2016 Hurricane WRF Tutorial, Jan 25-27, 2014, College Park, MD

Introduction to Data Assimilation & Gridpoint Statistical Interpolation System (GSI)

Hui Shao

Developmental Testbed Center

Outline

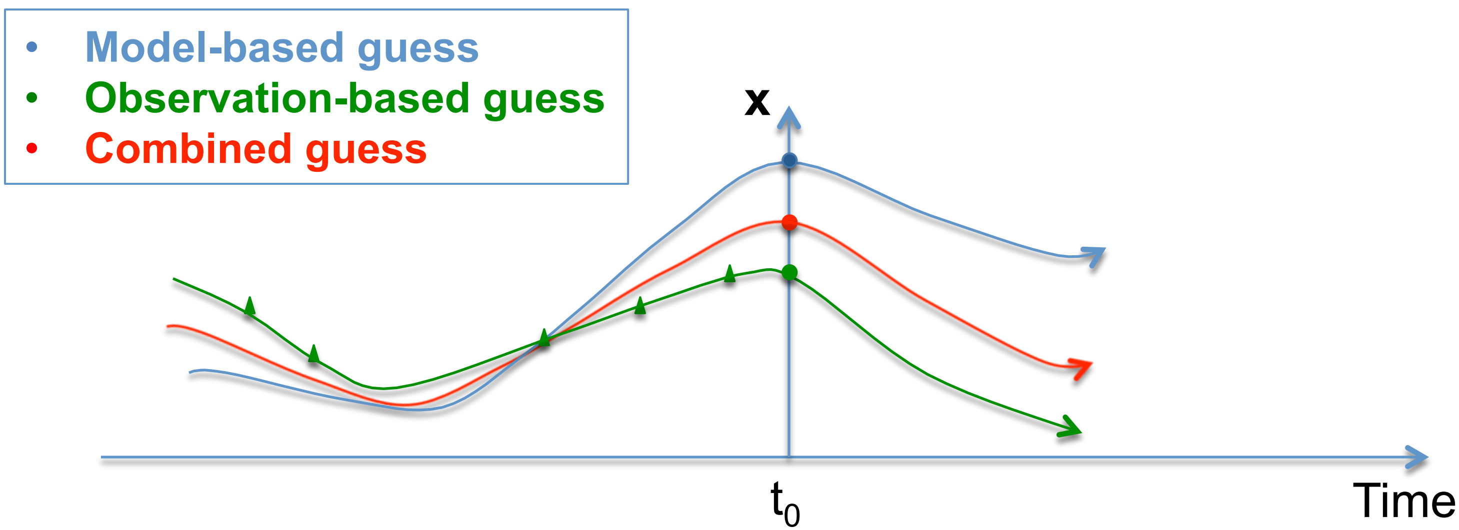
- What is data assimilation?
- GSI concepts and methods
- Community support and service

What is Data Assimilation

- Numerical Weather Prediction (NWP)

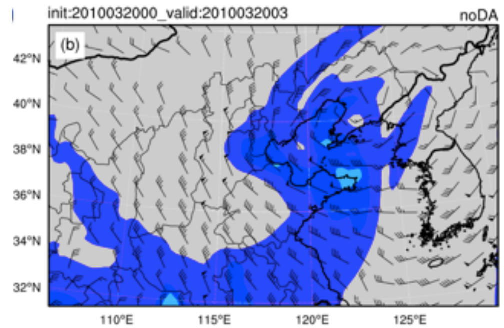
  Given an estimate of the current state of the atmosphere (initial conditions), and appropriate surface (and lateral, if regional) boundary conditions, the model simulates the atmospheric evolution (forecasts)

  - "Knowing the current state of the weather is as important as the numerical computer models processing the data."-NOAA National Climatic Data Center

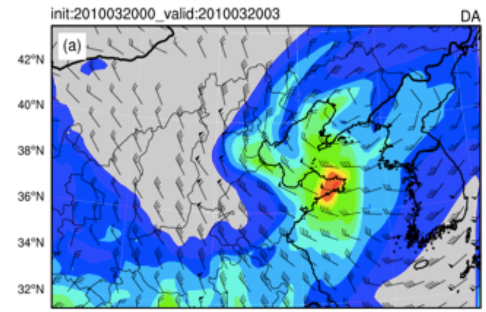


What is Data Assimilation (Cont.)

+



=



Observations (y): provide an incomplete description of the atmospheric state, but bring up to date information

Background (x_b): gives a complete description of the atmosphere, but errors grow rapidly in time

Analysis (x_a at $t=t_0$)

Initial conditions

Forecast model

Forecast (x_t)

Data assimilation: combines these two sources of information to produce an optimal (best) estimate of the atmospheric state

$$x_a = w_b x_b + w_o y = x_b + K(y - x_b)$$

✓ How to find optimal weighting for background information and observations?

Introduction to GSI

- GSI used in NCEP operations for
  - Regional: NAM, HWRF
  - Global: GFS
  - Analysis system: RTMA
- NOAA: RAP, HRRR
- GMAO collaboration: GEOS
- Operational at AFWA
- Modification to fit into WRF and NCEP infrastructure
- Evolution to Earth System Modeling Framework (ESMF)
- Community support and distribution are handled by the Developmental Testbed Center (DTC)

Methodology

- GSI is a variational(Var) data assimilation system, with hybrid options
- Variational methods are based on the maximum likelihood combination of observation and background information
- It can be shown that the most probable state of the atmosphere given a background X_b and some observations Y is that which minimizes a cost or penalty function J
- The solution obtained is optimal in that it fits the prior (or background) information and measured observations respecting the uncertainty in both

3D-Var Cost Function

Analysis (model state)

Background

Background error covariance

Observations

Observation error covariance

Observation operator

$$J(\mathbf{x}) = \frac{1}{2}(\mathbf{x}-\mathbf{x}_b)^T \mathbf{B}^{-1}(\mathbf{x}-\mathbf{x}_b) + \frac{1}{2}(\mathbf{y} - \mathrm{H}\mathbf{x})^T \mathbf{R}^{-1}(\mathbf{y} - \mathrm{H}\mathbf{x})$$

