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

Hui Shao

Joint Center for Satellite Data Assimilation (JCSDA)

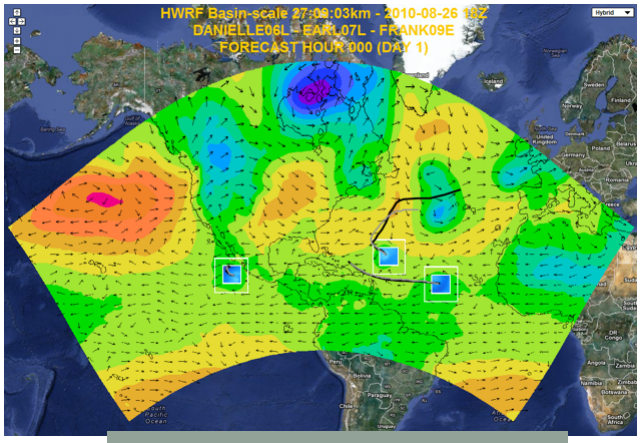
2018 Hurricane WRF Tutorial, Jan 23-25, 2018, College Park, MD

Outline

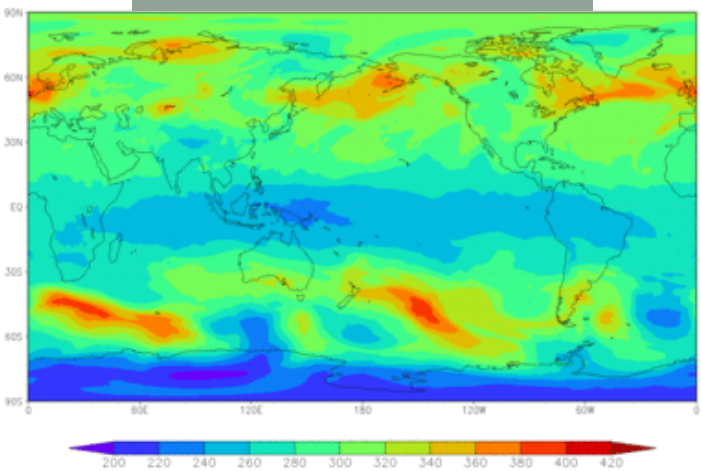
- Background of data assimilation
- GSI concepts and methods
- Community support and service

Numerical Weather Prediction (NWP)

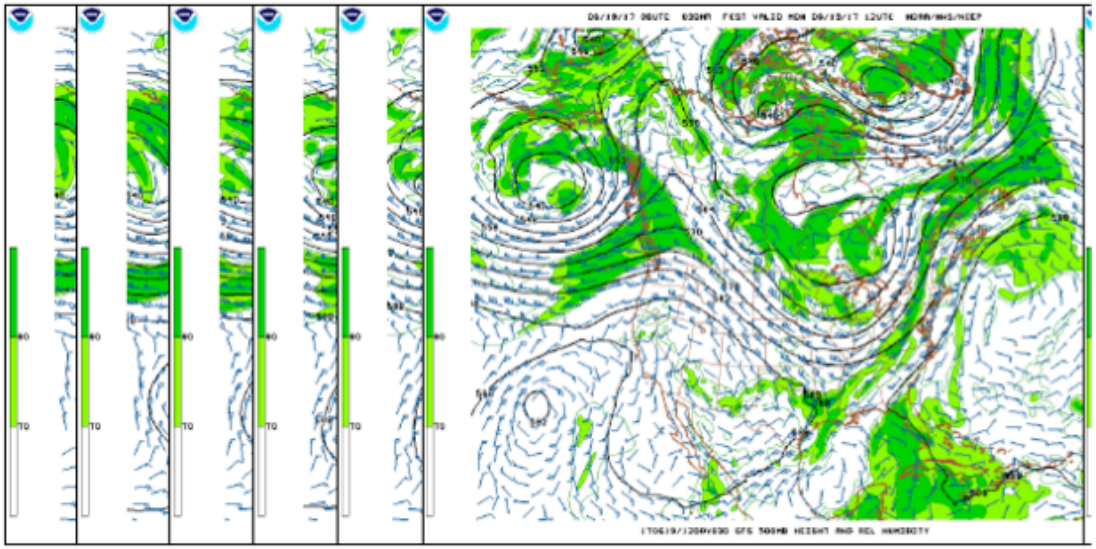
Given an estimate of the current state of the atmosphere (initial conditions), and appropriate surface (and lateral, if regional) boundary conditions, the computational model simulates the atmospheric evolution (forecasts).



Hurricane WRF forecasts



GFS Total Ozone



NOAA Global Forecast System (GFS) forecasts (hours)

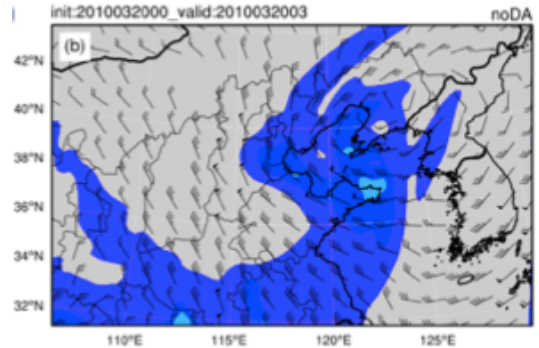
Target

- As accurately as possible for the longest lead time possible
- With uncertainty (confidence) estimates

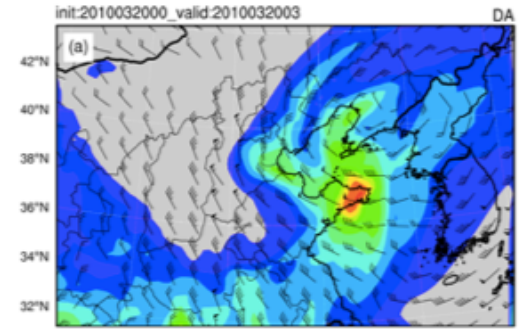
(Courtesy from various NOAA webpages)

Data Assimilation (DA)

+



=



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 $t=t_{k-1}$)

Initial conditions

Forecast model

Forecast (x_k)

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?

