Data Assimilation: Finding the Initial Conditions in Large Dynamical Systems

Eric Kostelich

Data Mining Seminar, Feb. 6, 2006

kostelich@asu.edu

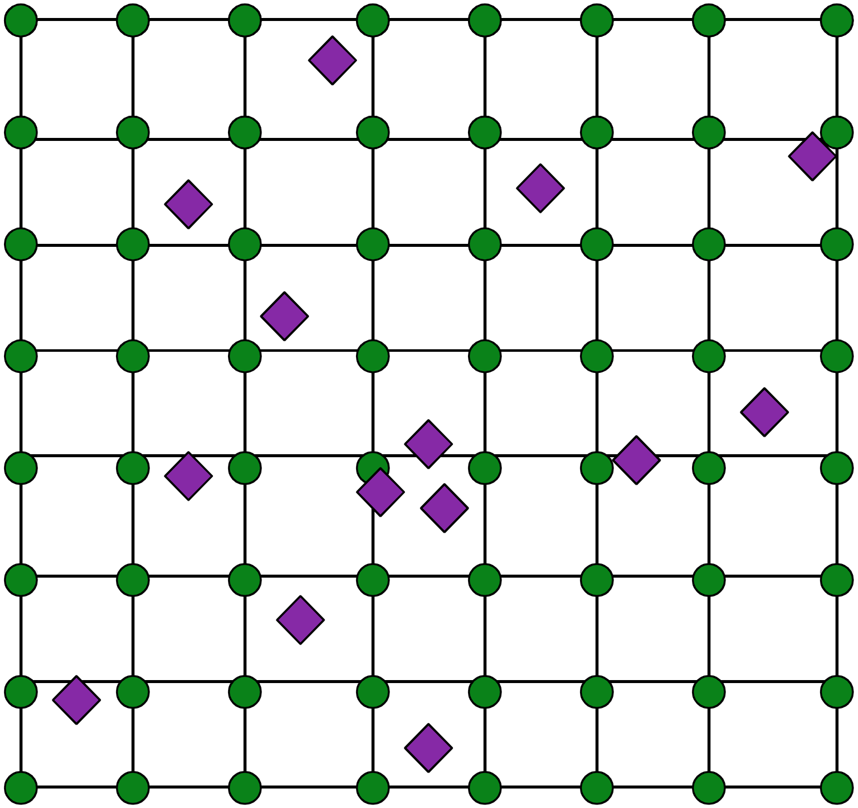
Co-Workers

Istvan Szunyogh, Gyorgyi Gyarmati, Ed Ott, Brian Hunt,
Eugenia Kalnay, D. J. Patil, Jim Yorke

Generous support from:
National Science Foundation, Army Research Office,
NASA, W. M. Keck Foundation, J. McDonnell Foundation,
IBM Corp., ASU

Preprints: http://keck2.umd.edu/weather/

The data assimilation problem



- Forecast model (PDE) predicts values of dynamical variables on a discretized grid (background)
- Observations are noisy and sparse
- What is the “true” current state?

The “data mining” challenge

- Data assimilation is currently the most expensive part of numerical weather prediction
- Current weather models have $\sim 10^7$ dynamical variables and $\sim 10^9$ in the future
- Current observing networks produce $\sim 10^5$ to $\sim 10^6$ measurements every 6 hr
- New satellite observing platforms will generate $\sim 10^7$ measurements every 6 hr

The mathematical challenge

- The dynamical variables in a spatio-temporal model can’t all be observed
- Probably the biggest impediment to better weather forecasts at the moment
- Can be forward in time (weather prediction) or backward in time (climate modeling)
- Methods must be fast to be practical
- Many potential applications: blood flow, cardiac and immune system dynamics

Why is weather so hard to predict?