Background term: J_b

Observation term: J_o

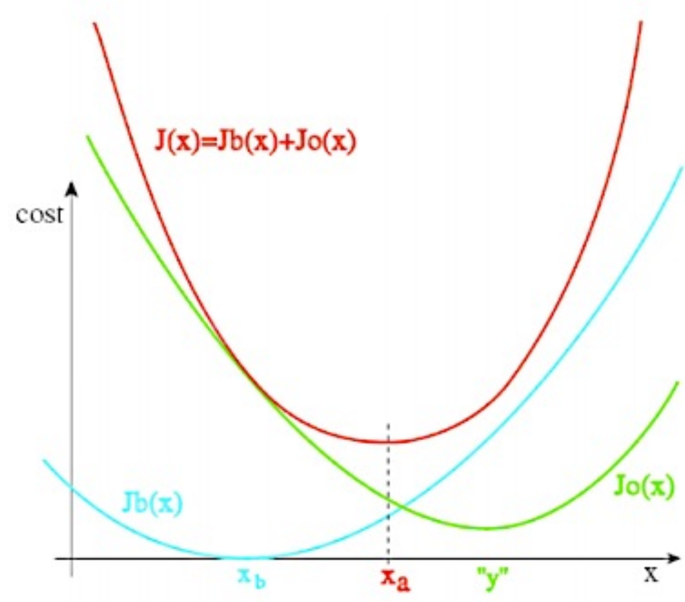
Observations within a specified time window are assimilated at one time

Minimizing Cost Function

- Optimal $\mathbf{x}_a$ is obtained by minimizing the cost function

$$\nabla J(\mathbf{x}) = \mathbf{B}^{-1}(\mathbf{x}-\mathbf{x}_b) - \mathbf{H}^T\mathbf{R}^{-1}(\mathbf{y} - H\mathbf{x}) = 0$$

$\mathbf{H}^T$ is called the Adjoint of the linearized observation operator



Assuming state variable **x** and the final analysis $\mathbf{x}_a$ remains close enough, we can derive

$$\mathbf{x}_a = \mathbf{x}_b + (\mathbf{B}^{-1} + \mathbf{H}^T\mathbf{R}^{-1}\mathbf{H})^{-1}\mathbf{H}^T\mathbf{R}^{-1}(\mathbf{y} - H\mathbf{x}_b)$$

Observation-based correction to background

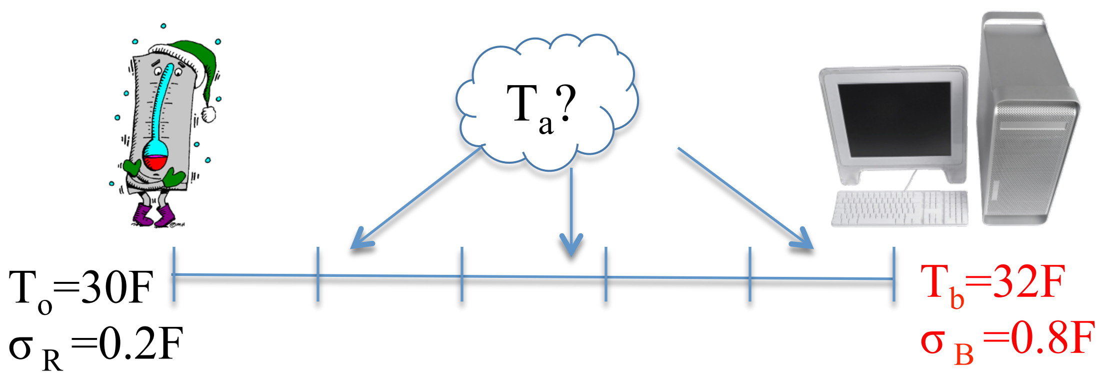
Developmental Testbed Center

Example: One-point Observation

$$J(\mathbf{x}) = \frac{1}{2}(\mathbf{x}-\mathbf{x}_b)^T \mathbf{B}^{-1}(\mathbf{x}-\mathbf{x}_b) + \frac{1}{2}(\mathbf{y}-H\mathbf{x})^T \mathbf{R}^{-1}(\mathbf{y}-H\mathbf{x}) = J_b + J_o$$

A scalar example: **x** here represents the temperature (T) outside

$$J(T) = ½(T-T_b)\sigma_B^{-1}(T-T_b) + ½(T_o-T)\sigma_R^{-1}(T_o-T)$$



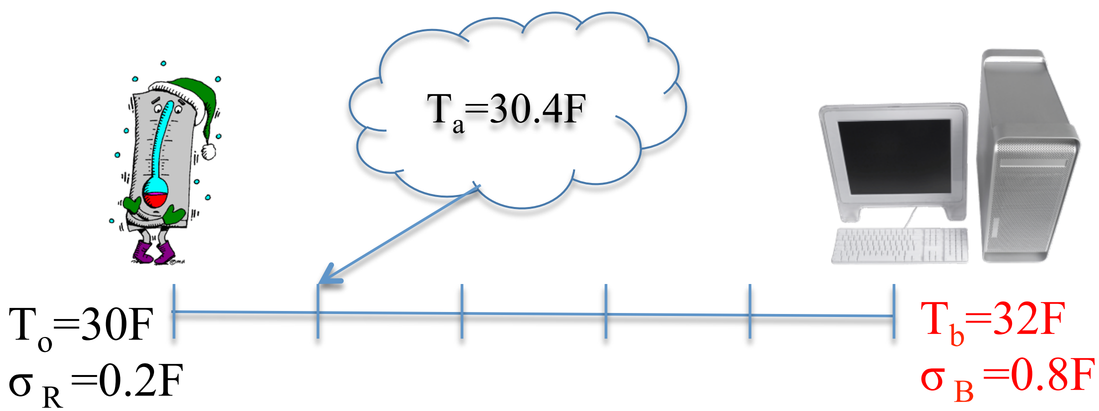
Example: One-point Observation (cont.)

$$\nabla J(\mathbf{x}) = \mathbf{B}^{-1}(\mathbf{x}-\mathbf{x}_b) - \mathbf{H}^T\mathbf{R}^{-1}(\mathbf{y}-\mathrm{H}\mathbf{x}) = 0$$

Scalar example: What is the temperature analysis ($\mathbf{x}_a$->T_a)?

$$\sigma_B^{-1}(T_a - T_b) - \sigma_R^{-1}(T_o - T_a) = 0$$

$$T_a = \sigma_B T_o(\sigma_B + \sigma_R)^{-1} + \sigma_R T_b(\sigma_B + \sigma_R)^{-1} = 30.4\mathrm{F}$$