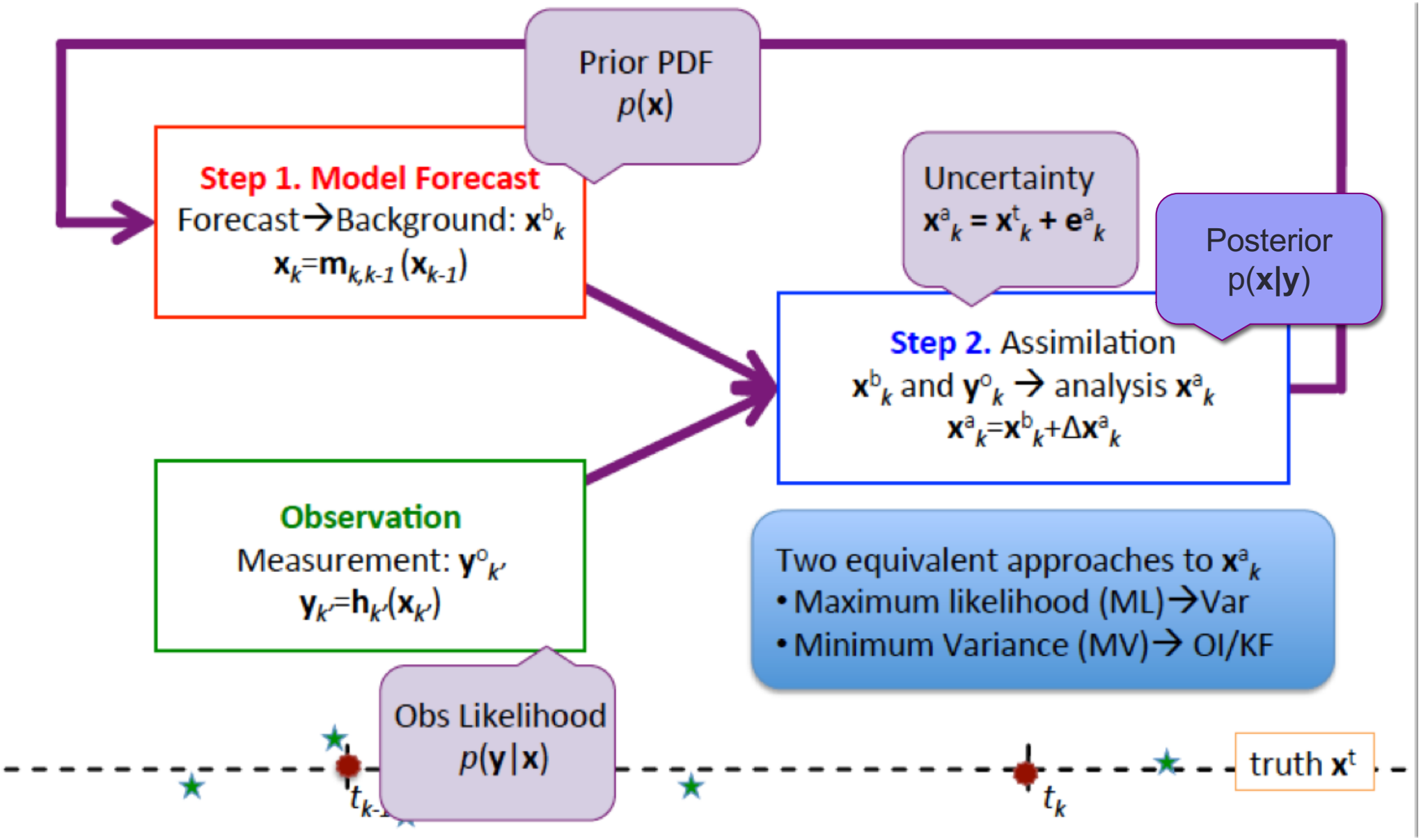
From Empirical to Statistical Methods

- Successive Correction Method
- Nudging
- Physical Initialization (PI), Latent Heat Nudging (LHN)

- Relaxation functions are somewhat arbitrary
- Good forecast can be replaced by bad observations
- Noisy observations can create unphysical analysis

- Modern DA techniques are usually statistical, e.g.,
  - Variational
  - Kalman filter based

Two Approaches to DA



Methodology: 3D-Var

- p(x|y) is maximized given p(x) & p(y|x) using Bayes theorem

  p(x|y)=p(y|x)p(x)/p(y)

- x is obtained by minimizing the cost function J(x)

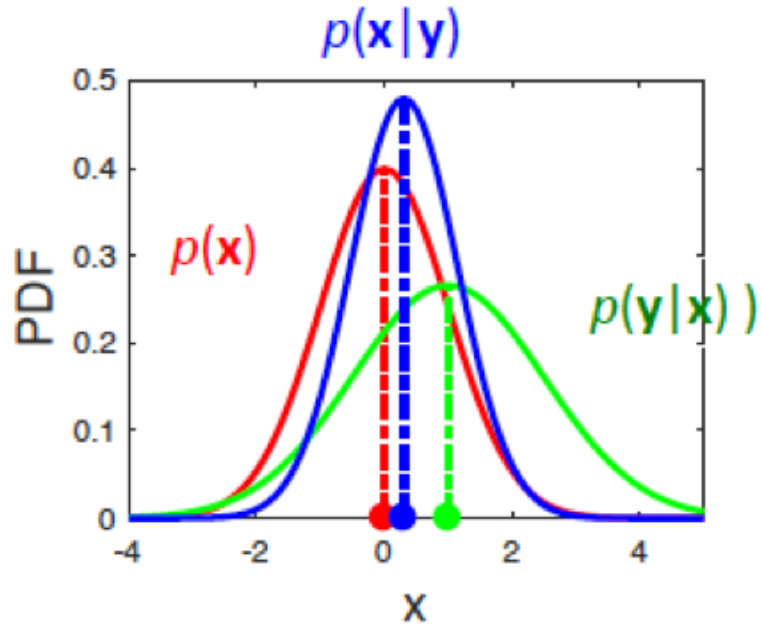
$$J(\mathbf{x}) = \frac{1}{2}(\mathbf{x}-\mathbf{x}_b)^T \mathbf{B}^{-1}(\mathbf{x}-\mathbf{x}_b) + \frac{1}{2}(\mathbf{y}-\mathrm{H}\mathbf{x})^T \mathbf{R}^{-1}(\mathbf{y}-\mathrm{H}\mathbf{x})$$