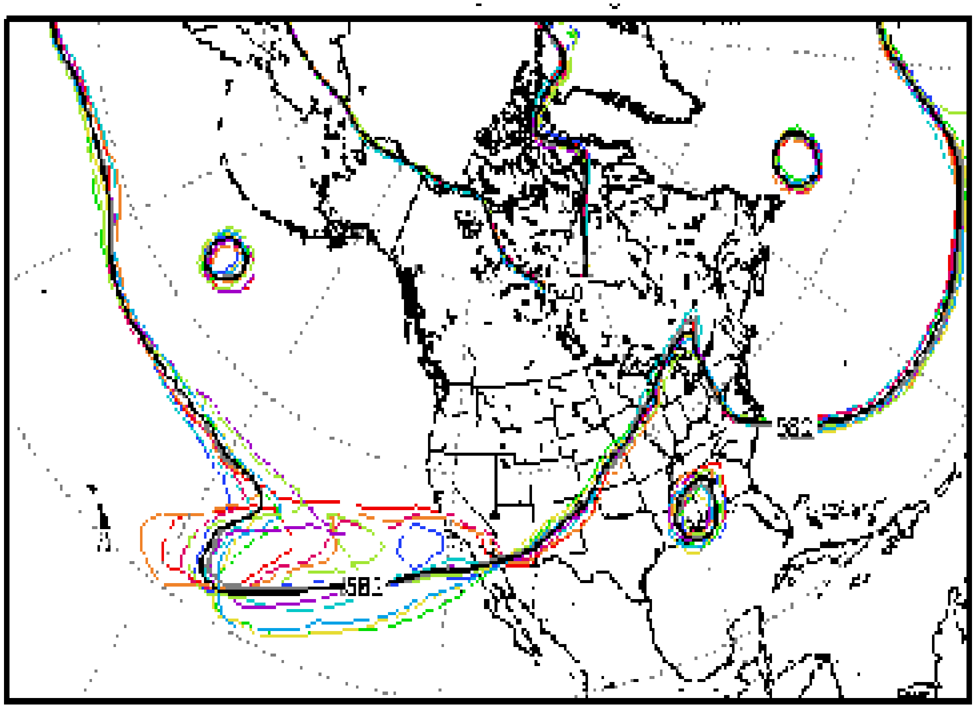
- Dynamics occur at multiple scales
- Dynamics are chaotic (“butterfly effect”)
- Global forecast uncertainty roughly doubles every 24-36 hours
- Uncertainty varies in space and time (“errors of the day”)

Ensemble forecasting

- Simple (but effective) way to assess the uncertainty in a weather forecast
- Basic idea: run many forecasts from statistically equivalent estimates of the current atmospheric state vector
- Assess covariance as function of space and forecast time

“Spaghetti plot”

- Contours reflect uncertainties in atmospheric pressure in this 72-hour forecast



The NCEP Global Forecast System

Spectral model: 3-d Navier-Stokes, plus:

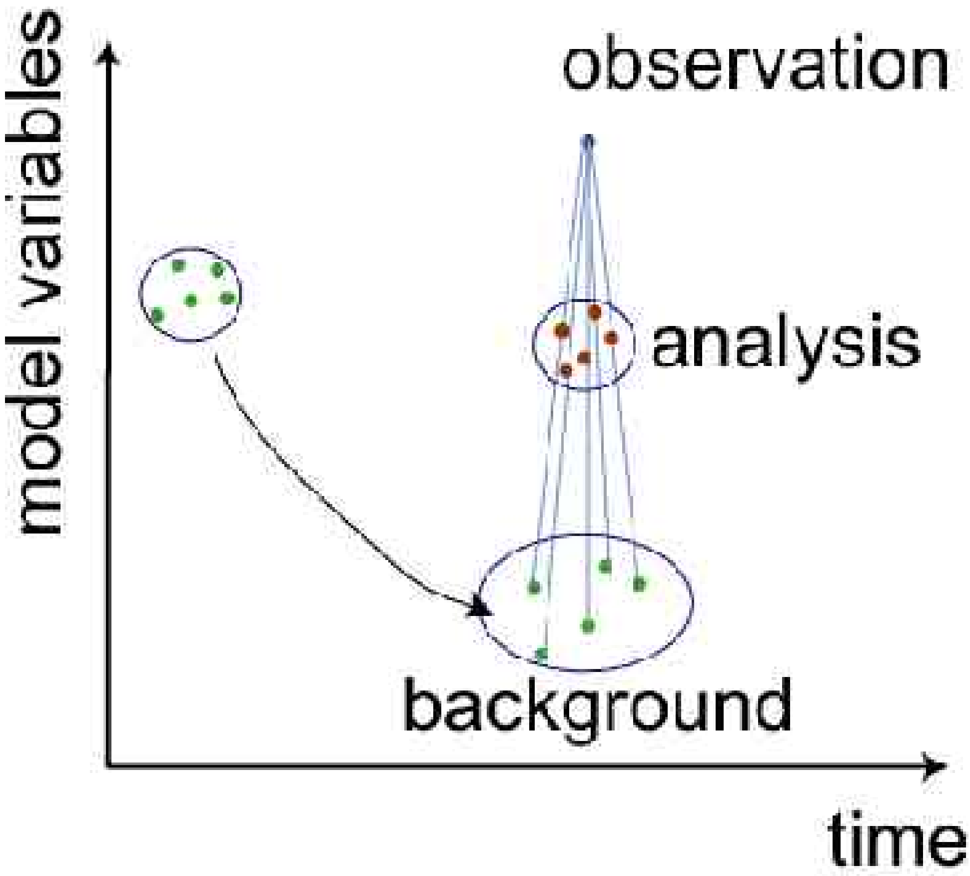
- Atmospheric chemistry (ozone, aerosols)
- Cloud physics (active research area)
- Complex boundary conditions (sea surface, mountains, plants, soils, etc.)

- Principal dynamical variables:
  - Surface pressure
  - Virtual temperature
  - Vorticity and divergence of the wind field