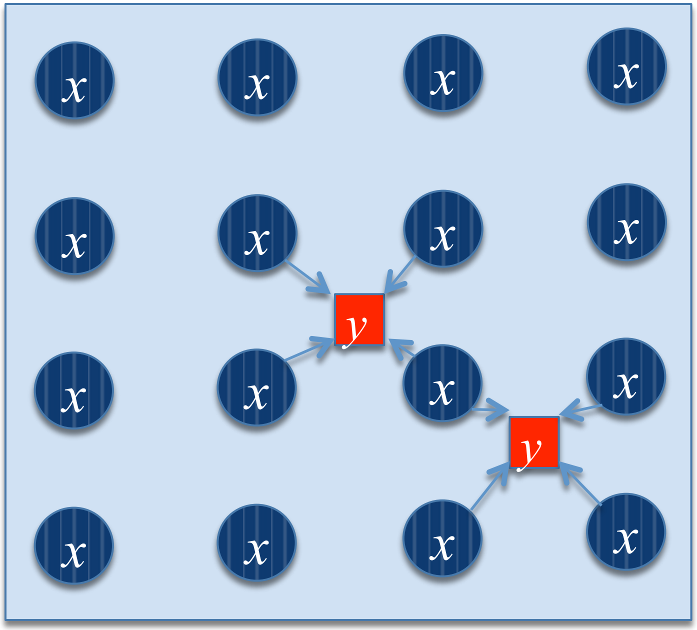
Hypotheses Assumed

- **Linearized observation operator**: the variations of the observation operator in the vicinity of the background state are linear:
  - for any $\mathbf{x}$ close enough to $\mathbf{x}_b$:
    $H(\mathbf{x}) - H(\mathbf{x}_b) = H(\mathbf{x} - \mathbf{x}_b)$, where H is a linear operator
- **Non-trivial errors**: **B** and **R** are positive definite matrices
- **Unbiased errors**: the expectation of the background and observation errors is zero, i.e., $< \mathbf{x}_b\text{-}\mathbf{x}_t > = < \mathbf{y}\text{-}H(\mathbf{x}_t) > = 0$
- **Uncorrelated errors**: observation and background errors are mutually uncorrelated i.e. $< (\mathbf{x}_b\text{-}\mathbf{x}_t)(\mathbf{y}\text{-}H[\mathbf{x}_t])^T > = 0$
- **Linear analysis**: we look for an analysis defined by corrections to the background which depend linearly on background observation departures.
- **Optimal analysis**: we look for an analysis state which is as close as possible to the true state in an r.m.s. sense
  - i.e. it is a minimum variance estimate
  - it is closest in an r.m.s. sense to the true state $\mathbf{x}_t$
  - If the background and observation error pdfs are Gaussian, then $\mathbf{x}_a$ is also the maximum likelihood estimator of $\mathbf{x}_t$

Observation Term (J_o)

$$J(\mathbf{x}) = \frac{1}{2}(\mathbf{x}-\mathbf{x}_b)^T \mathbf{B}^{-1}(\mathbf{x}-\mathbf{x}_b) + \frac{1}{2}(\mathbf{y}-\mathrm{H}\mathbf{x})^T \mathbf{R}^{-1}(\mathbf{y}-\mathrm{H}\mathbf{x}) = J_b + J_o$$

- Observation operator: H
  - Most (traditional measurements)
    - 3D interpolation
  - Some (non-traditional)
    - Complex function, e.g.,
      - Radiance= f(t,q), where f is a radiative transfer model
      - Radar Reflectivity = $f(q_r,q_s,q_h)$
- Observation innovation: **y**-H**x**
- Observation error covariance: **R**
  - Instrument errors + representation errors
  - No correlation between two observations (Typically assumed to be diagonal)

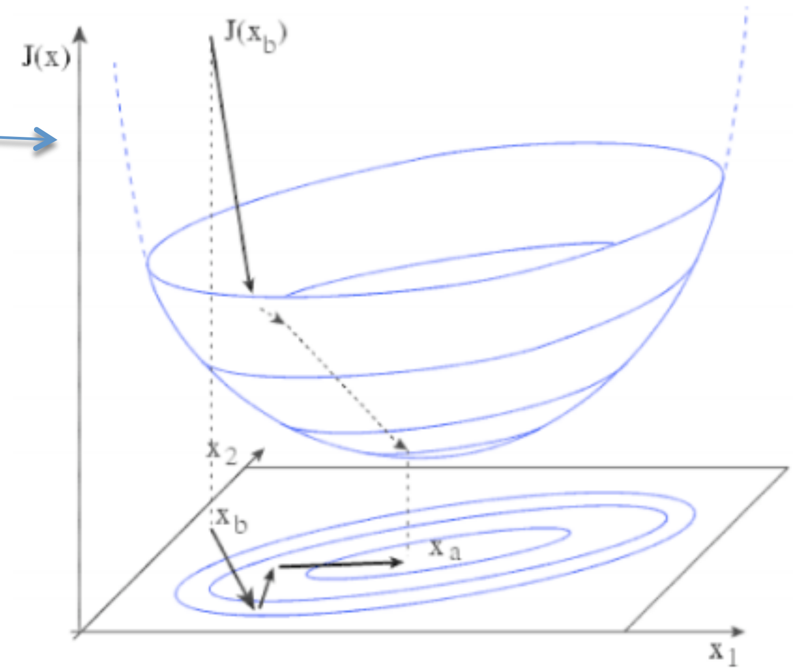


Developmental Testbed Center

Background Term

$$J(\mathbf{x}) = \frac{1}{2}(\mathbf{x}-\mathbf{x}_b)^T \mathbf{B}^{-1}(\mathbf{x}-\mathbf{x}_b) + \frac{1}{2}(\mathbf{y}-\mathrm{H}\mathbf{x})^T \mathbf{R}^{-1}(\mathbf{y}-\mathrm{H}\mathbf{x}) = J_b + J_o$$

- Analysis: **x**
  - Start from $\mathbf{x} = \mathbf{x}_b$
- Analysis increment: $\mathbf{x}-\mathbf{x}_b$
- Background error covariance: **B**
  - Controls influence distance
  - Contains multivariate information
  - Controls amplitude of correction to background
  - For NWP, matrix is prohibitively large
    - Many components are modeled or ignored



