2014 Hurricane WRF Tutorial, Jan 14-16, 2014, College Park,
MD

Introduction to Data Assimilation and Community Gridpoint Statistical Interpolation System (GSI)

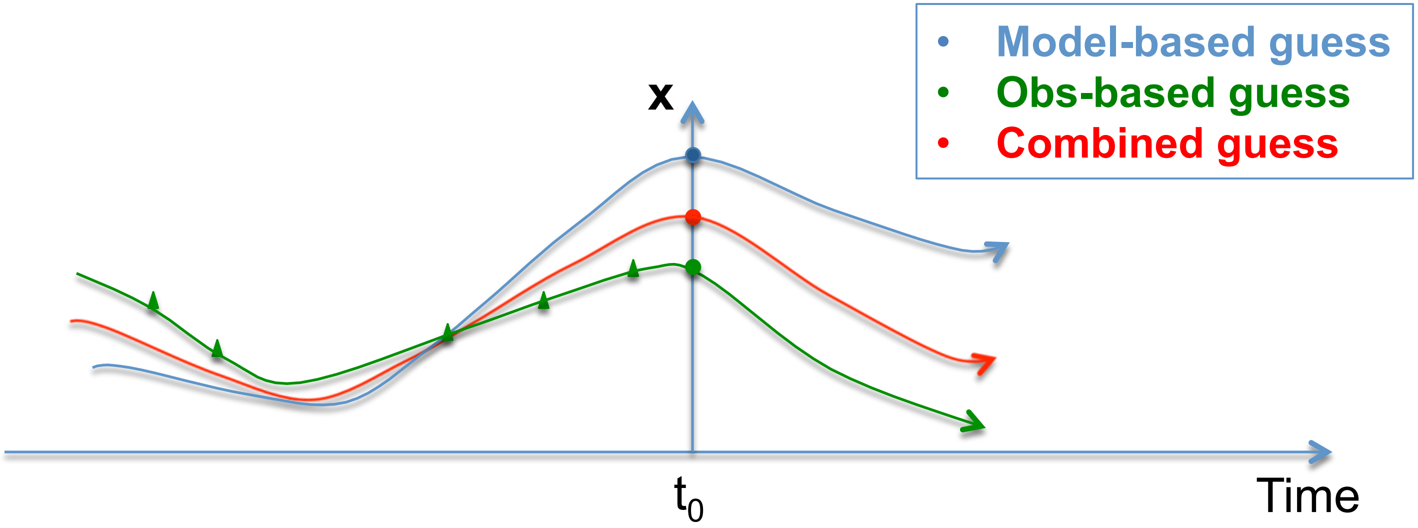
Hui Shao & Ming Hu
Developmental Testbed Center

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)

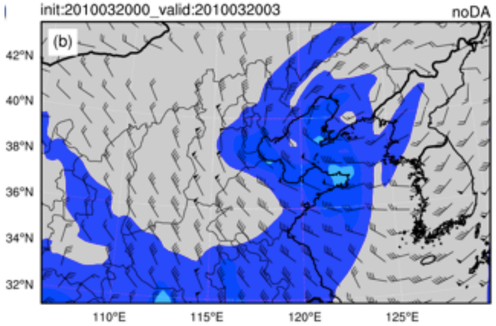
  - “Knowing the current state of the weather is as important as the numerical computer models processing the data.”-NOAA National Climatic Data Center



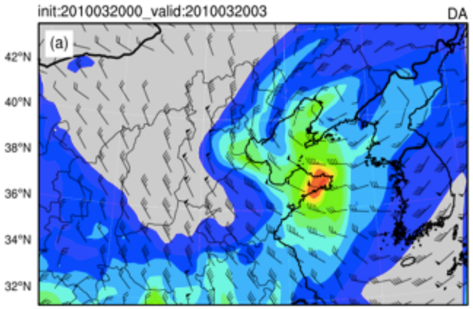
What is Data Assimilation (Cont.)

昨天北方沙尘来到南京，使南京蒙上灰蒙蒙的“沙帐”。张筠 摄

\+



=



- “**Data assimilation** is an analysis technique in which the observed information is accumulated into the model state by taking advantage of consistency constraints with laws of time evolution and physical properties”-F. Bouttier and P. Courtier, ECWMF Data Assimilation Training, 1999
  - Initial conditions for NWP
  - Calibration and validation
  - Observing system design, monitoring and assessment
  - Reanalysis
  - Better understanding (Model errors, Data errors, Physical process interactions, etc)

Hybrid Ensemble-Variational Data Analysis Concepts and Methods

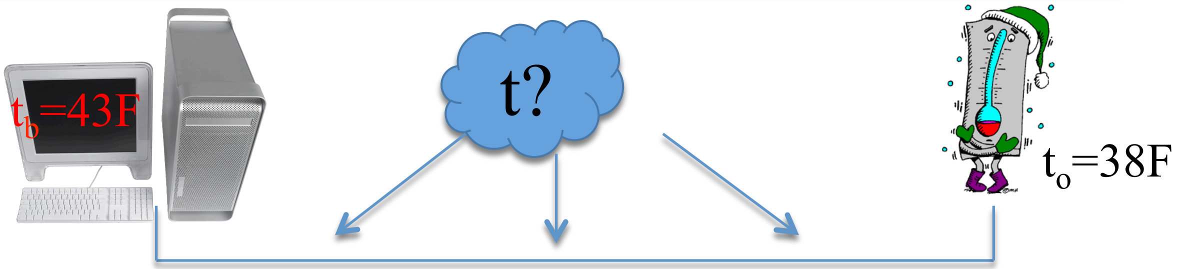
Three Dimensional Variational (3D-Var) Data Assimilation

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

- J: Cost function (penalty) = fit to background (J_b) + fit to observations (J_o)
- **x**: Analysis vector
- $\mathbf{x}_{b:}$ background vector
- $\mathbf{B}_{Var}$**:** Background error (BE) covariance matrix(estimated offline)
- H**:** Observation operator (for 4D-var, H->forward operator HM, where M is a forward model)
- **R:** Observation error covariance
- **y**: Observation vector

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

$$J_{Var}(t) = ½(t-t_b)\sigma_B^{-1}(t-t_b) + ½(t_o-t)\sigma_R^{-1}(t_o-t)$$



3D-Var Data Assimilation (cont.)

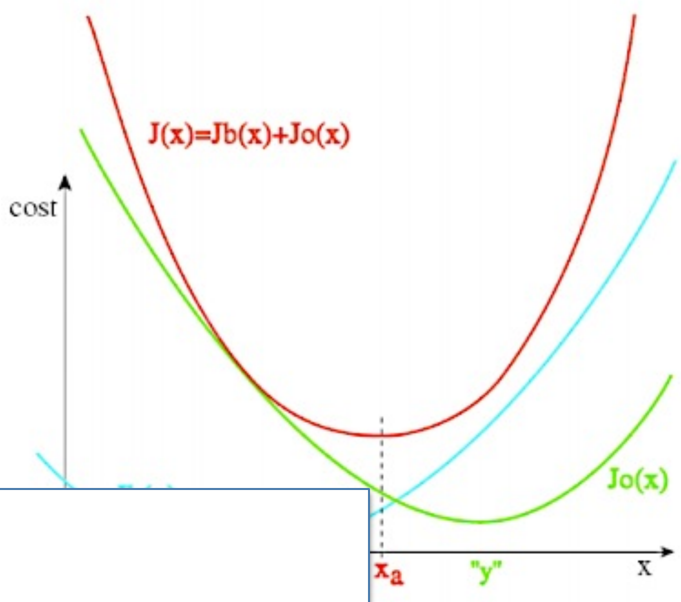
$$J_{\mathrm{Var}}(\mathbf{x})=\frac{1}{2}(\mathbf{x}-\mathbf{x}_b)^{\mathrm{T}}\mathbf{B}_{\mathrm{Var}}^{-1}(\mathbf{x}-\mathbf{x}_b)+\frac{1}{2}(\mathbf{y}-\mathrm{H}\mathbf{x})^{\mathrm{T}}\mathbf{R}^{-1}(\mathbf{y}-\mathrm{H}\mathbf{x})=\mathrm{J}_{\mathrm{b}}+\mathrm{J}_{\mathrm{o}}$$

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

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

$$\nabla J_{\mathrm{Var}}(\mathbf{x})=0$$

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



Scalar example: What is the temperature (t)?