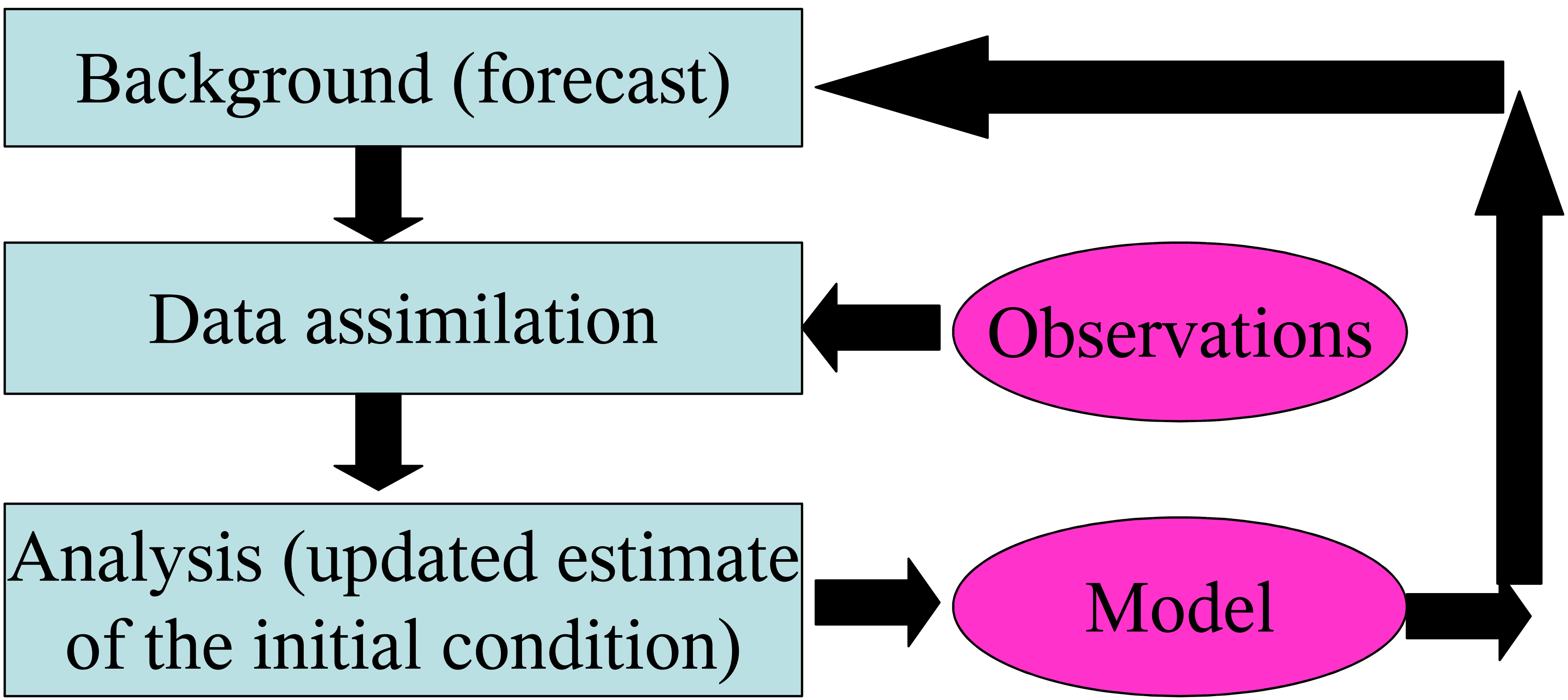
Data assimilation: Basic approach

- Treat the observations and initial condition as random variables
- Statistically interpolate between the model grid and observations to make "best guess" of the true initial condition
- Estimate the uncertainty in the guess
- Need *a priori* estimates of the uncertainties in both the measurements and the background (forecast)

Sequential assimilation



Basic algorithm



The estimation problem

observations: $\mathbf{y} \in \mathbf{R}^{p}, \mathbf{y} = \mathbf{H}\mathbf{x}_{\mathbf{t}} + \boldsymbol{\varepsilon}$

observation errors: $\mathbf{E}(\boldsymbol{\varepsilon}) = \mathbf{0}, \mathbf{E}(\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}^{\mathbf{T}}) = \boldsymbol{\Sigma}$

model variables: $\mathbf{x} \in \mathbf{R}^{n}, \mathbf{x}_{\mathbf{b}} = \mathbf{x}_{\mathbf{t}} + \boldsymbol{\eta}.$

$$\mathbf{E}(\boldsymbol{\eta}) = \mathbf{0}, \mathbf{E}(\boldsymbol{\eta}\boldsymbol{\eta}^{\mathbf{T}}) = \mathbf{P}_{\mathbf{b}}$$

minimize the objective function:

$$\mathbf{J}(\mathbf{x}) = (\mathbf{H}\mathbf{x} - \mathbf{y})^{\mathbf{T}} \boldsymbol{\Sigma}^{-\mathbf{1}} (\mathbf{H}\mathbf{x} - \mathbf{y})$$

$$+ (\mathbf{x} - \mathbf{x}_{\mathbf{b}})^{\mathbf{T}} \mathbf{P}_{\mathbf{b}}^{-\mathbf{1}} (\mathbf{x} - \mathbf{x}_{\mathbf{b}})$$

The estimation problem

- When the errors are Gaussian and the underlying dynamics are linear, the minimizer of $\mathbf{J}$ is “optimal” (unbiased, minimum variance)
- The forecast uncertainty $\mathbf{P_b}$ can be estimated using ensemble forecasts
- Weather service uses seasonally averaged $\mathbf{P_b}$ (ignores errors of the day)

The dimensionality problem

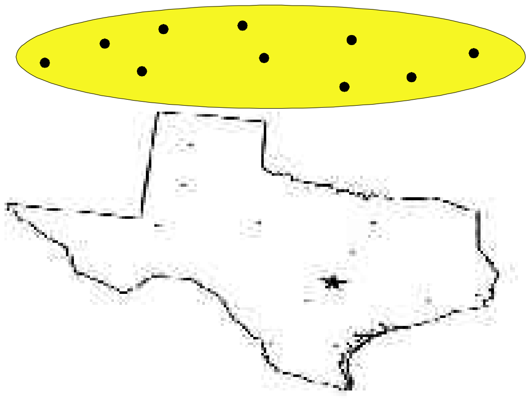
- To evaluate J, we must invert Σ and P_b .