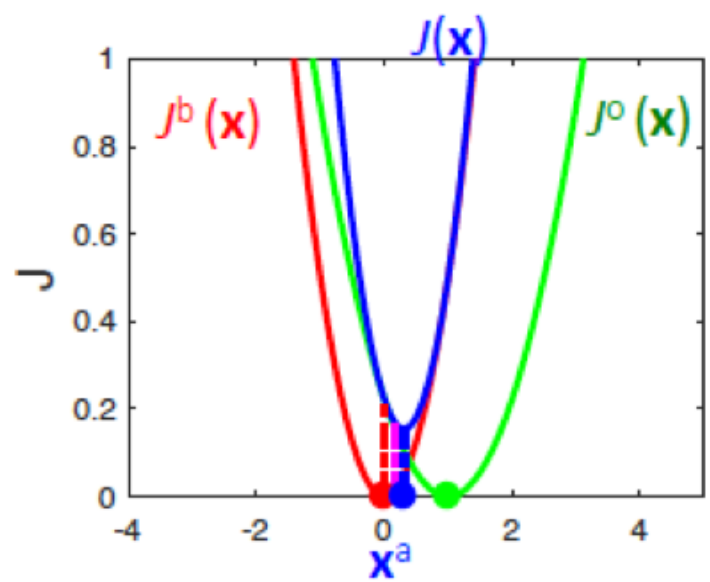
$$= J_b + J_o$$

Minimizing

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

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





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., $< \mathbf{x}_b - \mathbf{x}_t >= < \mathbf{y} - H(\mathbf{x}_t) > = 0$
- **Uncorrelated errors**: observation and background errors are mutually uncorrelated i.e. $< (\mathbf{x}_b - \mathbf{x}_t)(\mathbf{y} - H[\mathbf{x}_t])^T >= 0$
- **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 $\mathbf{x}_t$

Gridpoint Statistical Interpolation (GSI)

- GSI is a variational(Var) data assimilation system, with hybrid options
  - 2D-Var: static background error ($\mathbf{B}_{var}$)
  - 3D-Var: static $\mathbf{B}_{var}$
  - 3D EnVar: using ensemble estimated $\mathbf{B}_{Ens}$
  - 3D Hybrid EnVar: using combination of $\mathbf{B}_{var}$ and $\mathbf{B}_{Ens}$
  - 4D (hybrid) EnVar: ensemble is used to estimate $\mathbf{B}_{Ens}$ & 4D increments
- Original developed by NCEP based on the previous Spectral Statistical Interpolation (SSI) analysis system, replacing spectral definition for background error with grid point version based on recursive filters
- Used in NOAA/NCEP operations: regional, global, hurricane, real-time mesoscale analysis, rapid refresh
- Operational at the Air Force
- NASA/GMAO operations
- Modification to fit into FV3, WRF and NCEP infrastructure
- Evolution to Joint Effort for Data Assimilation Integration (JEDI) (lead by JCSDA)

Input and Output

Analysis (model state)

Background

Background error covariance

Observations

Observation error covariance

Observation operator

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

Background term: J_b

Observation term: J_o

Analysis (x)

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

- Analysis variables
  - Streamfunction (Ψ)
  - Unbalanced Velocity Potential (χ')
  - Unbalanced Temperature (T')
  - Unbalanced Surface Pressure (Ps')
  - Ozone – Clouds – etc.
  - Satellite bias correction coefficients

- Currently $\mathbf{X}_b$ from following systems
  - NCEP GFS
  - NCEP NMM(B) – binary and netcdf
  - NCEP RTMA
  - NCEP Hurricane
  - NOAA FV3
  - GMAO FV3 global
  - ARW – binary and netcdf

- Size of problem
  - N_X x N_Y x N_Z x N_{VAR}
  - Global = 600 million component control vector
  - Requires multi-tasking to fit on computers
  - Also, often requires analysis increment to be done at a different resolution than the model to achieve run-time requirements.
    - GSI has 3 resolutions, analysis, background and ensemble.

Background Errors (B)

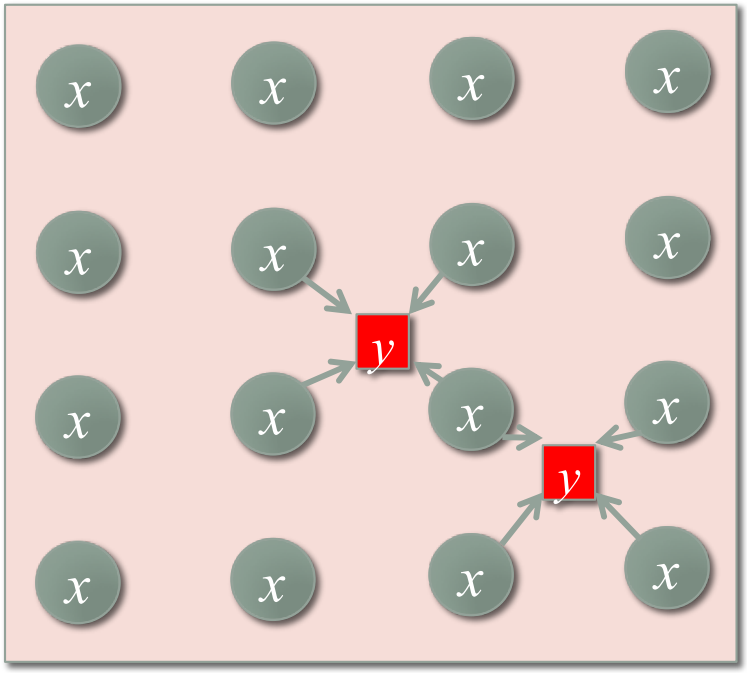
$$J(\mathbf{x}) = \frac{1}{2}(\mathbf{x}-\mathbf{x}_b)^T \boxed{\mathbf{B}^{-1}} (\mathbf{x}-\mathbf{x}_b) + \frac{1}{2}(\mathbf{y}-\mathrm{H}\mathbf{x})^T \mathbf{R}^{-1} (\mathbf{y}-\mathrm{H}\mathbf{x}) = J_b + J_o$$

- Three options in GSI
  - Isotropic/homogeneous
    - Most common usage
    - Function of latitude/height
    - Vertical and horizontal scales separable
    - Variances can be location dependent
  - Anisotropic/inhomogeneous
    - Function of location /state
    - Can be full 3-D covariances
    - Still relatively immature
  - Hybrid
    - Includes ensembles along with one of the two options above