$$\sigma_B^{-1}(t-t_b)-\sigma_R^{-1}(t_o-t)=0$$

$$t=\sigma_B t_0(\sigma_B+\sigma_R)^{-1}+\sigma_R t_b(\sigma_B+\sigma_R)^{-1}$$

Hypotheses assumed

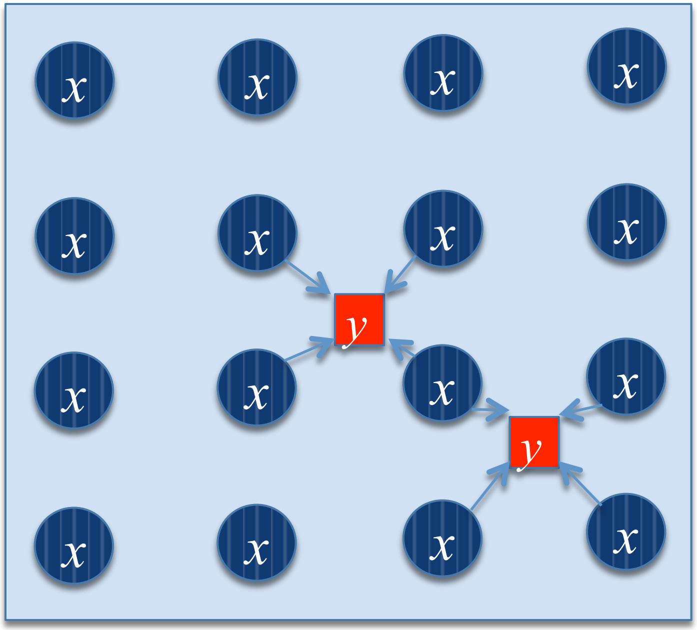
- **Linearized observation operator**: the variations of the observation operator in the vicinity of the background state are linear:
  - for any $\mathbf{x}$ close enough to $\mathbf{x}_b$:

    $H(\mathbf{x}) - H(\mathbf{x}_b) = H(\mathbf{x} - \mathbf{x}_b)$, 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)

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

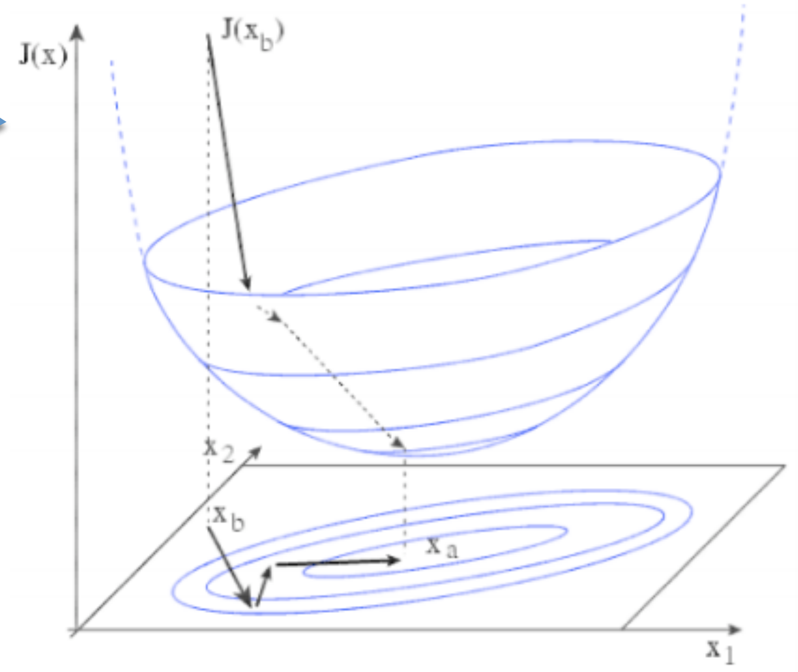
- Observation: **y**
- Observation operator: H
  - Most (traditional measurements)
    - 3D interpolation
  - Some (non-traditional)
    - Complex function, e.g.,
      - Radiance= f(t,q), where f is a radiative transfer model
      - Radar Reflectivity = f(q_r,q_s,q_h)
- Observation innovation: **y**-H**x**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



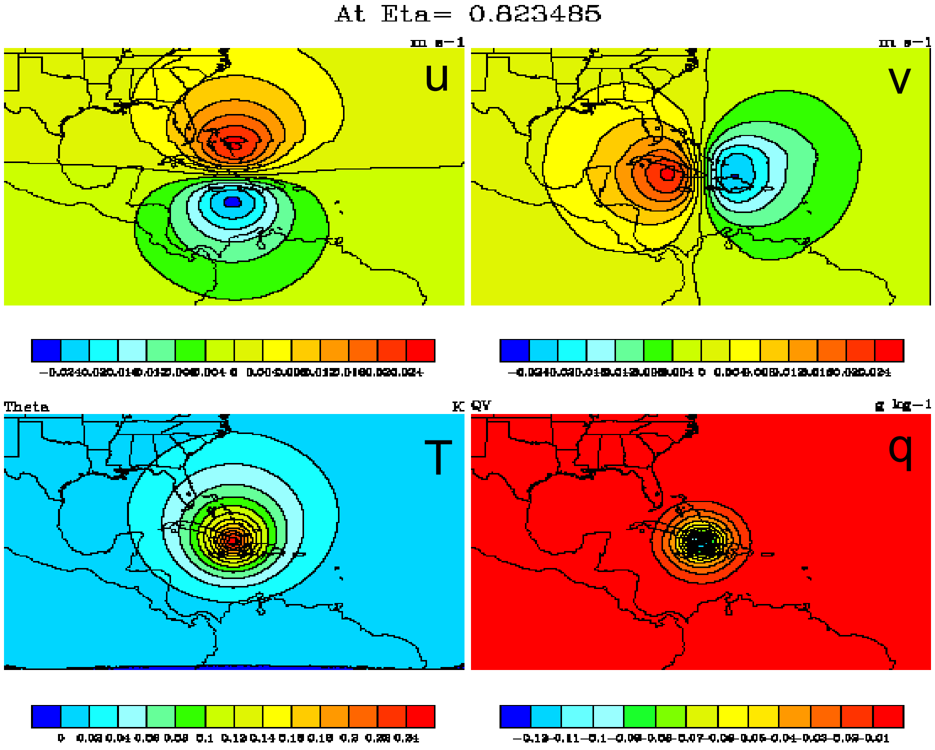
Background Term

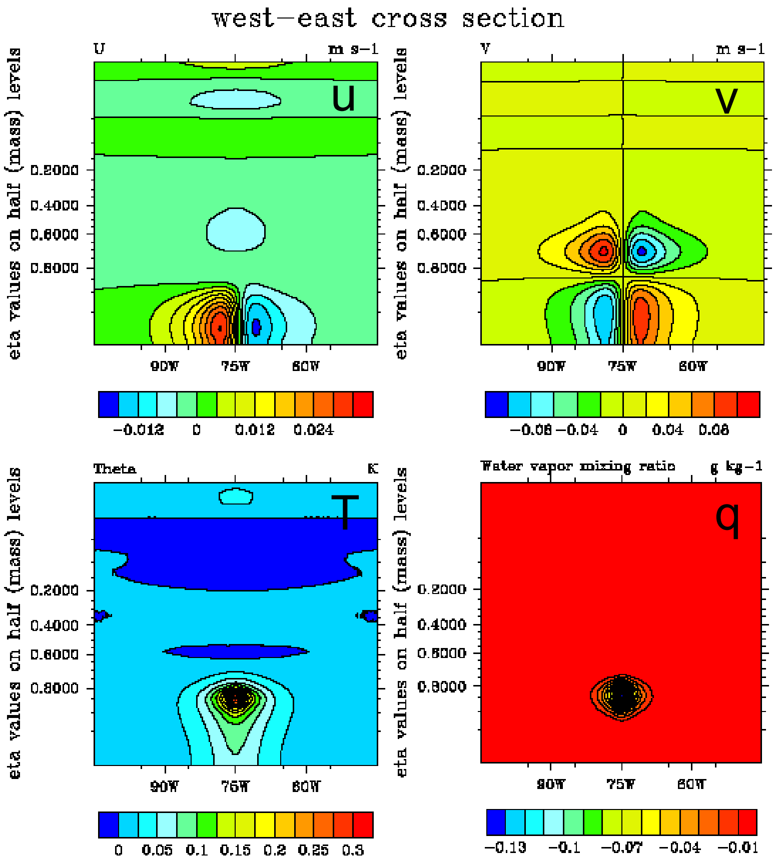
$$J_{\mathrm{Var}}(\mathbf{x})=\boxed{\frac{1}{2}(\mathbf{x}\text{-}\mathbf{x}_b)^{\mathrm{T}}\mathbf{B}_{\mathrm{Var}}^{-1}(\mathbf{x}\text{-}\mathbf{x}_b)}+\frac{1}{2}(\mathbf{y}-\mathrm{H}\mathbf{x})^{\mathrm{T}}\mathbf{R}^{-1}(\mathbf{y}-\mathrm{H}\mathbf{x})=\mathrm{J}_{\mathrm{b}}+\mathrm{J}_{\mathrm{o}}$$