  Σ is $p \times p$ and P_b is $n \times n$.
- For typical weather models, $n \sim 10^7$ to 10^9 and $p \sim 10^5$ to 10^7!
- The computational complexity of matrix inversion is $O(n^3)$.
- Inverting a 100×100 matrix takes ~1 sec.
- A $10^7 \times 10^7$ matrix takes $\sim 10^{15}$ sec!

Maryland/ASU idea: use chaos to reduce the dimensionality

- A medium-resolution weather model has ~3000 variables in a typical 1000×1000 km synoptic region (~Texas)
- Find the dimension of the subspace spanned by a typical ensemble of 100-200 forecast vectors over a Texas-sized region
- The forecast uncertainty evolves along a ~40 dimensional “unstable manifold” (Patil et al., 2001)

The local ensemble idea



- Take ensemble of 100-200 forecast vectors over Texas-sized patch
- Each forecast vector is ~3000 dimensional
- Their span is typically ~40 dimensional for 6-24 hr forecasts

Important implications

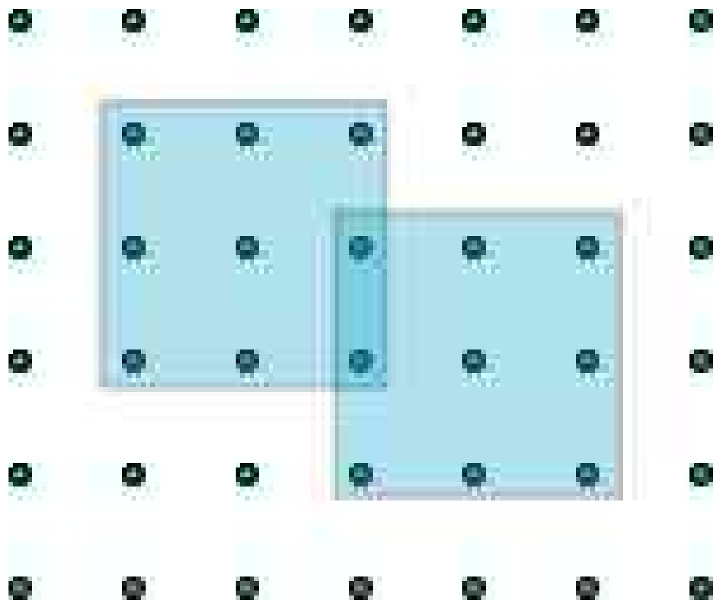
- The “weather attractor” is locally low-dimensional over typical synoptic regions
- The spread in the forecast ensemble is in the direction of most rapidly increasing uncertainty
- A data assimilation algorithm need only reduce the uncertainty in this low-dimensional subspace in any given synoptic region
- The relevant covariance matrix is only 40×40 and can be determined by ensemble forecasts
- Leads to an embarrassingly parallel algorithm

The local ensemble transform Kalman filter (LETKF)

- Perform the data assimilation step independently in each local region
- The grid point in the center of each patch has the most accurate analysis
- Assemble the center-point local analyses into a global grid, then advance to the next forecast time