"Hybrid" Methods

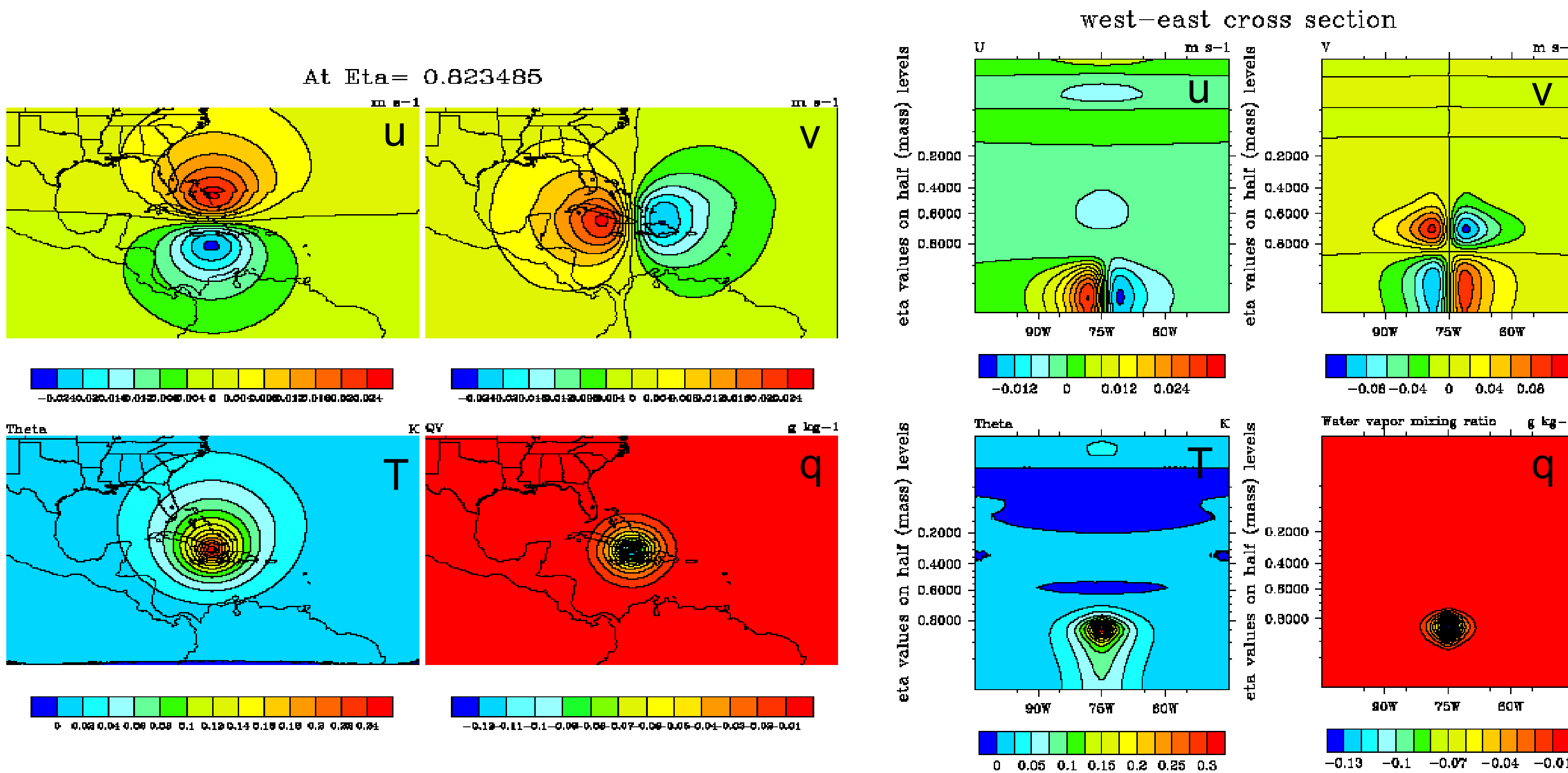
EnVar: Variational methods using **ensemble** background error covariances
Hybrid: **Variational** methods that combine **static** and **ensemble** background error covariances

$$J(\mathbf{x}) = \frac{\beta}{2}(\mathbf{x}-\mathbf{x}_b)^{T}\mathbf{B}_{\mathrm{Var}}^{-1}(\mathbf{x}-\mathbf{x}_b) + \frac{1-\beta}{2}(\mathbf{x}-\mathbf{x}_b)^{T}\mathbf{B}_{Ens}^{-1}(\mathbf{x}-\mathbf{x}_b) + J_o$$

- **B_{Var}**: (Static) background error (BE) covariance matrix(estimated offline)
- **B_{Ens}**: (Flow dependent) background error covariance matrix (estimated from ensemble at each analysis time)
- β: Associated with relative weightings of B_{Var} and B_{Ens}

Traditional Variational DA: $\mathbf{B}_{Var}$

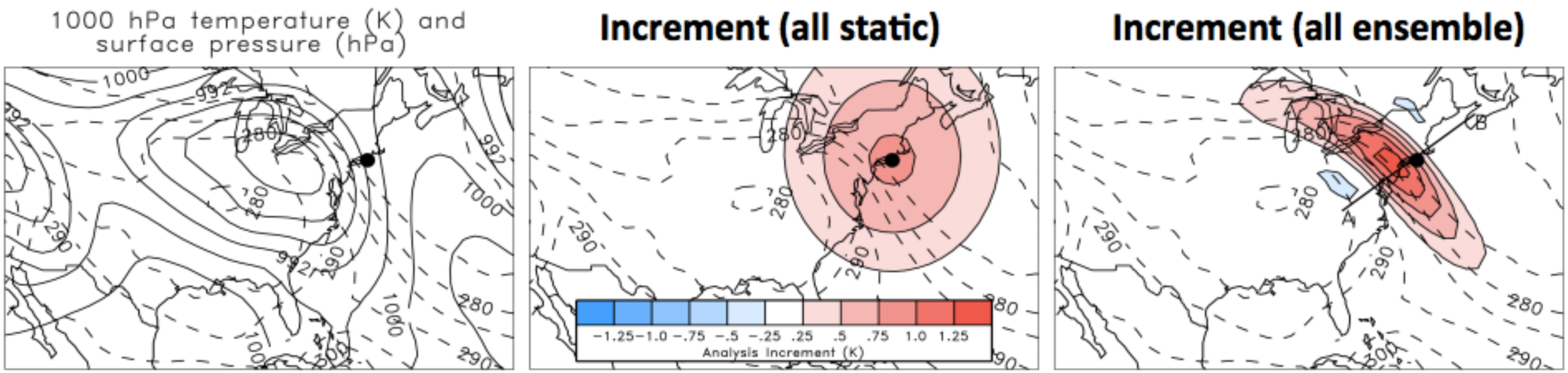
Analysis increments due to single observation



Reflection of static background error covariances:
Error, horizontal scale, vertical scale, univariable/cross-variable correlation

What Does $\mathbf{B}_{Ens}$ Do?

