i.e. separate) DA system that updates every ensemble member with information from observations each analysis time using the prior/posterior ensemble to represent the error covariances. Google "ensemble based atmospheric assimilation" for a good review article by Tom Hamill.

Dual-Res Coupled Hybrid Var/EnKF Cycling

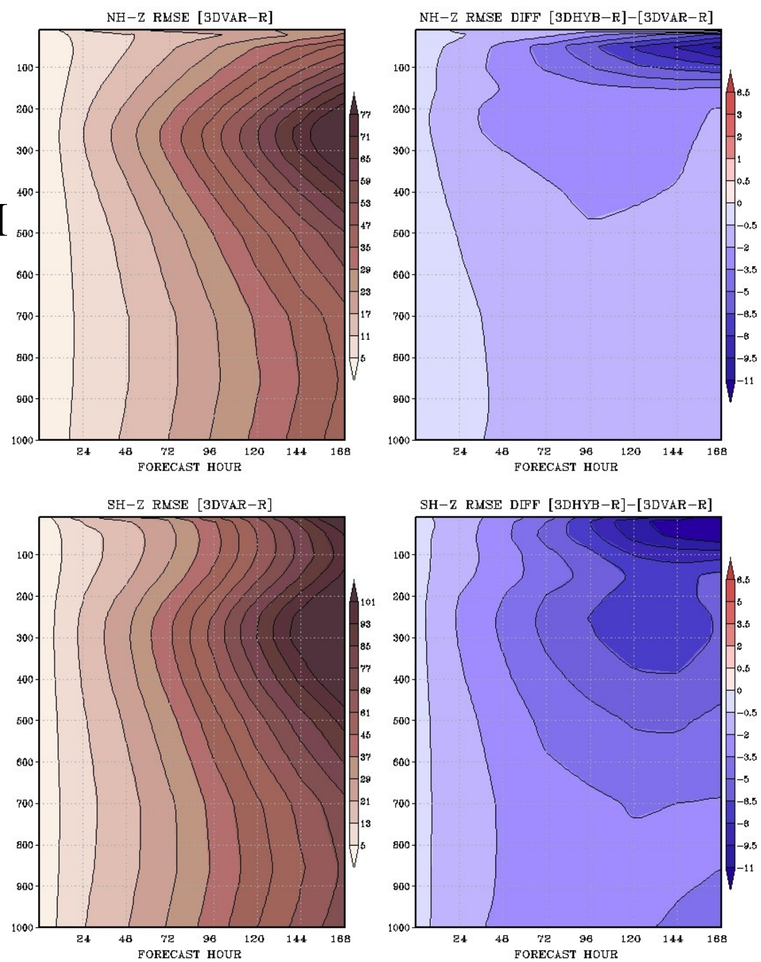


Operational (3D) Hybrid

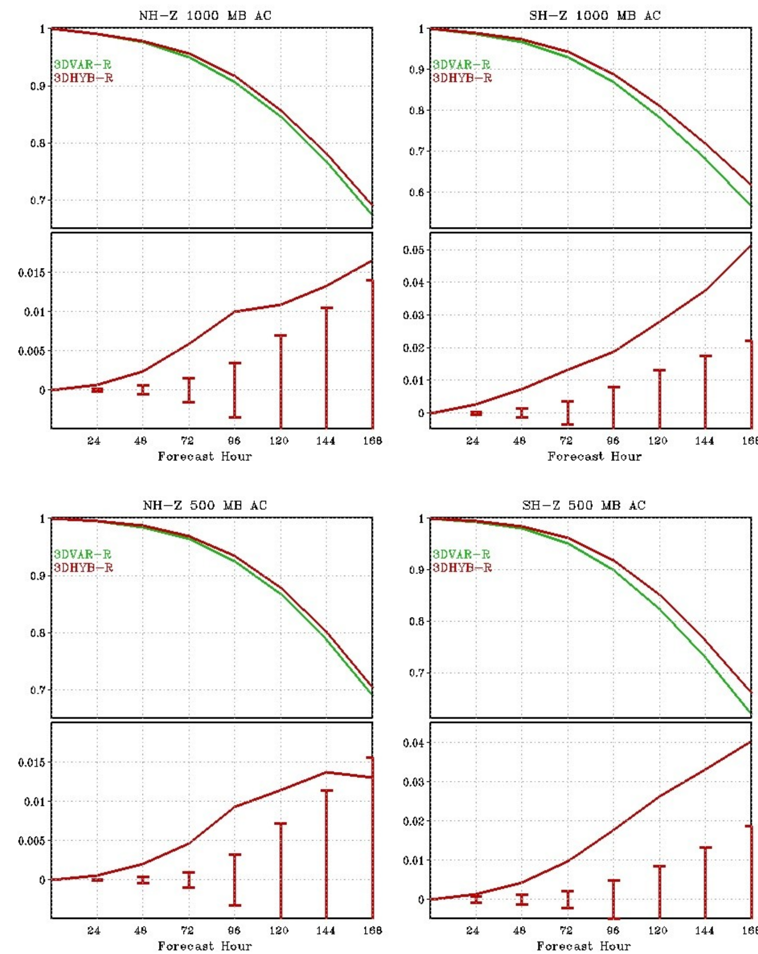
- Implemented into GDAS/GFS at 12z 22 May 2012

NH

SH



3DHYB-3DVAR



1000 mb

500 mb

15

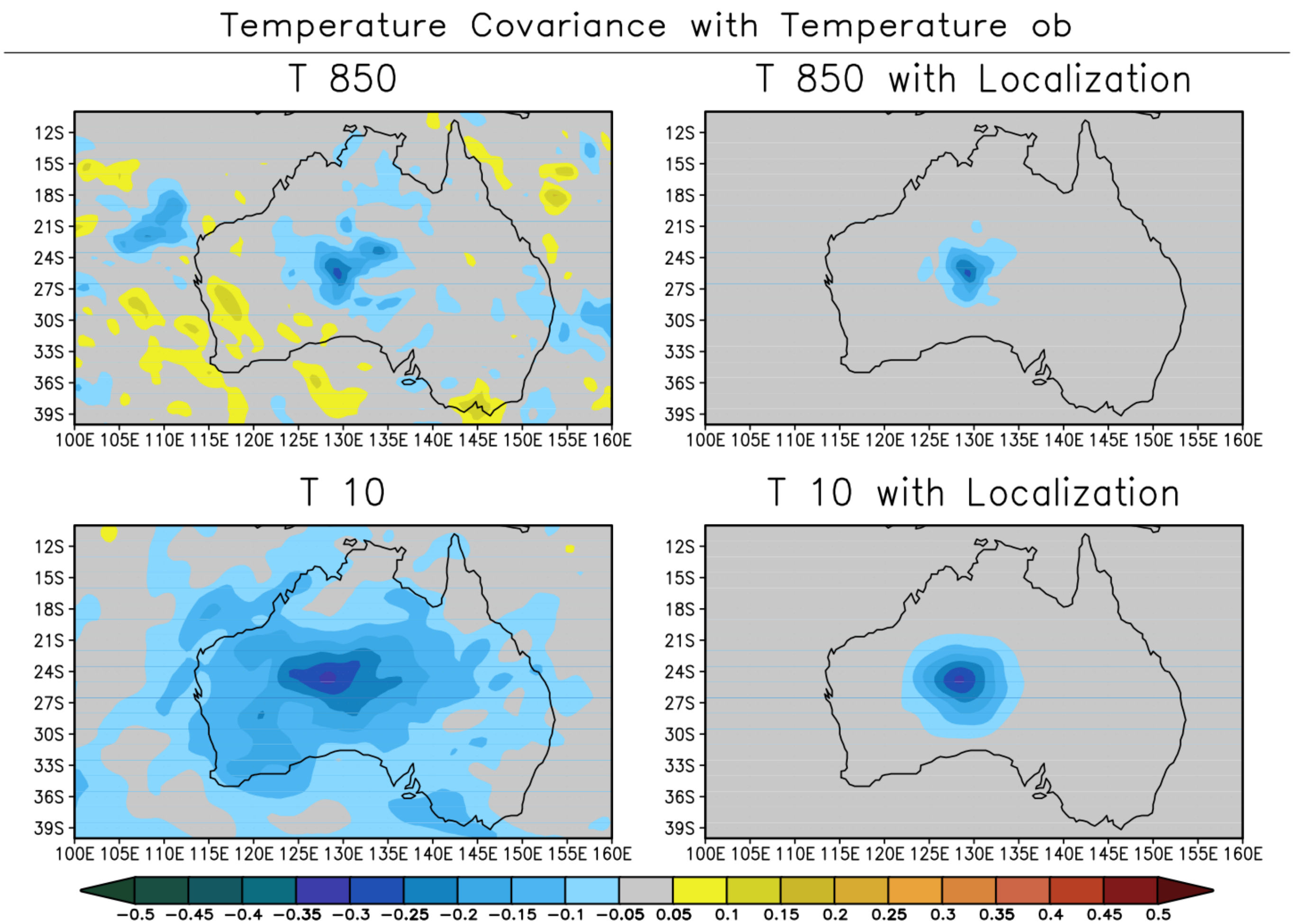
What if you don't have an EnKF?