Observation Term (J_o)

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

- Observational data (**y**) is expected to be in BUFR format (WMO standard)
  - PrepBUFR (with NCEP flavored header): conventional
  - BUFR: satellite
- Observation operator: H
  - Conversion from model space to observational space
  - 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)
  - Should be tuned for each observation type/instrument



Input data

Conventional

- Radiosondes
- Pibal winds
- Synthetic tropical cyclone winds
- wind profilers
- conventional aircraft reports
- ASDAR aircraft reports
- MDCARS aircraft reports
- dropsondes
- MODIS IR and water vapor winds
- GMS, JMA, METEOSAT and GOES cloud drift IR and visible winds
- GOES water vapor cloud top winds
- Surface land observations
- Surface ship and buoy observation
- SSM/I wind speeds
- QuikScat and ASCATwind speed and direction
- SSM/I and TRMM TMI precipitation estimates
- Doppler radial velocities
- VAD (NEXRAD) winds
- GPS precipitable water estimates
- GPS Radio occultation refractivity and bending angle profiles
- SBUV ozone profiles and OMI total ozone

Satellite radiance

- AMSU-A
  - NOAA-15 Channels 1-5,7-10, 12-13, 15
  - NOAA-18 Channels 1-4,6-7, 10-13, 15
  - NOAA-19 Channels 1-6, 9-13, 15
  - METOP-A Channels 1-6, 9-13, 15
  - METOP-B Channels 7-11(13)
  - AQUA Channels 6, 8-13
- ATMS
  - NPP Channels 1-11,16-19
- MHS
  - NOAA-19 Channels 1-2,4-5
  - METOP-A Channels 1-5
  - METOP-B Channels 1-5
- Geo Sounder
  - GOES-15 Channels 1-15
  - SEVERI M10 Channels 2-3
- Hyperspectral
  - AIRS AQUA 148 Channels
  - IASI METOP-A 165 Channels
  - IASI METOP-B 165 Channels
  - NPP CrIS 84 Channels
- SSMIS
  - F17 Channels 1-3,5-7,24

"Hybrid" EnVar Methods

EnVar: Variational methods using **ensemble** background error covariances

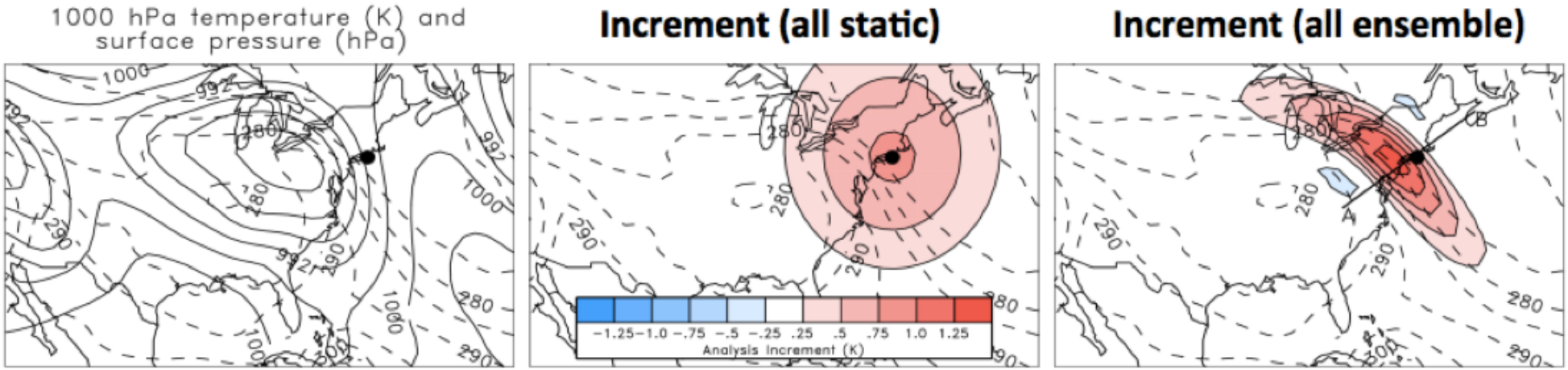
Hybrid: **Variational** methods that combine **static** and **ensemble** background error covariances

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

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

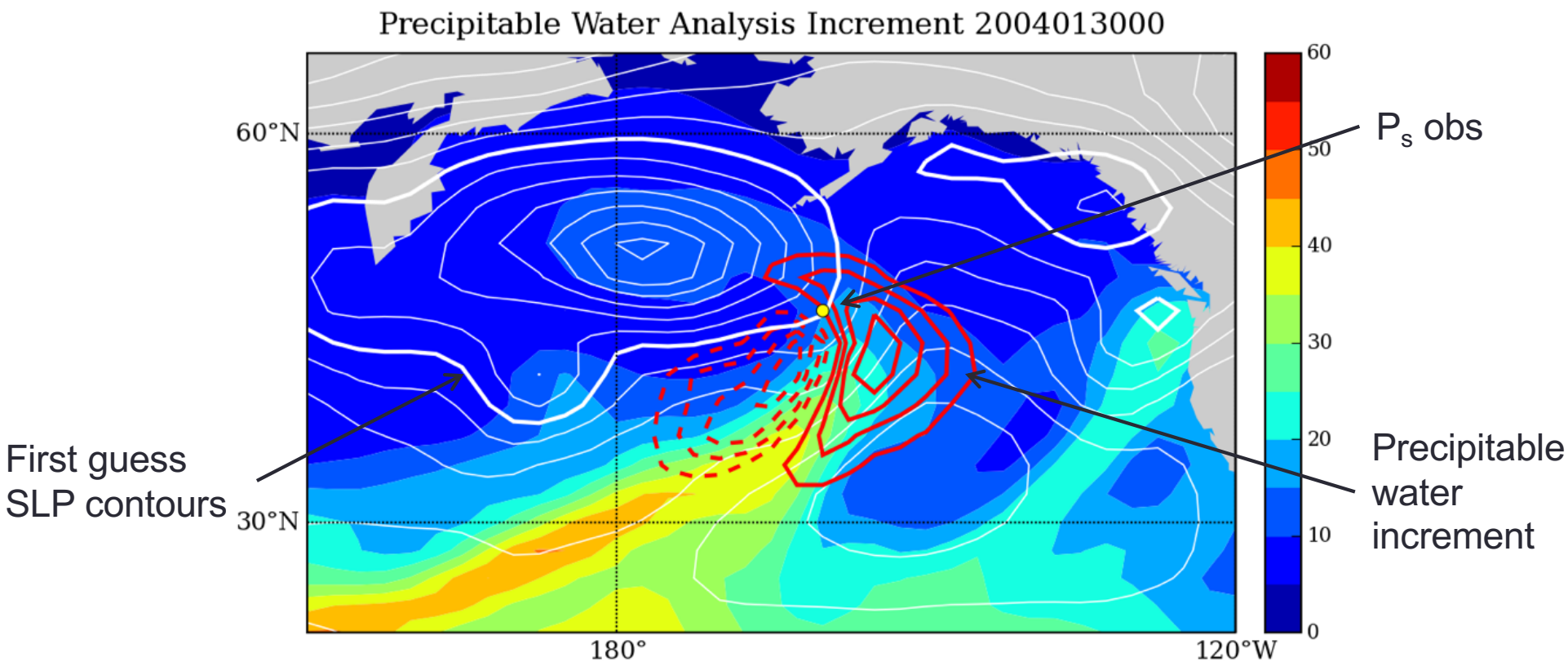
What Does $\mathbf{B}_{Ens}$ Do?