- Background: $\mathbf{x}_b$
- Analysis: $\mathbf{x}$
  - Start from $\mathbf{x}=\mathbf{x}_b$
- Analysis increment: $\mathbf{x}\text{-}\mathbf{x}_b$
- 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.
    - Computed a priori



Pseudo Single-Observation Test





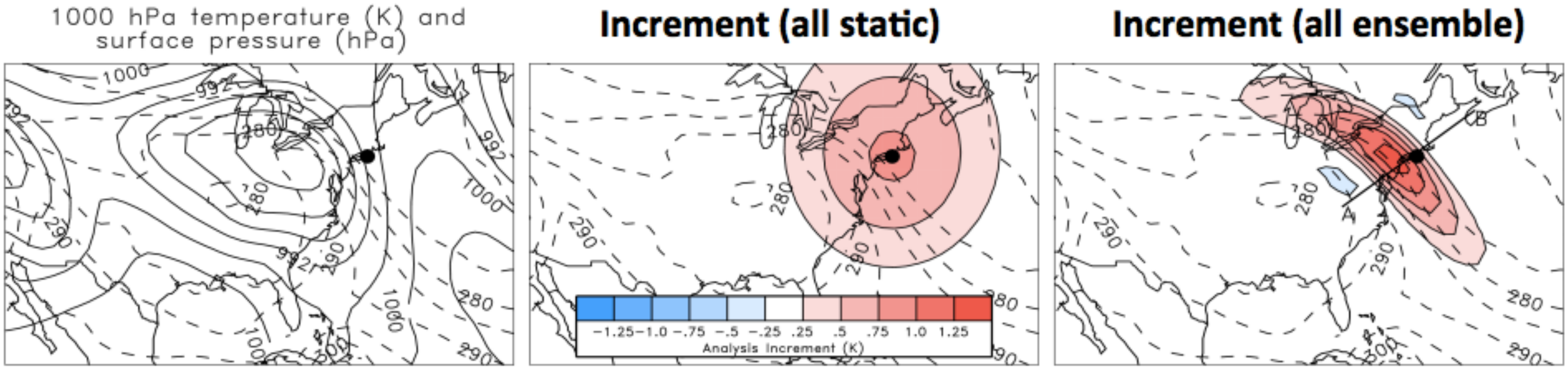
Hybrid Ensemble 3D-Var

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

- **B**$_{\mathrm{Var}}$**:** (Static) background error (BE) covariance matrix(estimated offline)
- **B**$_{\mathrm{Ens}}$**:** (Flow dependent) background error covariance matrix (estimated from ensemble)
- *β*: Weighting factor (0.25 means total **B** is ¾ ensemble)

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

Temperature observation near a warm front



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

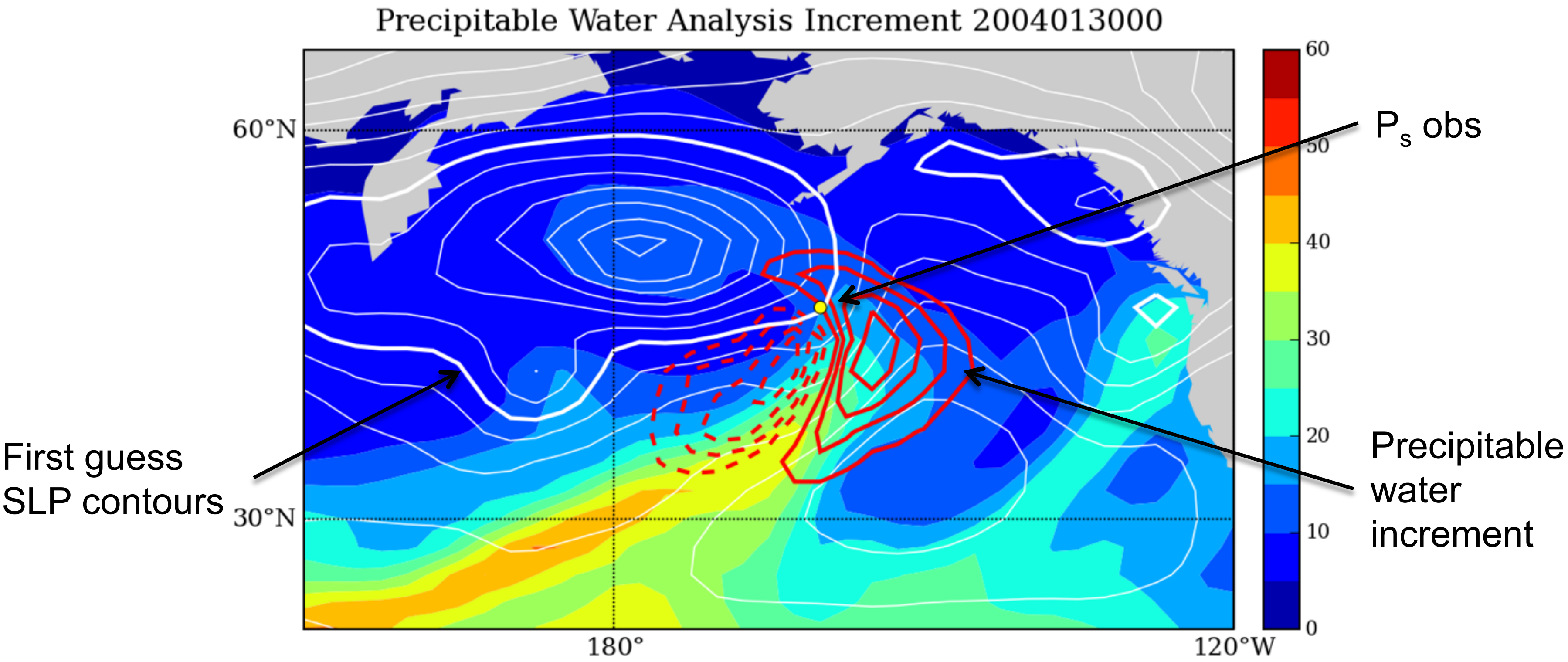
Increment (all static)

Increment (all ensemble)

- ✓ Allows for flow-dependence/errors of the day

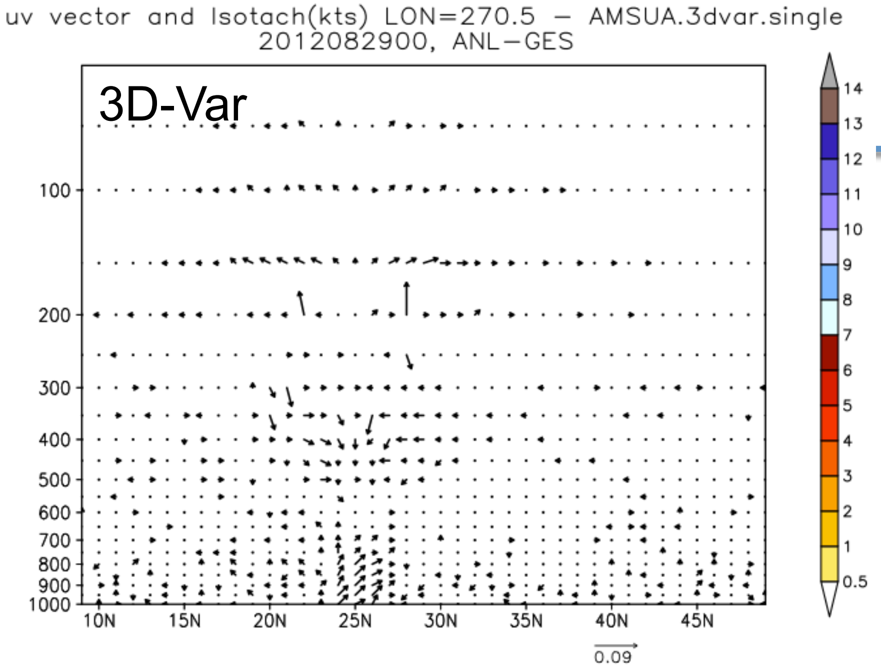
What Does $\mathbf{B}_{ens}$ Do?