Computational implementation

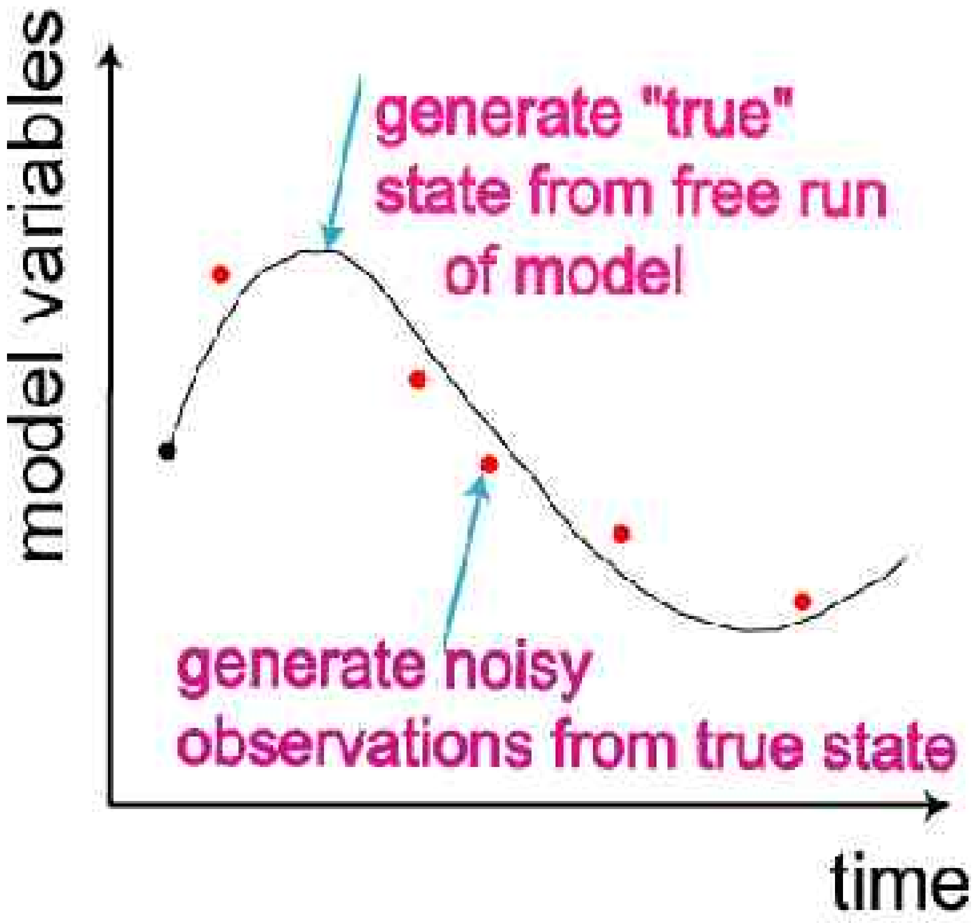
- Patches centered at each point of horizontal grid
- Update the initial condition at center of each patch