Temperature observation near a warm front



✓ Allows for flow-dependence/errors of the day

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



3D-Var increment would be zero

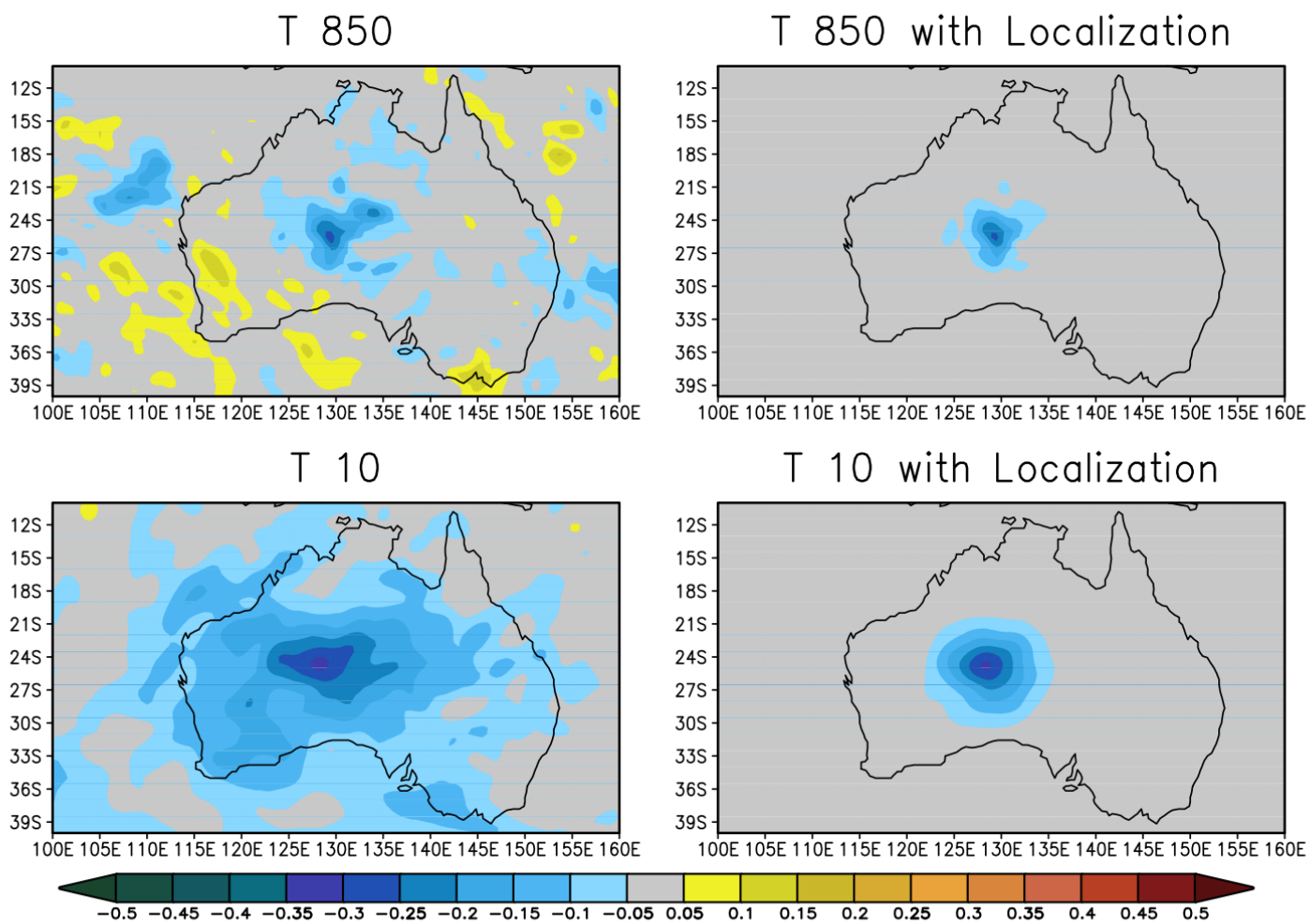
(Cross-variable covariances hard to model with static $\mathbf{B}_{var}$)

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!