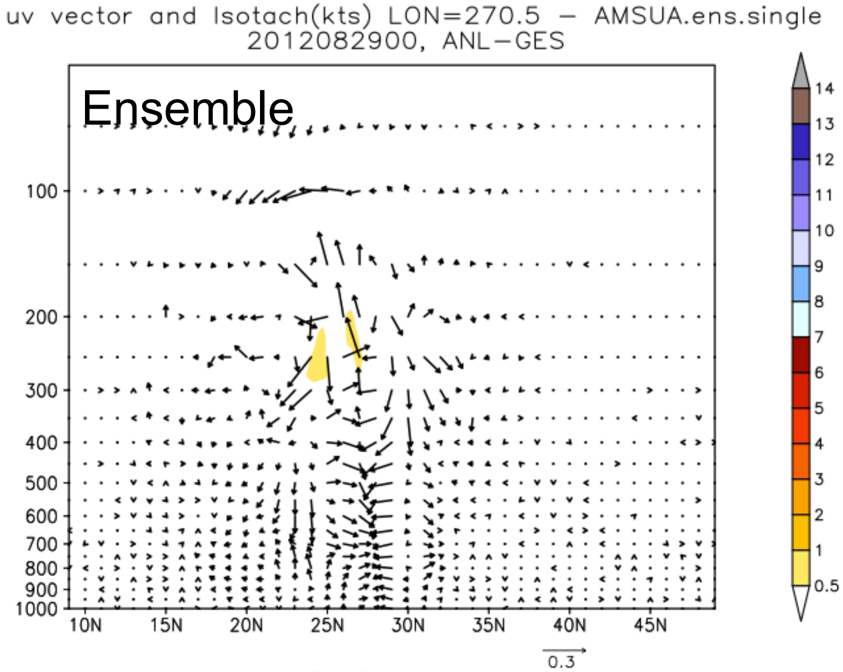
Surface pressure observation near "atmospheric river"