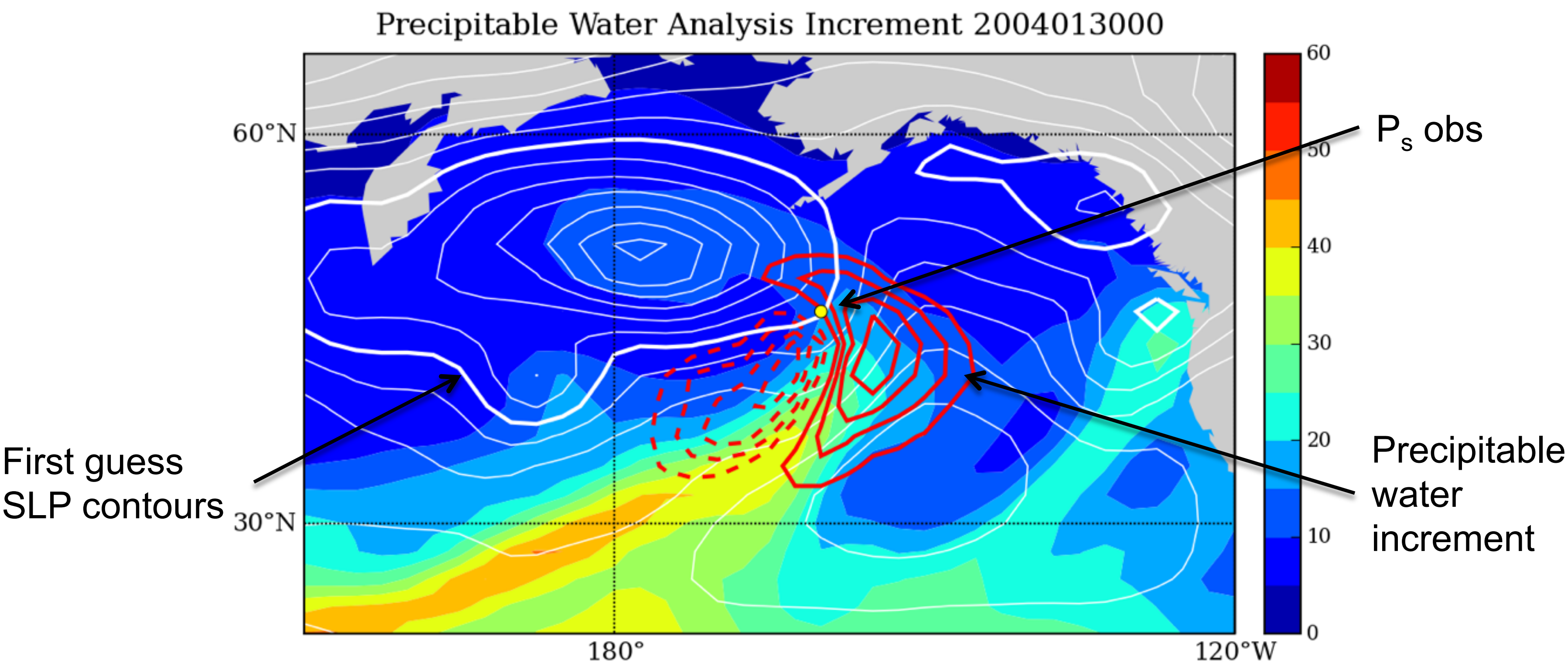
Temperature observation near a warm front



✓ Allows for flow-dependence/errors of the day

What Does $\mathbf{B}_{Ens}$ Do?

Surface pressure observation near "atmospheric river"



3D-Var increment would be zero

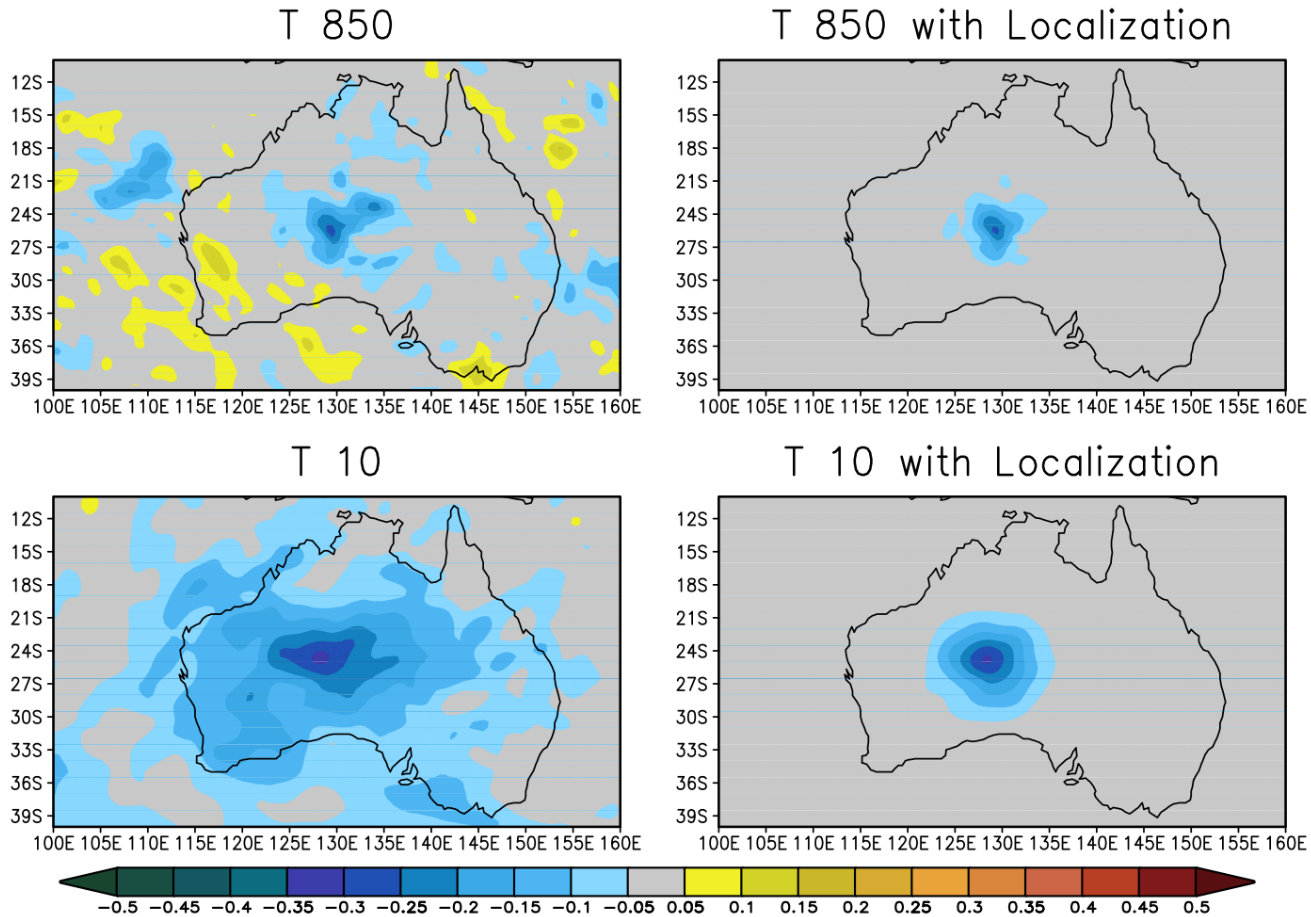
(Cross-variable covariances hard to model with static $\mathbf{B}_{var}$)