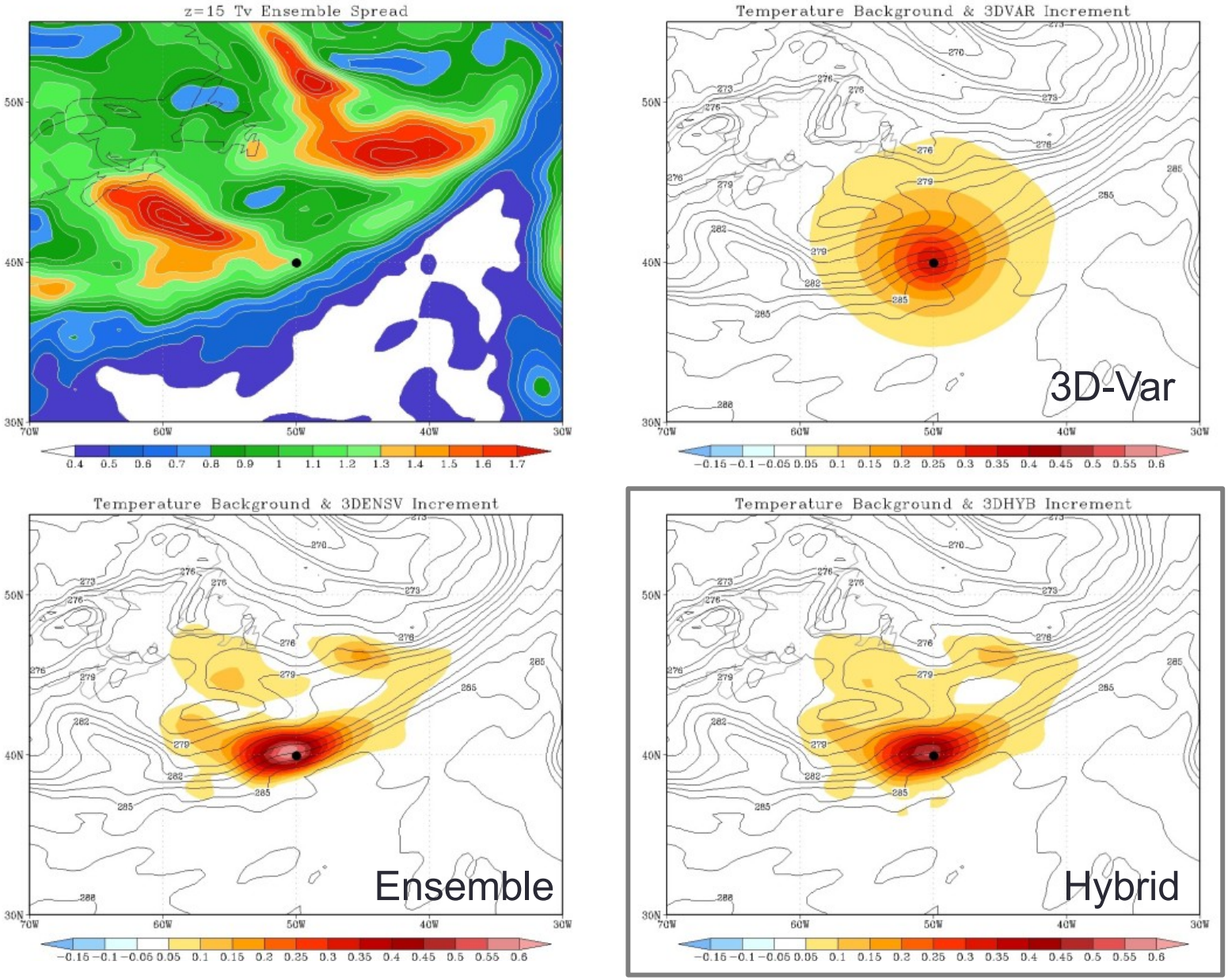
Localization of $\mathbf{B}_{Ens}$



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

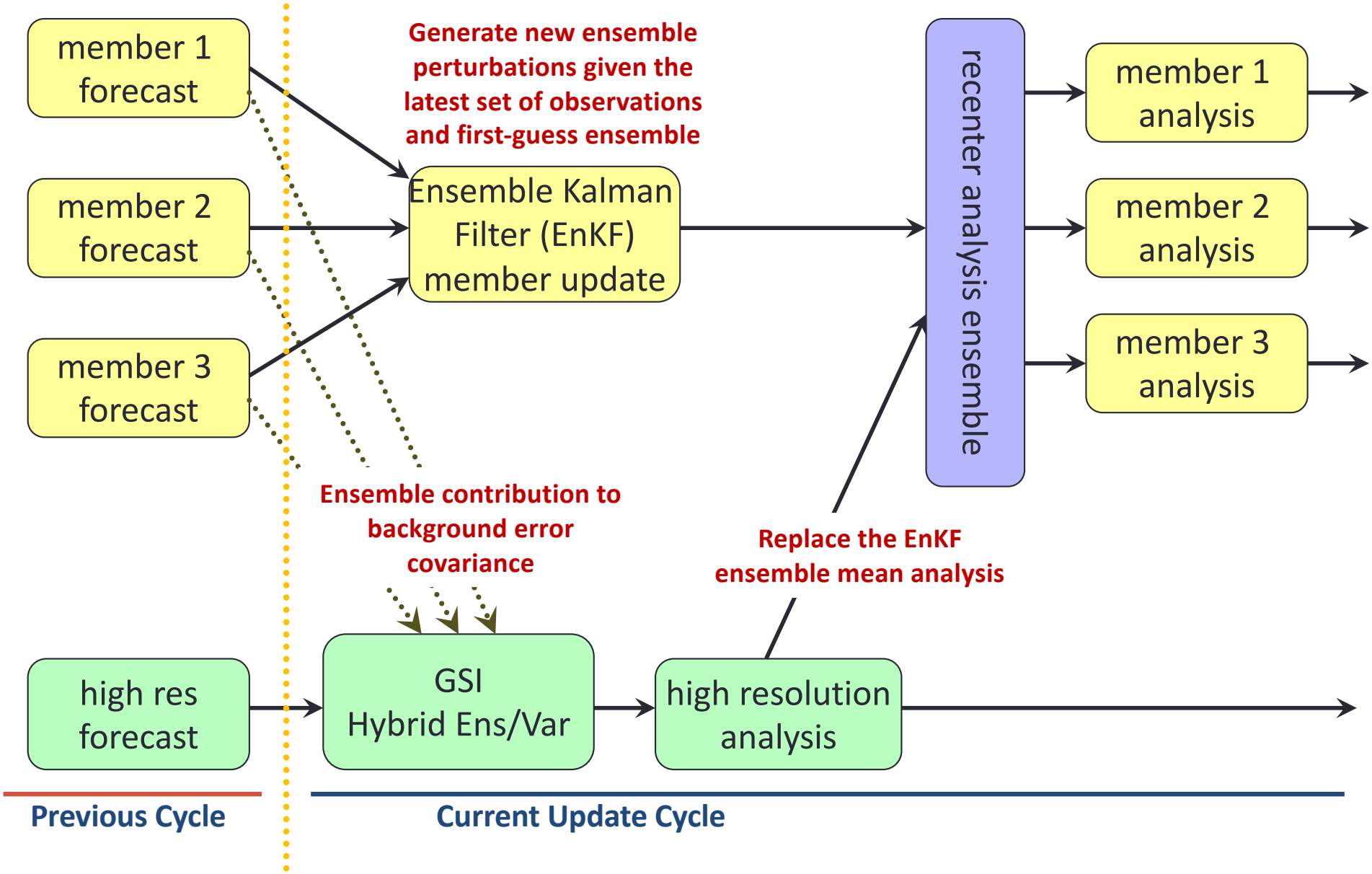
Single Temperature Observation Tests



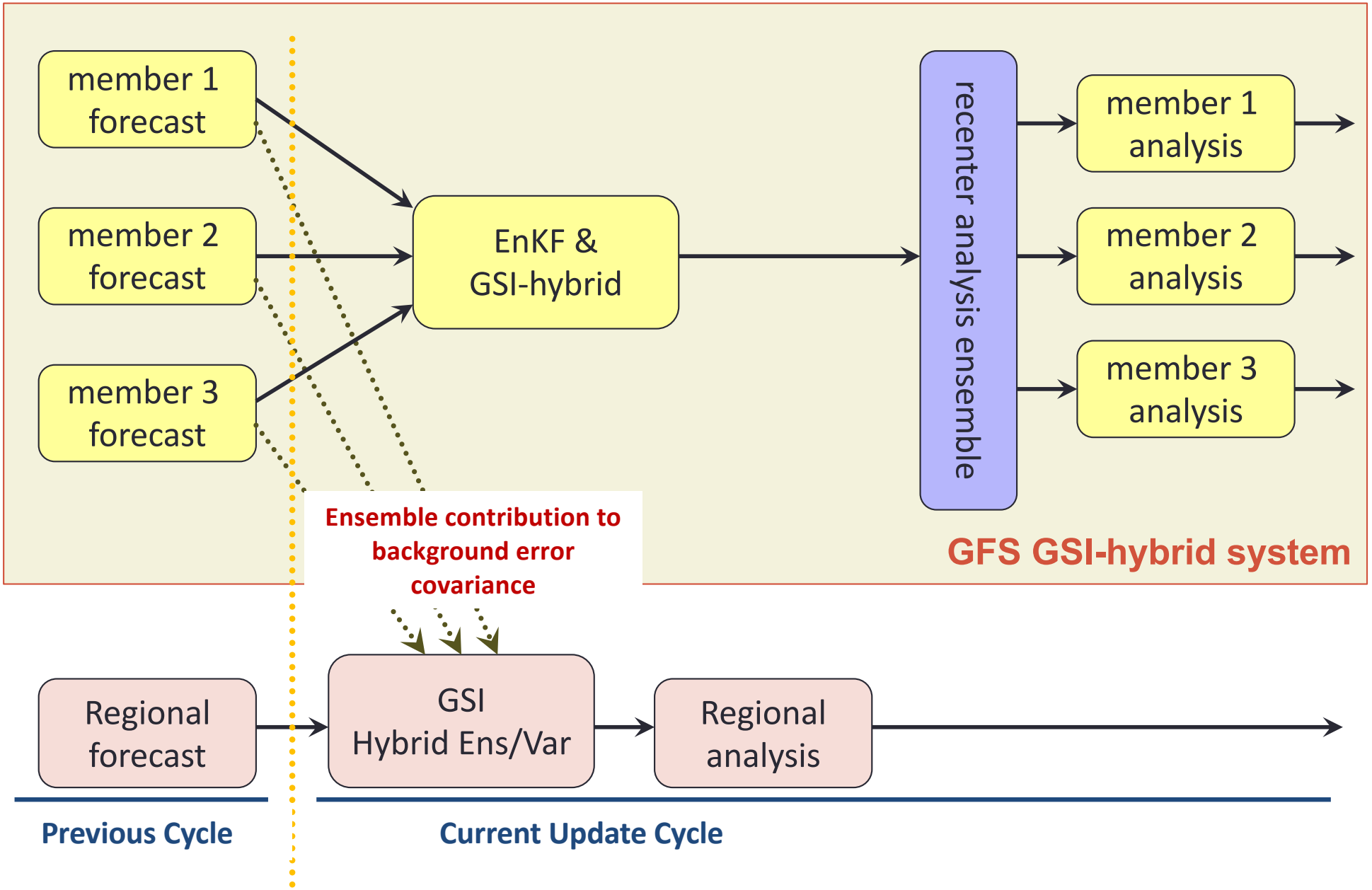
So What's the Catch?

- 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 (how NCEP does it?)
  - Global: an Ensemble Kalman Filter is currently used for NCEP 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

M: forecast model
k: observation time