- In principle, any ensemble can be used
  - However, ensemble should represent well the forecast errors
- GSI can ingest GFS (global spectral) ensemble to update regional models (WRF ARF/NMM)
  - Has been highly successful in NAM, RR, HWRF applications
- 80 member GFS/EnKF 6h ensemble forecasts are archived at NCEP since May 2012
  - Real-time ensemble should also be publicly available

GSI Hybrid Configuration

- Several hybrid related parameters controlled via GSI namelist &hybrid_ensemble
  - Logical to turn on/off hybrid_ensemble option (**l_hyb_ens**)
  - Ensemble size (**n_ens**), resolution (**jcap_ens, nlat_ens, nlon_ens**)
  - Source of ensemble: GFS spectral, native model, etc. (**regional_ensemble_option**)
  - Weighting factor for static contribution to increment (**beta1_inv**)
  - Horizontal and vertical distances for localization, via **L** on augmented control variable (**s_ens_h, s_ens_v**)
    - Localization distances are the same for all variables since operating on α
  - Option to specify different localization distances as a function of vertical level (**readin_localization**)
    - Instead of single parameters, read in ascii file containing a value for each layer
    - Example for global in fix directory (global_hybens_locinfo.l64.txt)
  - Other regional options related to resolution, pseudo ensemble, etc.

Localization



Hybrid 4D EnVar [No Adjoint]

The 4D EnVar cost function can be easily expanded to include a static contribution