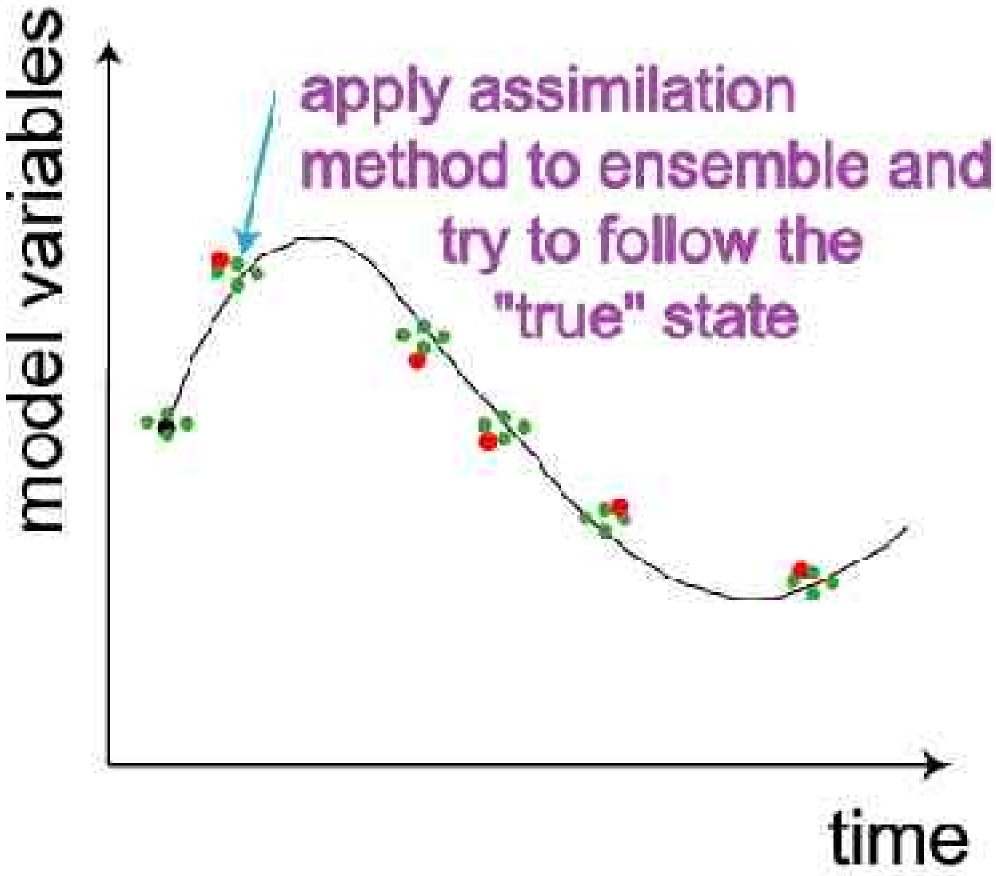
Fast, parallel implementation

- Only operations on ~40×40 matrices are needed in the analyses
- Assimilation of 500,000 observations into 3-million variable model takes 10 min on 20-cpu Beowulf cluster
- Model independent approach: the same algorithm has been applied to three different weather models (NCEP GFS, NASA fvGCM, regional NAM)