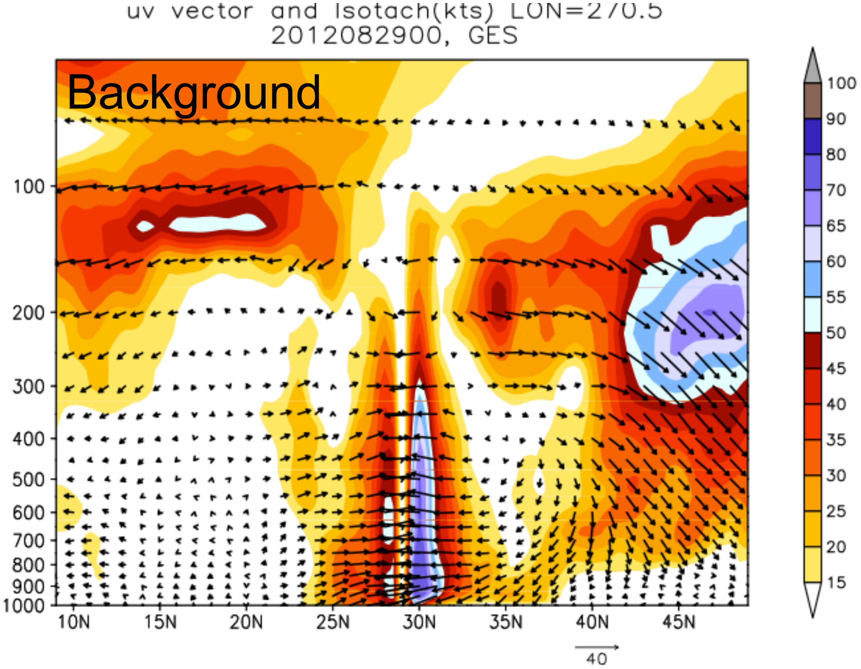


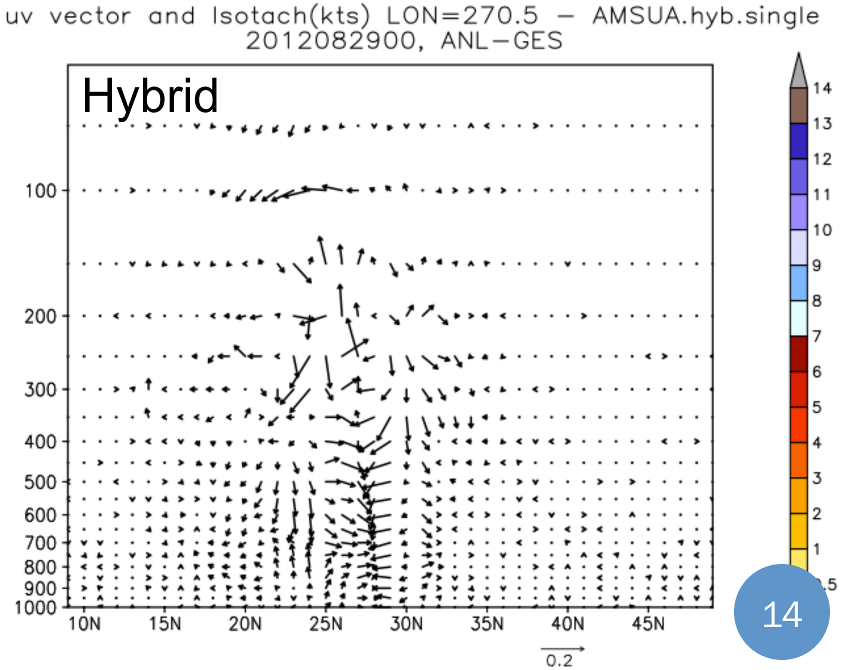
3D-Var increment would be zero

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


uv vector and Isotach(kts) LON=270.5 – AMSUA.3dvar.single
2012082900, ANL–GES


uv vector and Isotach(kts) LON=270.5 – AMSUA.ens.single
2012082900, ANL–GES


uv vector and Isotach(kts) LON=270.5
2012082900, GES

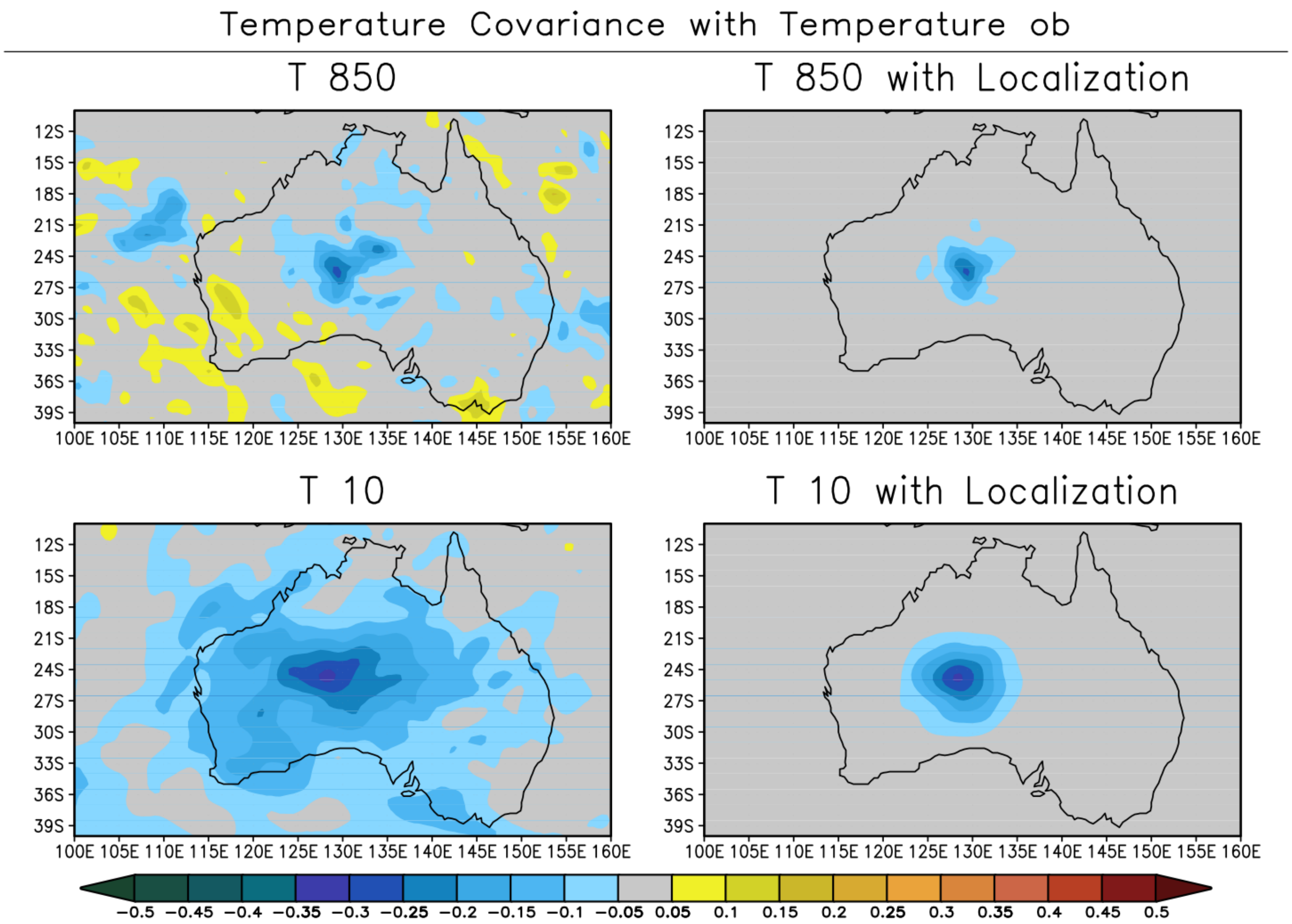

uv vector and Isotach(kts) LON=270.5 – AMSUA.hyb.single
2012082900, ANL–GES

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!

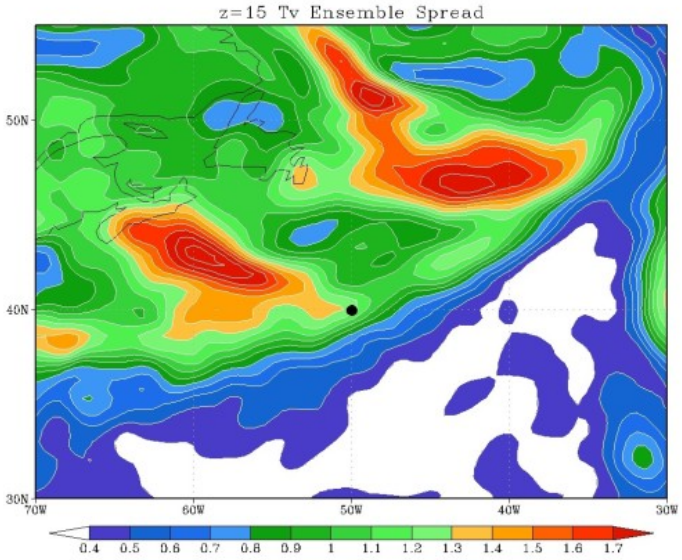
Localization of $\mathbf{B}_{\text{Ens}}$



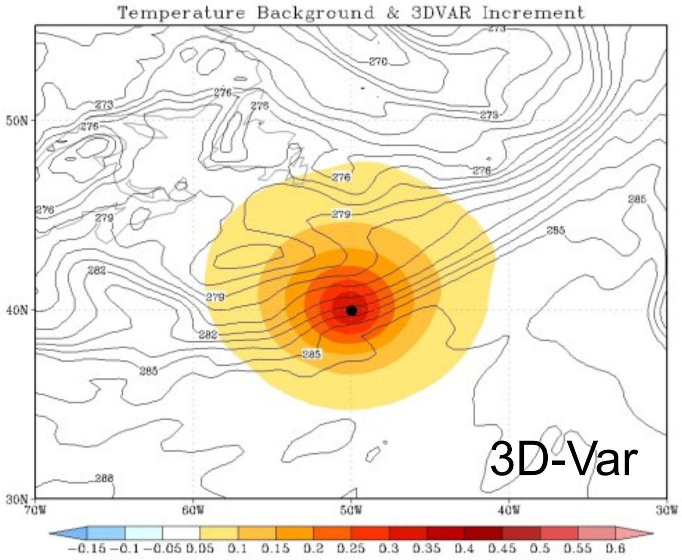
Why Hybrid?

	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

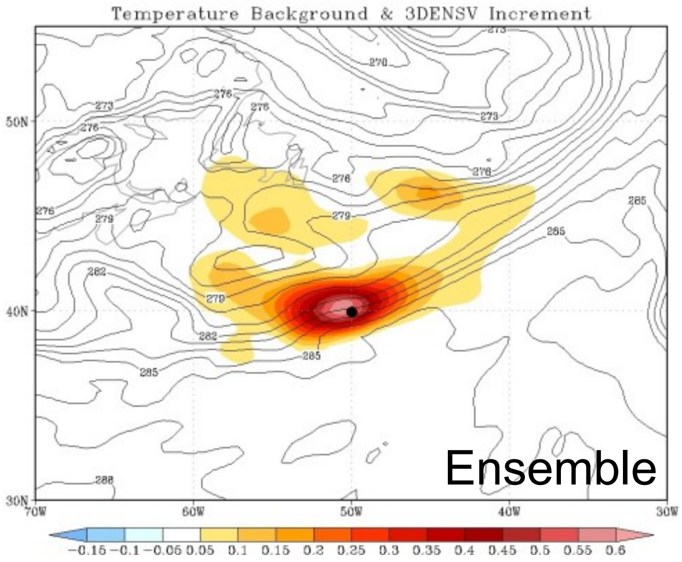


z=15 Tv Ensemble Spread



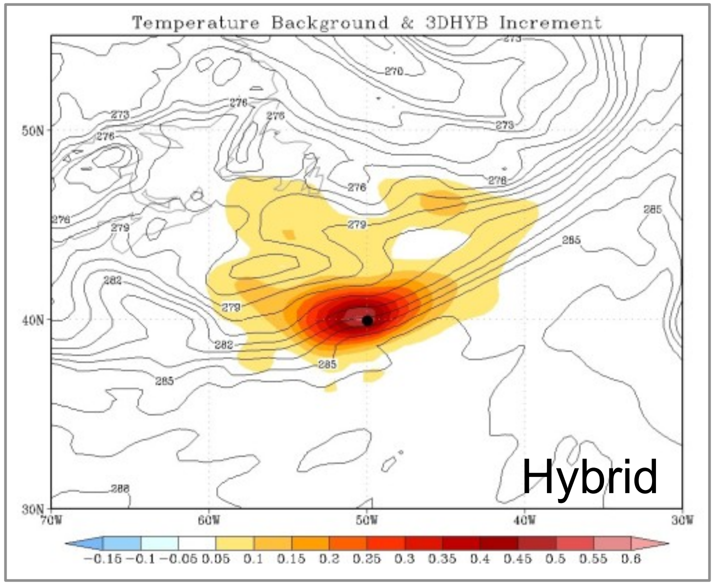
Temperature Background & 3DVAR Increment