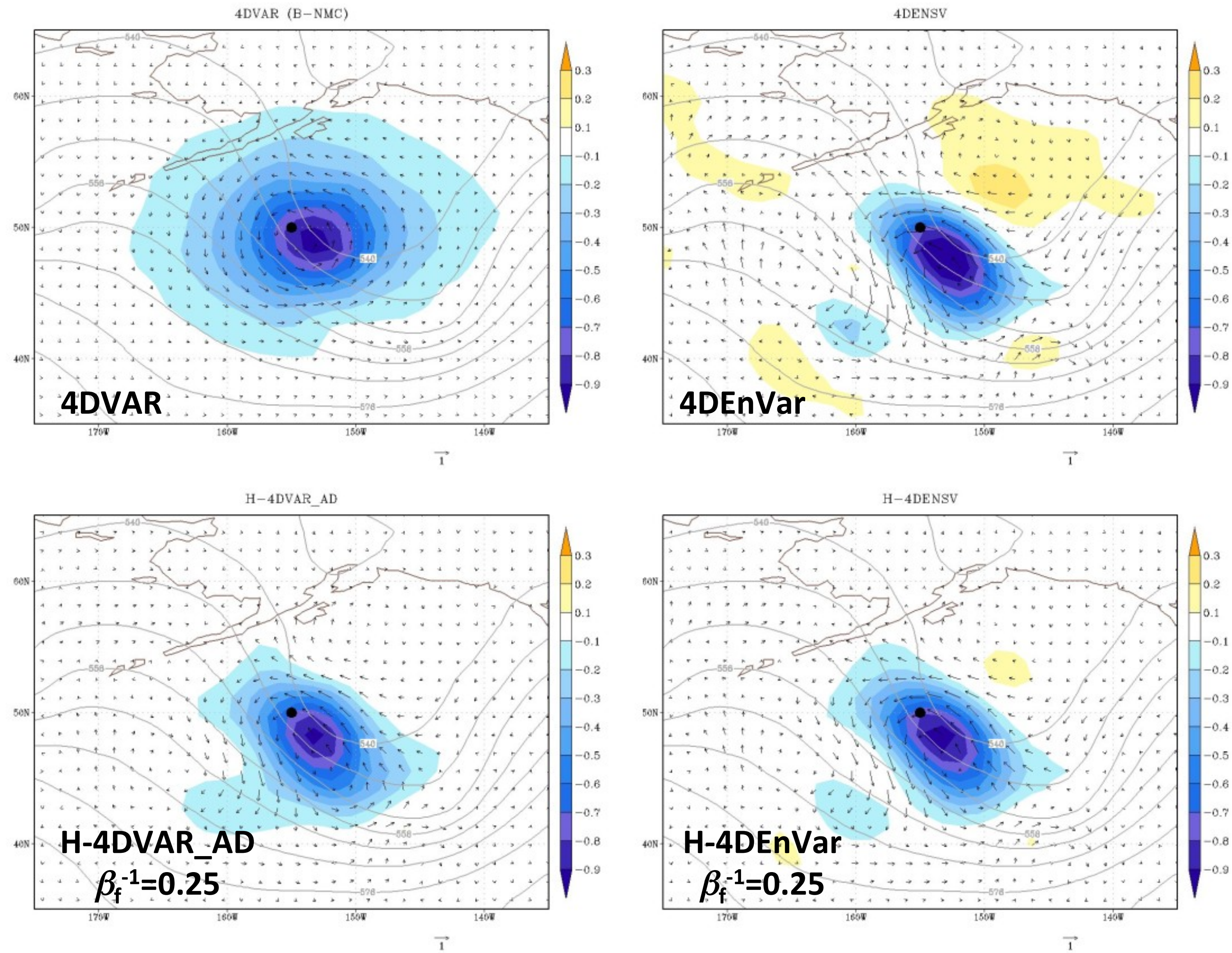
$$J(\mathbf{x}_f', \boldsymbol{\alpha}) = \beta_f \frac{1}{2}(\mathbf{x}_f')^T \mathbf{B}_f^{-1}(\mathbf{x}_f') + \beta_e \frac{1}{2}\sum_{n=1}^{N}(\boldsymbol{\alpha}^n)^T \mathbf{L}^{-1}(\boldsymbol{\alpha}^n) + \frac{1}{2}\sum_{k=1}^{K}(\mathbf{H}_k \mathbf{x}_k' - \mathbf{y}_k')^T \mathbf{R}_k^{-1}(\mathbf{H}_k \mathbf{x}_k' - \mathbf{y}_k')$$

Where the 4D increment is prescribed exclusively through linear combinations of the 4D ensemble perturbations plus static contribution