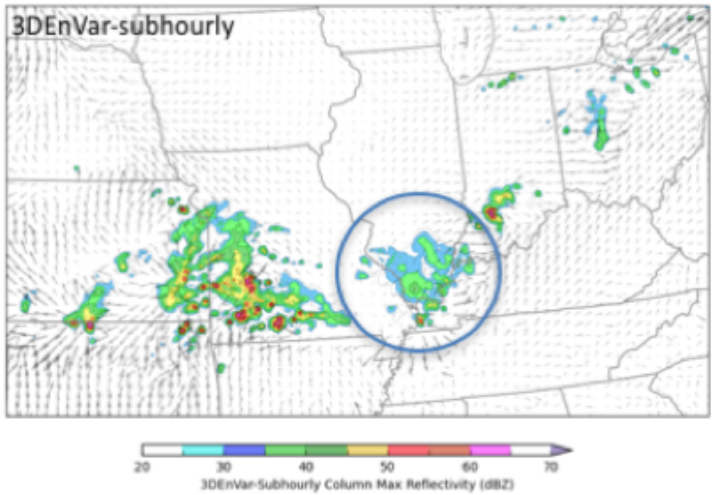
$$J(\mathbf{x}) = \frac{\beta}{2}\left(\mathbf{x}-\mathbf{x}_b\right)^{\mathrm{T}} \mathbf{B}_{Var}^{-1}\left(\mathbf{x}-\mathbf{x}_b\right) + \frac{1-\beta}{2}\left(\mathbf{x}-\mathbf{x}_b\right)^{\mathrm{T}} \mathbf{B}_{Ens}^{-1}\left(\mathbf{x}-\mathbf{x}_b\right)$$

$$+\frac{1}{2}\sum_{k=1}^{K}\left(\mathbf{y}_k - \mathrm{H}_k \mathrm{M}_{0\to k}(\mathbf{x}_0)\right)^{\mathrm{T}} \mathbf{R}_k^{-1}\left(\mathbf{y}_k - \mathrm{H}_k \mathrm{M}_{0\to k}(\mathbf{x}_0)\right)$$