3D-Var



Temperature Background & 3DENSV Increment

Ensemble



Temperature Background & 3DHYB Increment

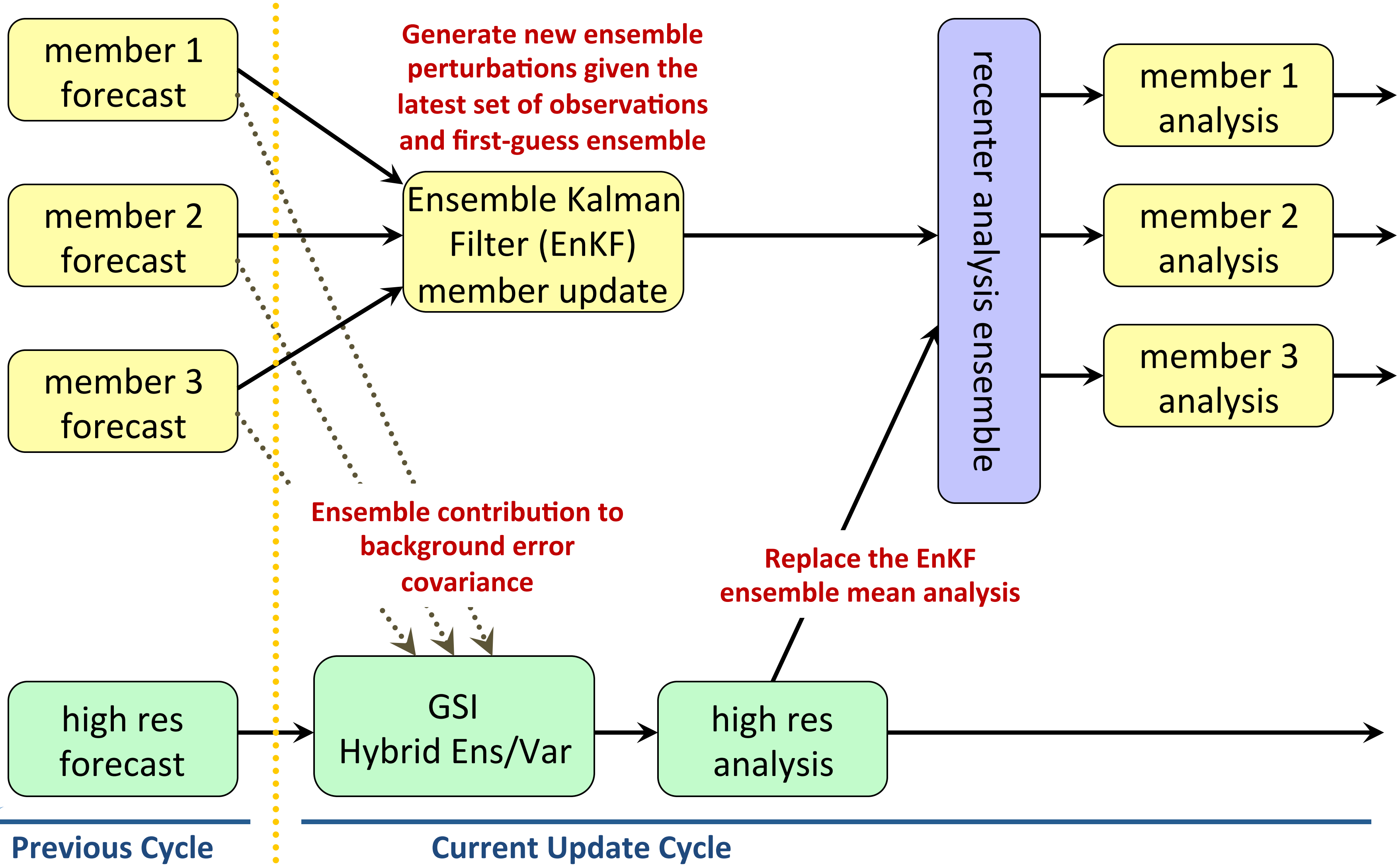
Hybrid

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 $\mathbf{B}_{\mathrm{Var}}$)
  - Larger ensembles have increased computational expense
- Updating the ensemble (NCEP)
  - Global only currently: 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)


member 1 forecast

member 2 forecast

member 3 forecast

Generate new ensemble perturbations given the latest set of observations and first-guess ensemble