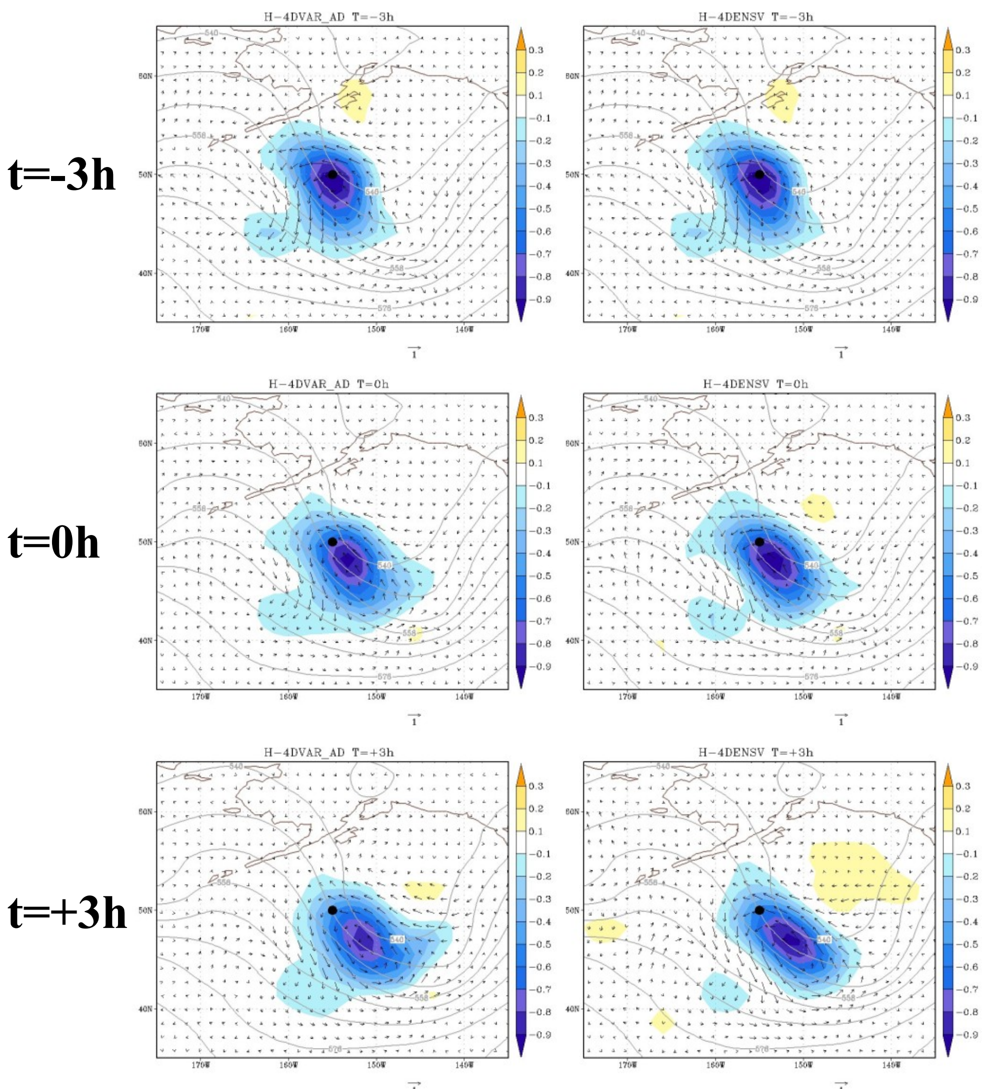
$$\mathbf{x}_k' = \mathbf{x}_f' + \sum_{n=1}^{N}\left(\boldsymbol{\alpha}^n \circ (\mathbf{x}_e)_k^n\right)$$

Here, the static contribution is considered time-invariant (i.e. from 3DVAR-FGAT). Weighting parameters exist just as in the other hybrid variants.

Single Observation (-3h) Example for 4D Variants



Time Evolution of Increment



Solution at beginning of window same to within round-off (because observation is taken at that time, and same weighting parameters used)

Evolution of increment qualitatively similar between dynamic and ensemble specification

** Current linear and adjoint models in GSI are computationally unfeasible for use in 4DVAR other than simple single observation testing at low resolution

Summary

- The "hybrid" EnVar option in GSI uses perturbations from an ensemble of short term forecasts to better estimate the background error covariance term.
  - Added expense (mostly IO)
  - Added complexity
    - Running/updating ensemble
    - Additional DA component (EnKF)?
- Although any ensemble can be used, should be representative of forecast/background error
  - Feel effects of observations
- More tuning parameters
  - Weights, localization, etc., which may be dependent upon model, resolution, and observing system.