- 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.
- Hybrid 4D EnVar: Same as 4D EnVar, except the background error covariance is a combination of both static and ensemble

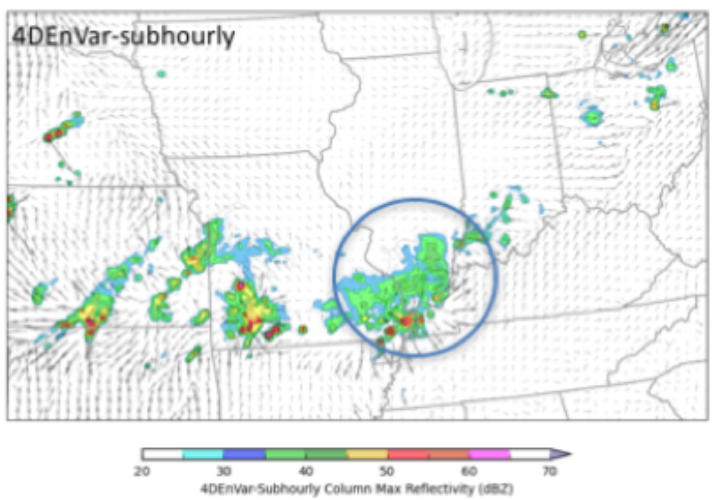
4D Experiments for Regional Applications

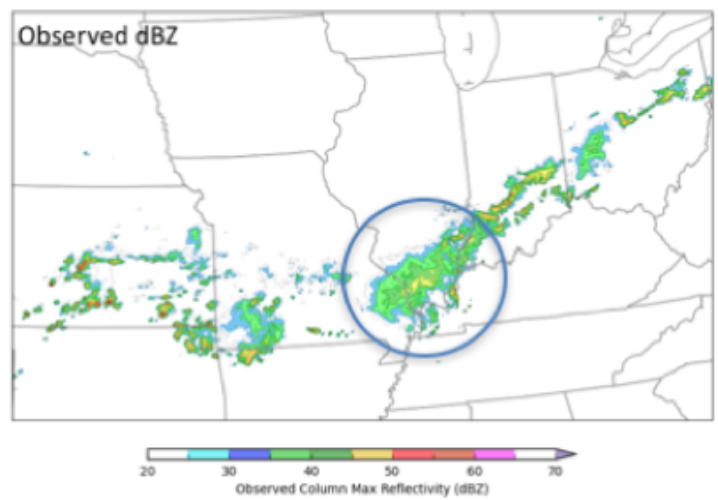
3D hybrid



Forecasts @2016090817 after 1-day of cycles (DA updates per 15 mins)

4D hybrid





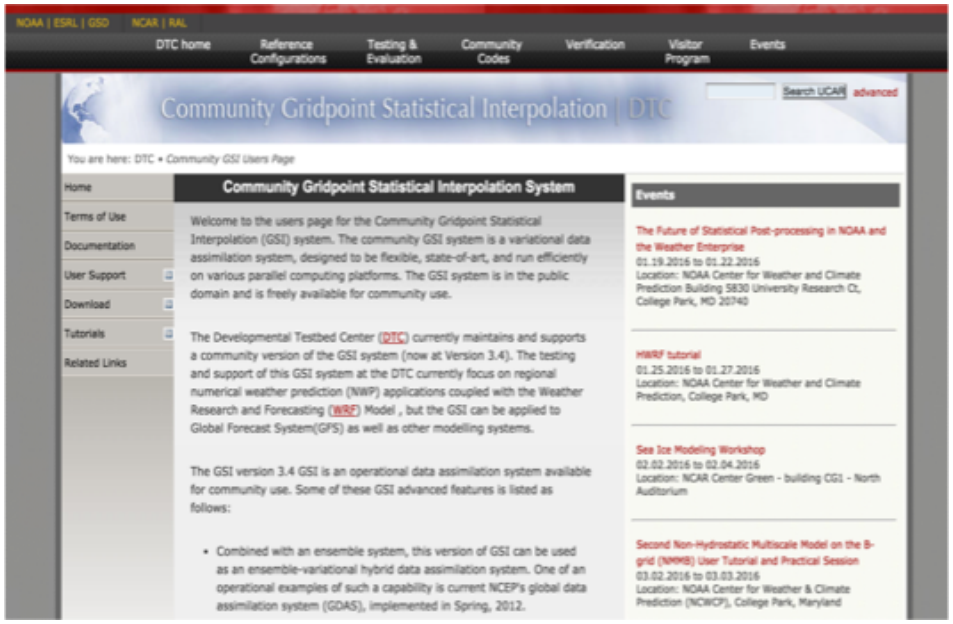
obs

Community GSI

- Managed by DA Review Committee, including members from NCEP/EMC, NASA/GMAO, NOAA/ESRL, NESDIS, AFWA, NCAR/MMM and Developmental Testbed Center (DTC)
- Central community support provided by DTC
  - Resources for general users
    - Webpage

    http://www.dtcenter.org/com-GSI/users/index.php
    - Annual code release
    - Users Guide
    - Annual onsite tutorial
    - Invited onsite tutorial
    - Online tutorial
    - Workshop
    - DTC visitor program:

    http://www.dtcenter.org/visitors/opportunity/



- For developers and advanced users: direct access to code repositories
  - Unified NOAA Vlab repository shared with all other GSI developers, including those from operational centers

- GSI community users: gsi-help@ucar.edu
- HWRF: hwrf-help@ucar.edu

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, Kayo Ide)
  - Background and Observation Errors (Daryl Kleist)
  - GSI Hybrid Data Assimilation (Jeff Whitaker, Daryl Kleist)
  - Aerosol Data Assimilation (Zhiquan Liu)