Ensemble Kalman Filter (EnKF) member update

recenter analysis ensemble

member 1 analysis

member 2 analysis

member 3 analysis

Ensemble contribution to background error covariance

Replace the EnKF ensemble mean analysis

high res forecast

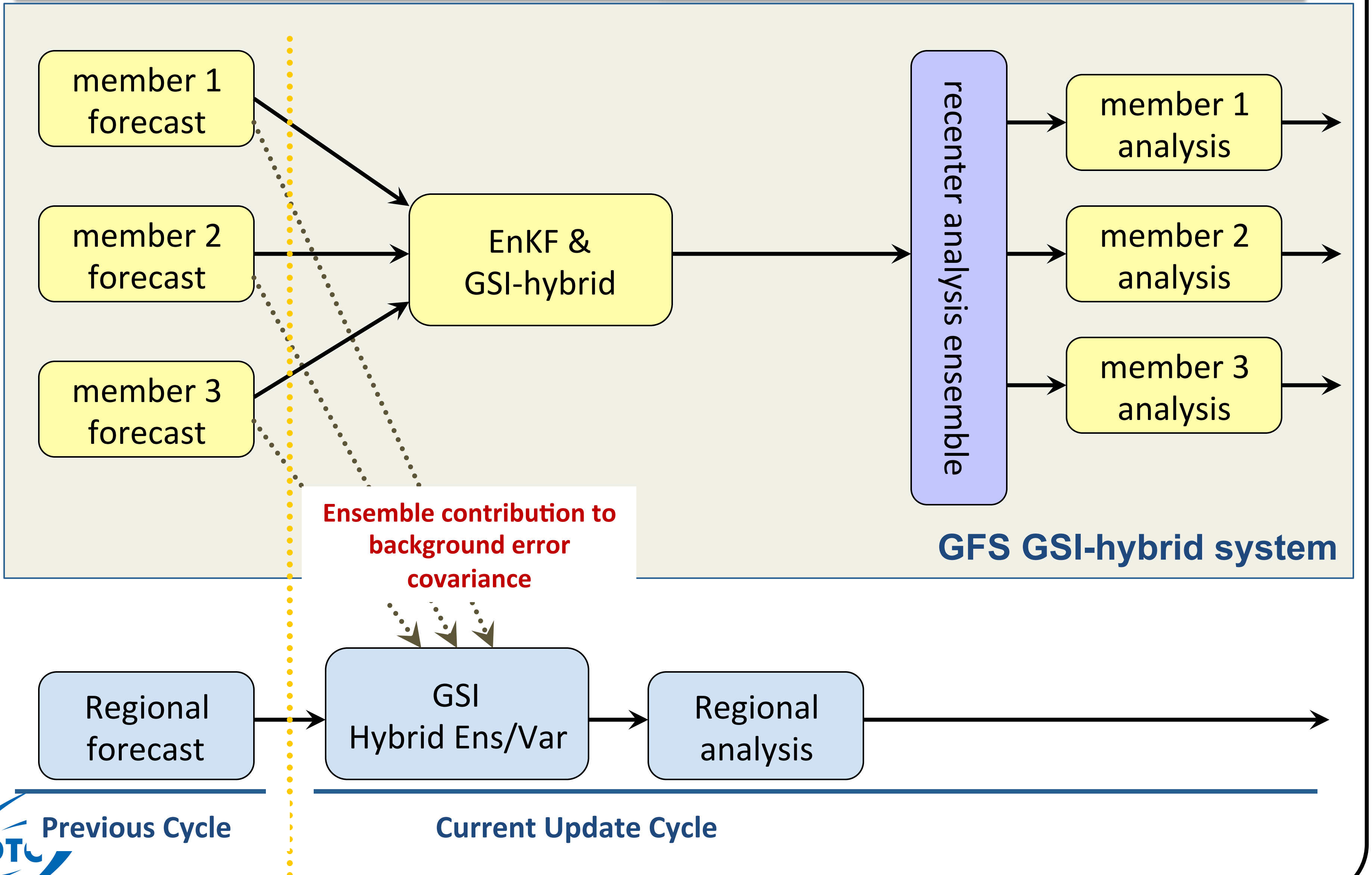
GSI Hybrid Ens/Var

high res analysis

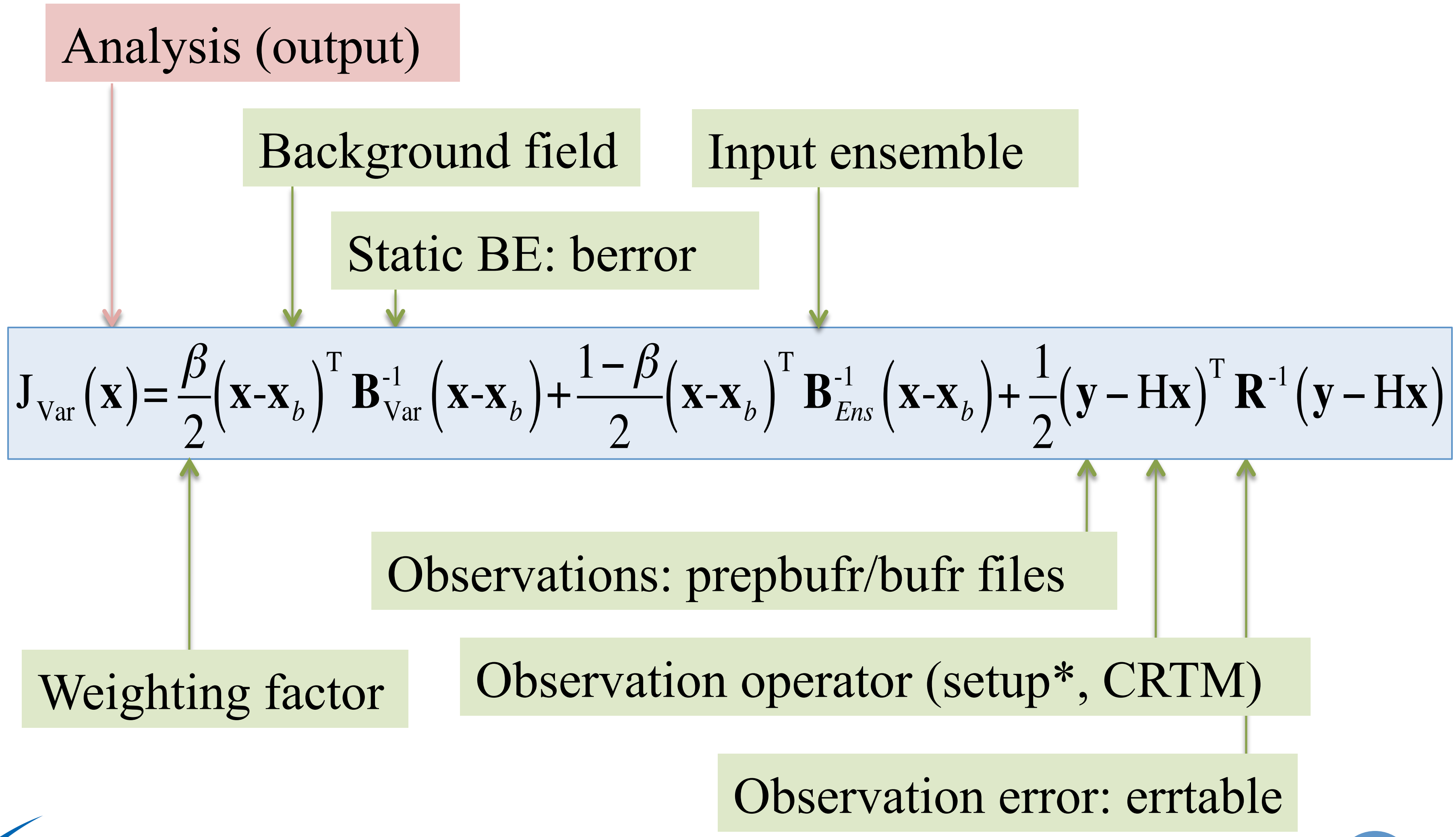
Previous Cycle

Current Update Cycle

Current Scheme for Regional GSI-hybrid



Summary with GSI



Analysis (output)

Background field

Input ensemble

Static BE: berror

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

Observations: prepbufr/bufr files

Weighting factor

Observation operator (setup*, CRTM)

Observation error: errtable

GSI and Its Community Support

History

- Optimal Interpolation (OI) Analysis system
  - First statistic data analysis system
- The Spectral Statistical Interpolation (SSI) analysis system was developed at NCEP in the late 1980's and early 1990's
  - First operational variational analysis system
  - Directly assimilate radiances
- The Gridpoint Statistical Interpolation (GSI) analysis system was developed as the next generation global/regional analysis system
  - Wan-Shu Wu, R. James Purser, David Parrish, 2002:Three-Dimensional Variational Analysis with spatially Inhomogeneous Covariances. *Mon. Wea. Rev., 130, 2905-2916.*
  - Based on SSI analysis system
  - Replace spectral definition for background errors with grid point version based on recursive filters
  - First implemented at NCEP in 2006

Current

- GSI used in NCEP operations for
  - Regional
  - Global
  - Hurricane
  - Real-Time Mesoscale Analysis
  - Rapid Refresh (ESRL/GSD)
- GMAO collaboration (NASA 4DVAR)
- Operational at AFWA
- Modification to fit into WRF and NCEP infrastructure
- Evolution to Earth System Modeling Framework (ESMF)

General Comments

- GSI analysis code is an evolving system.
  - Scientific advances (hybrid, Ens-Var, new data & analysis variables)
  - Improved coding
    - More efficient algorithms
    - Bundle structure
    - Generalizations of code (portability, multiple-models, …)
    - Improved documentation
- Code is intended to be used operationally as well as by the research community
  - Coding requirements, ops. Infrastructure, efficiency, ease of usage
  - External testing, user-friendly interface, transition to ops., distributed effort (potential duplication)

✓ HWRF is one of the GSI applications. Some other options/capabilities available or under development might not be introduced here.

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

Community GSI – User Support

- General GSI support through User's Page and help desk:

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

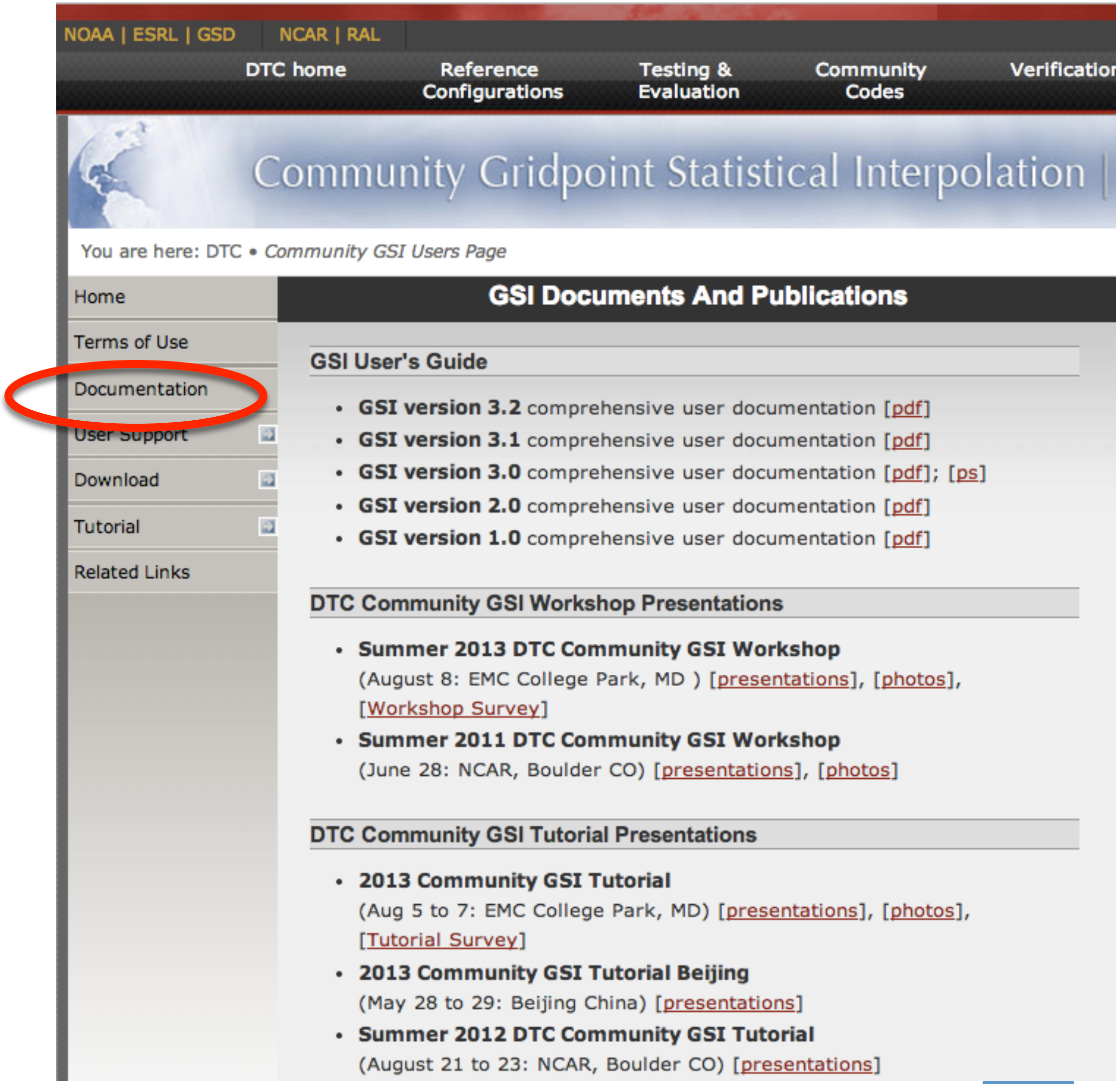
- Annual code release (general)
  - Tested with PGI, Intel, and GNU (gfortran) compilers
  - Latest version is V3.2 (July, 2013)
    - 2013 operational HWRF capabilities
  - Next release is planned for Spring 2014