How Does $\mathbf{B}_{Ens}$ Benefit 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

But $\mathbf{B}_{Ens}$ is not perfect…at least not yet!

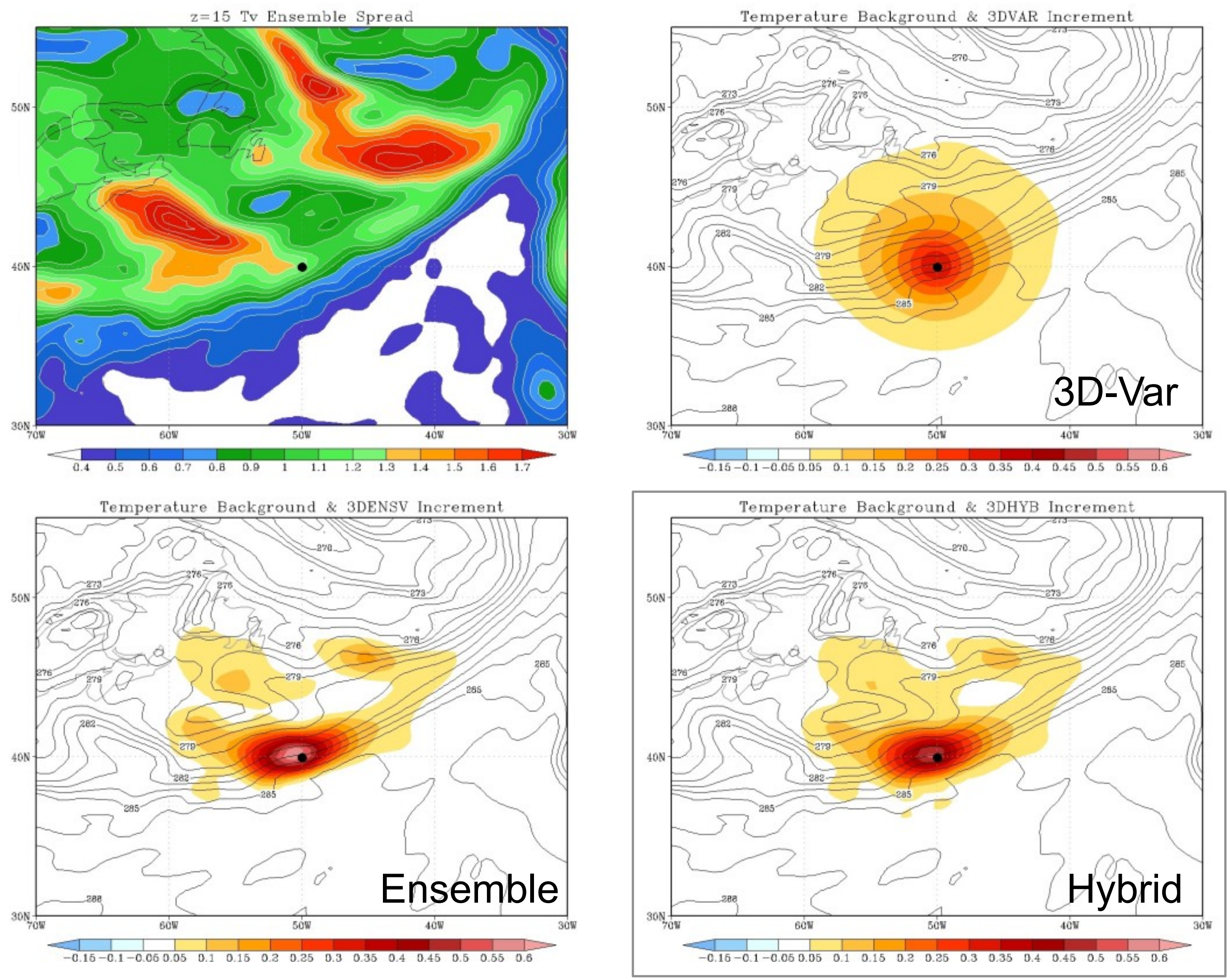
Localization of $\mathbf{B}_{Ens}$



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

Single Temperature Observation

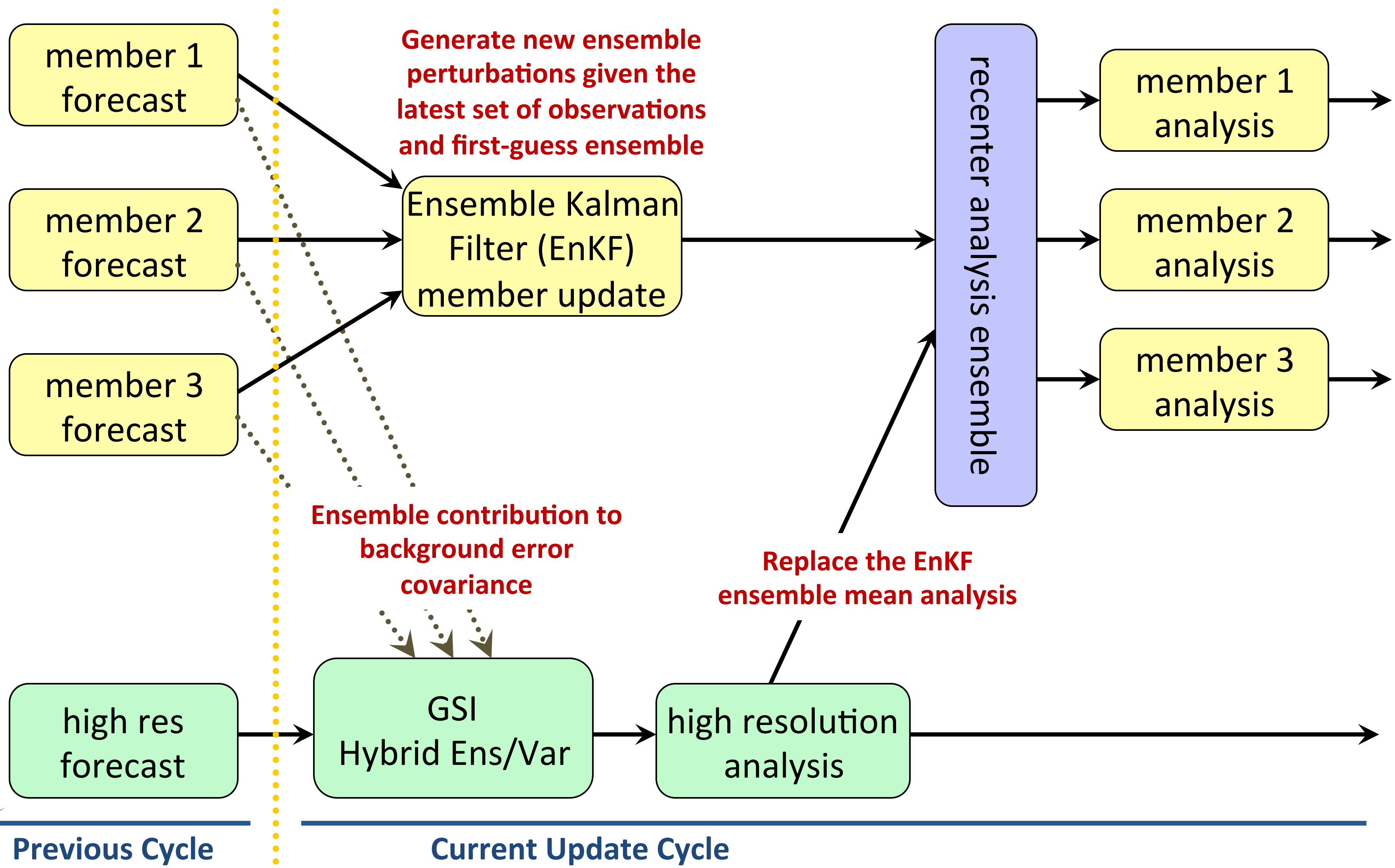


So What's the Catch?

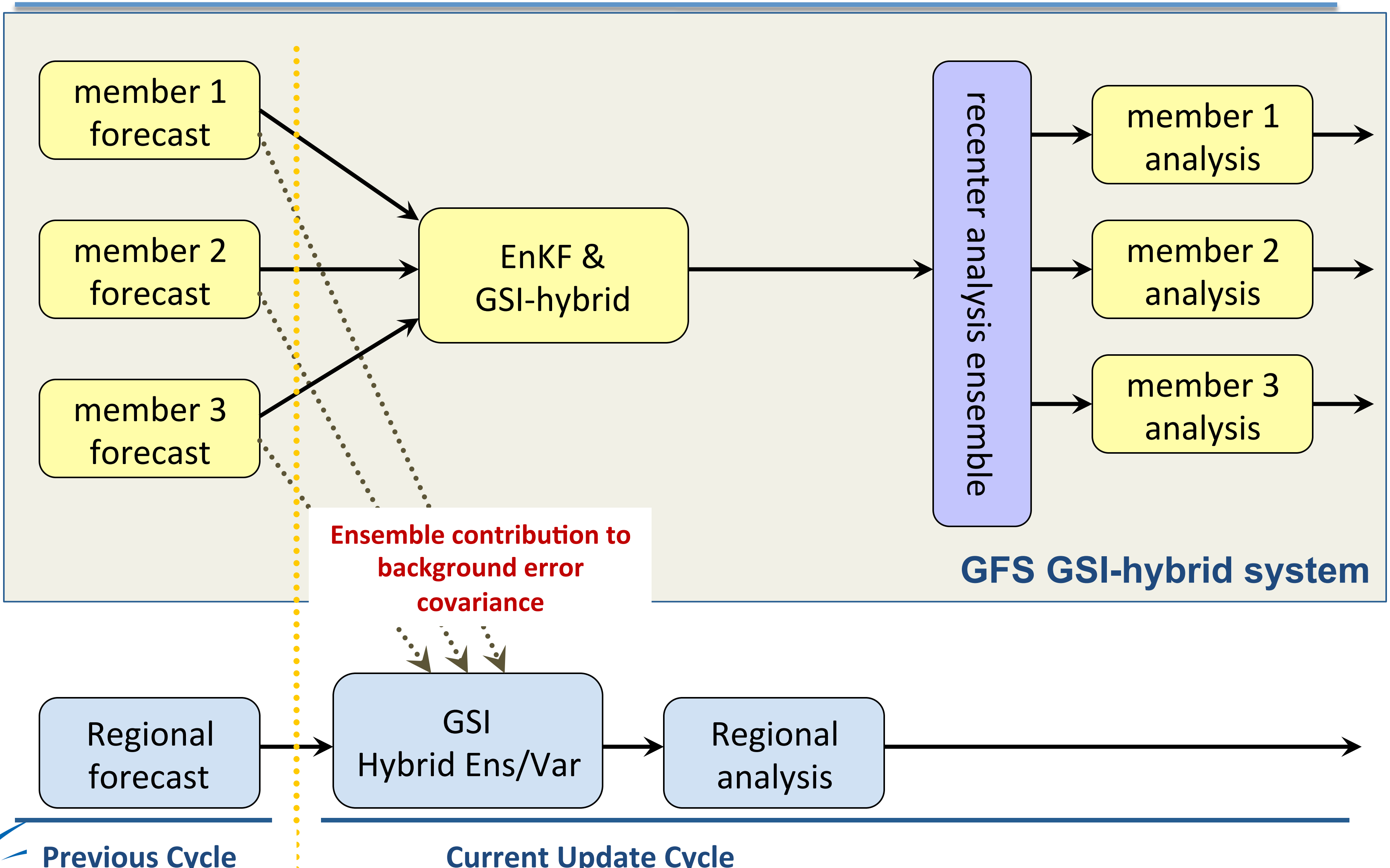
- Need an ensemble that represents first guess uncertainty (background error)
  - In principle, any ensemble can be used. However, ensemble should represent well the forecast errors
- This can mean O(50-100+) for NWP applications
  - Smaller ensembles have larger sampling error (rely more heavily on $\mathbf{B}_{Var}$)
  - Larger ensembles have increased computational expense
- Updating the ensemble (NCEP) (up to 2015)
  - Global: an Ensemble Kalman Filter is currently used for NCEP Global Forecasting System (GFS)
  - Regional: using the GFS ensemble generated by the GFS & GSI-hybrid system at each analysis time (ensemble members are updated during the GFS cycle)

.

Coupled GSI-Hybrid Cycling (GFS)



Current Scheme for Regional GSI-hybrid



Adding Time Dimension

$$J(\mathbf{x}) = \frac{1}{2}(\mathbf{x}-\mathbf{x}_b)^T \mathbf{B}^{-1}(\mathbf{x}-\mathbf{x}_b) + \frac{1}{2}\sum_{k=1}^{K}\left(\mathbf{y}_k - H_k M_{0\rightarrow k}(\mathbf{x}_0)\right)^T \mathbf{R}_k^{-1}\left(\mathbf{y}_k - H_k M_{0\rightarrow k}(\mathbf{x}_0)\right)$$

M: forecast model
k: observation time

4D-Var: using forecast model and its adjoint model for the Kalman Gain (observation based correction to background)
4D EnVar: using model ensemble forecast to replace the temporal propagation of perturbations by the tangent linear model and its adjoint. Backgorund
Hybrid 4D EnVar: Same as 4D EnVar, except the background error covariance is a combination of both static and ensemble

Hybrid 4D EnVar is being tested for GFS implementation

Community GSI

- Objective:
  - Provide current operational GSI capabilities to the research community (O2R) and a pathway for the research community to contribute to operational GSI (R2O)
  - Provide a framework to enhance the collaboration from distributed GSI developers
- GSI Code support:
  - Community GSI repository
  - User's webpage
  - Annual code release with user's guide
  - Annual residential tutorial
  - Help desk
- GSI code management
  - Unified code review-commit procedure
  - Development coordinated through GSI Review Committee

Public code distribution and support are through the Developmental Testbed Center (DTC)

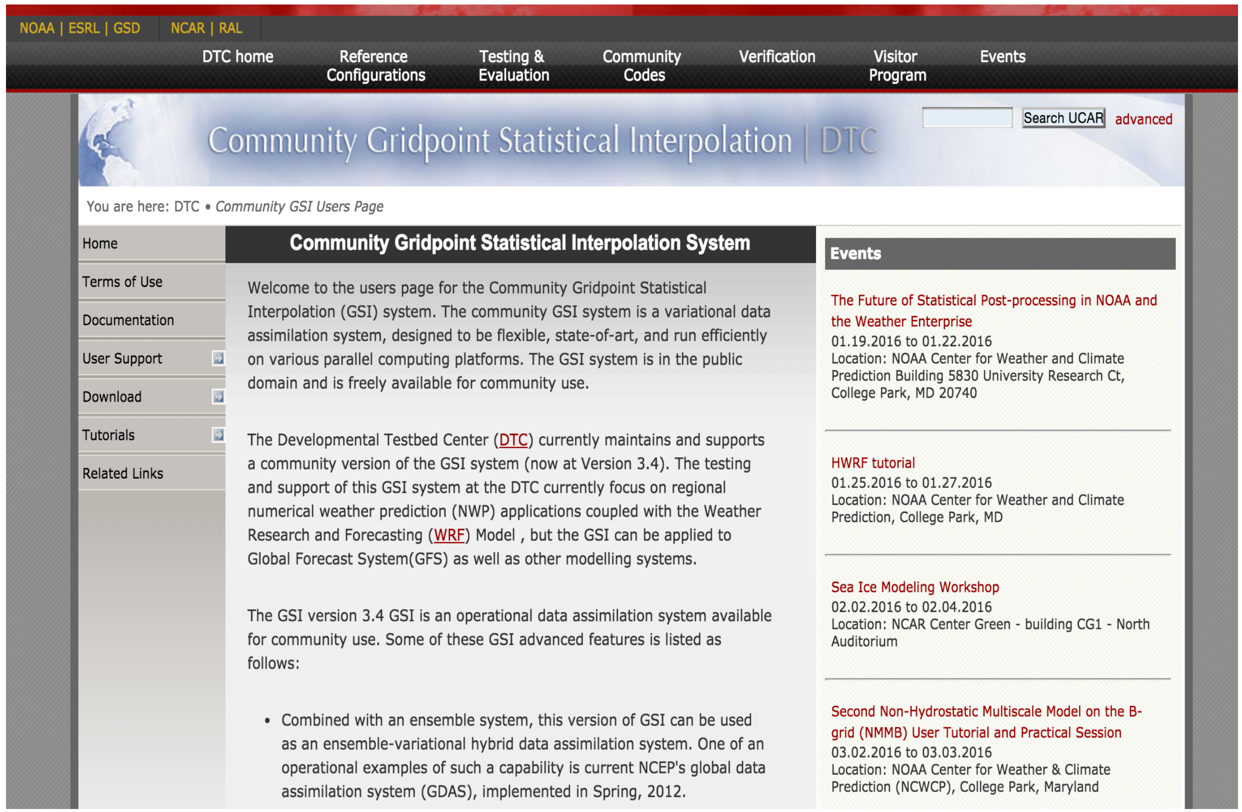
Community GSI Public Release

- Stand alone release
  - Latest version is V3.4 (July, 2015)
  - Next release is planned for Summer 2016
  - Currently GSI is released together with another ensemble based DA system (GSI and EnKF)
- HWRF release
  - Include GSI and other compoments

DTC supports
- GSI community users: gsi_help@ucar.edu
- HWRF: wrfhelp@ucar.edu

Community GSI Resources

- Code download (stand alone)
- User's Guide
  - Match each official release
- Workshop presentations
- Tutorial lectures
- Online tutorial and practical cases
- Code browser



http://www.dtcenter.org/com-GSI/users/index.php

Reference

- Data Assimilation Concept and Methods (ECMWF Training Course, Bouttier & Courtier)
- GSI Tutorial Lectures:
  - GSI overview (John Derber, Ming Hu)
  - Fundamentals of Data Assimilation (Tom Auligne)
  - Background and Observation Errors (Daryl Kleist)
  - GSI Hybrid Data Assimilation (Jeff Whitaker, Daryl Kleist)
  - Aerosol Data Assimilation (Zhiquan Liu)