GSI for HWRF: wrfhelp@ucar.edu

- Bundled with HWRF release
- Consistent with general GSI code release
- Next release is planned for Summer 2014

Community GSI - Documents

- User's Guide
  - Match each official release
- Workshop presentations
- Tutorial lectures
- Technique documentations
- Code browser
  - Calling tree
- Key publications



You are here: DTC • *Community GSI Users Page*

GSI Documents And Publications

GSI User's Guide

- **GSI version 3.2** comprehensive user documentation [pdf]
- **GSI version 3.1** comprehensive user documentation [pdf]
- **GSI version 3.0** comprehensive user documentation [pdf]; [ps]
- **GSI version 2.0** comprehensive user documentation [pdf]
- **GSI version 1.0** comprehensive user documentation [pdf]

DTC Community GSI Workshop Presentations

- **Summer 2013 DTC Community GSI Workshop** (August 8: EMC College Park, MD) [presentations], [photos], [Workshop Survey]
- **Summer 2011 DTC Community GSI Workshop** (June 28: NCAR, Boulder CO) [presentations], [photos]

DTC Community GSI Tutorial Presentations

- **2013 Community GSI Tutorial** (Aug 5 to 7: EMC College Park, MD) [presentations], [photos], [Tutorial Survey]
- **2013 Community GSI Tutorial Beijing** (May 28 to 29: Beijing China) [presentations]
- **Summer 2012 DTC Community GSI Tutorial** (August 21 to 23: NCAR, Boulder CO) [presentations]

Community GSI - Practice

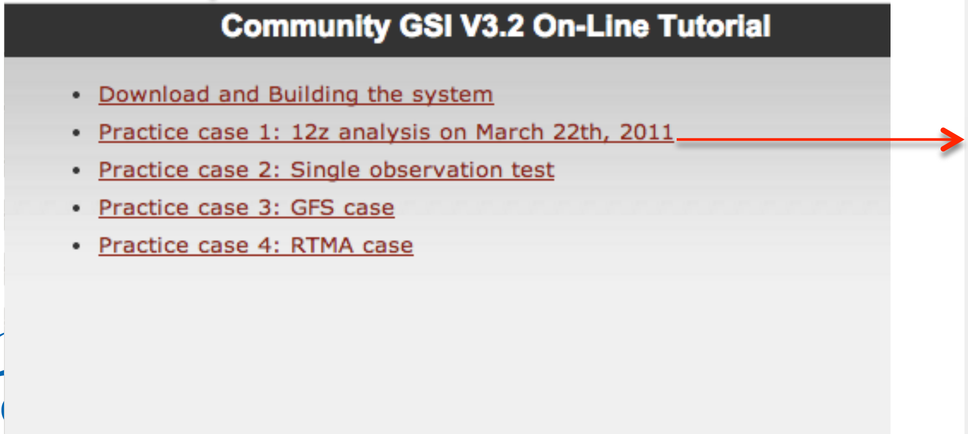
- On-line tutorial for each release
- Residential tutorial practice cases

GSI Tutorials

On-line Tutorial

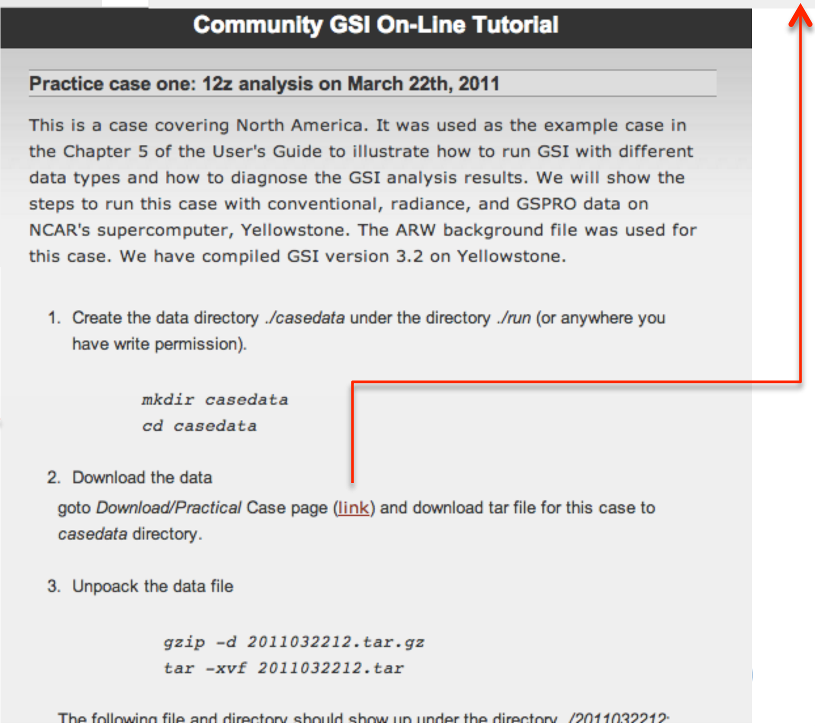
The GSI On-line Tutorial provides the basic steps to compile the GSI system, and simple cases for testing GSI code to run.

- On-line tutorial for community GSI Version 3.2
- On-line tutorial for community GSI Version 3.1
- On-line tutorial for community GSI Version 3.0
- On-line tutorial for community GSI Version 2.0
- On-line tutorial for community GSI Version 1.0



Community GSI V3.2 On-Line Tutorial

- Download and Building the system
- Practice case 1: 12z analysis on March 22th, 2011
- Practice case 2: Single observation test
- Practice case 3: GFS case
- Practice case 4: RTMA case



Community GSI On-Line Tutorial

Practice case one: 12z analysis on March 22th, 2011

This is a case covering North America. It was used as the example case in the Chapter 5 of the User's Guide to illustrate how to run GSI with different data types and how to diagnose the GSI analysis results. We will show the steps to run this case with conventional, radiance, and GSPRO data on NCAR's supercomputer, Yellowstone. The ARW background file was used for this case. We have compiled GSI version 3.2 on Yellowstone.

1. Create the data directory *./casedata* under the directory *./run* (or anywhere you have write permission).

```
mkdir casedata
cd casedata
```

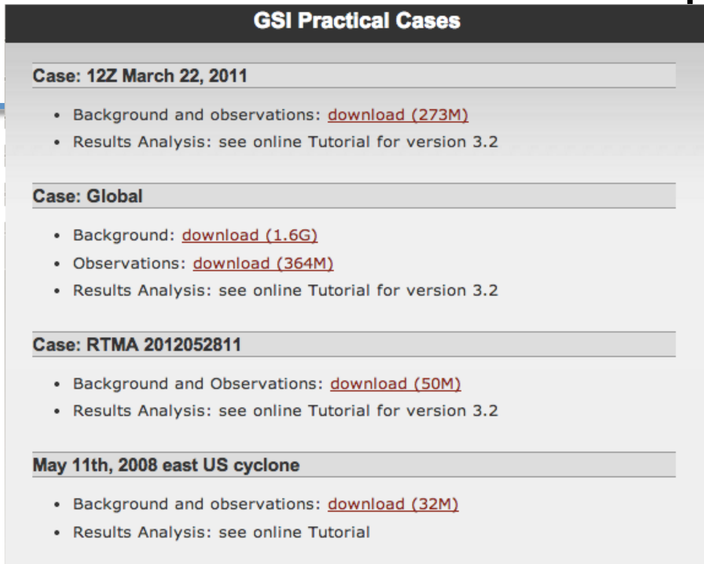
2. Download the data

goto *Download/Practical* Case page (link) and download tar file for this case to *casedata* directory.

3. Unpoack the data file

```
gzip -d 2011032212.tar.gz
tar -xvf 2011032212.tar
```

The following file and directory should show up under the directory ./2011032212:



GSI Practical Cases

Case: 12Z March 22, 2011

- Background and observations: download (273M)
- Results Analysis: see online Tutorial for version 3.2

Case: Global

- Background: download (1.6G)
- Observations: download (364M)
- Results Analysis: see online Tutorial for version 3.2

Case: RTMA 2012052811

- Background and Observations: download (50M)
- Results Analysis: see online Tutorial for version 3.2

May 11th, 2008 east US cyclone

- Background and observations: download (32M)
- Results Analysis: see online Tutorial

Reference

- Data Assimilation Concept and Methods (ECMWF Training Course, Bouttier & Courtier)
- GSI Tutorial Lectures:
  - 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)