“Perfect model” scenario



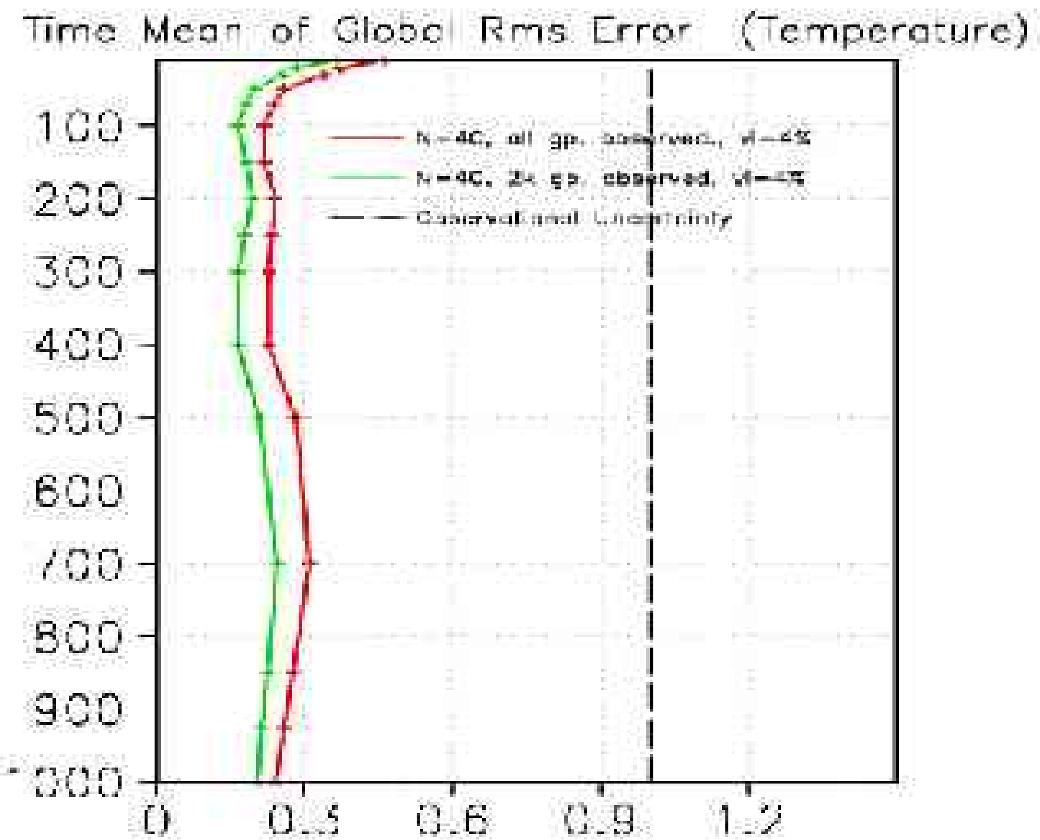
Evaluation method



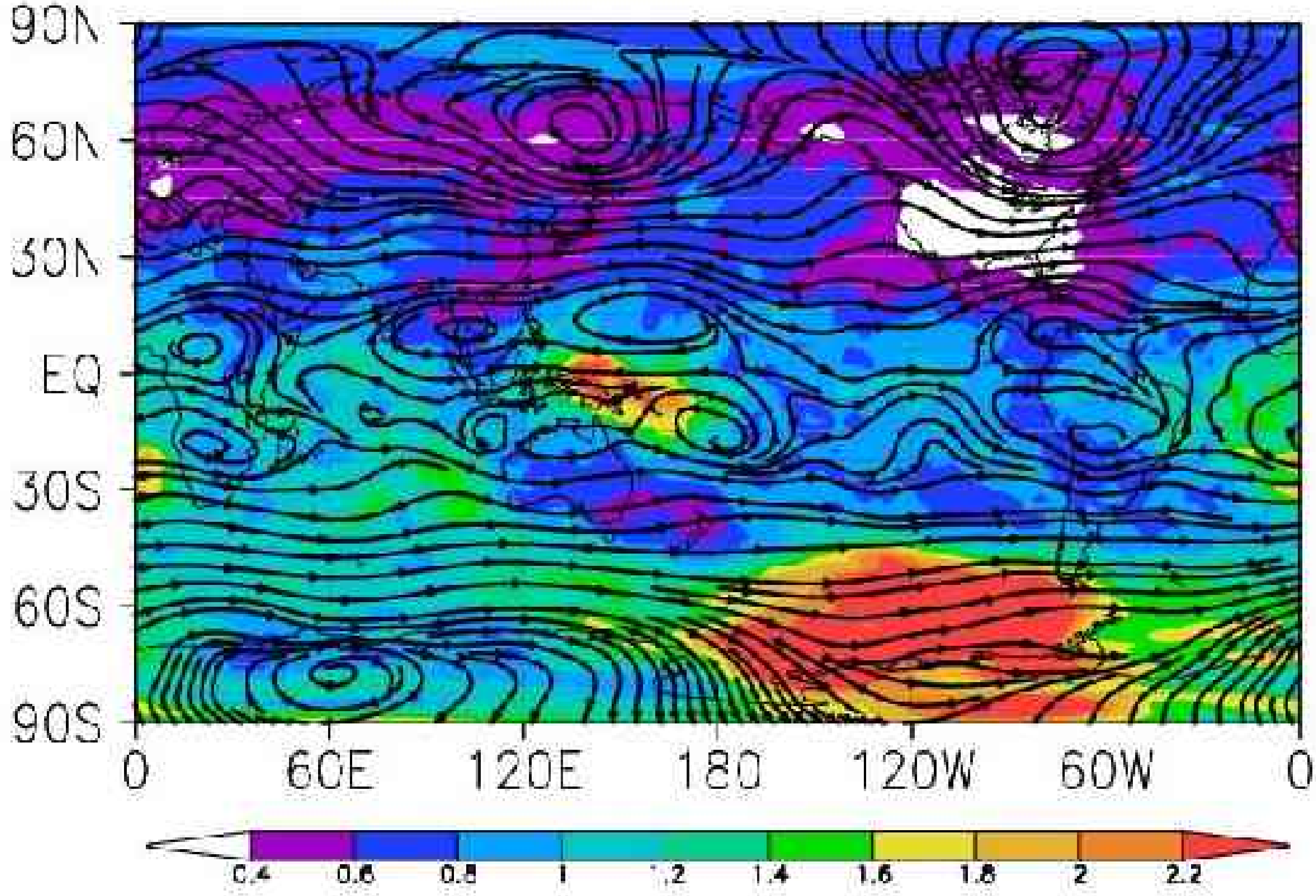
Results with simulated observations

- Observations are created by adding Gaussian random noise to the true state (1 K for temperature, 1 m/s for wind vector components, and 1 hPa for surface pressure)
- No asynchronous observations
- Full and realistic observing networks
- Compare the resulting analysis to the “true” state consisting of 45-60 days of simulated weather

Representative results: Temperature



Error in the u-wind analysis at 300 hPa



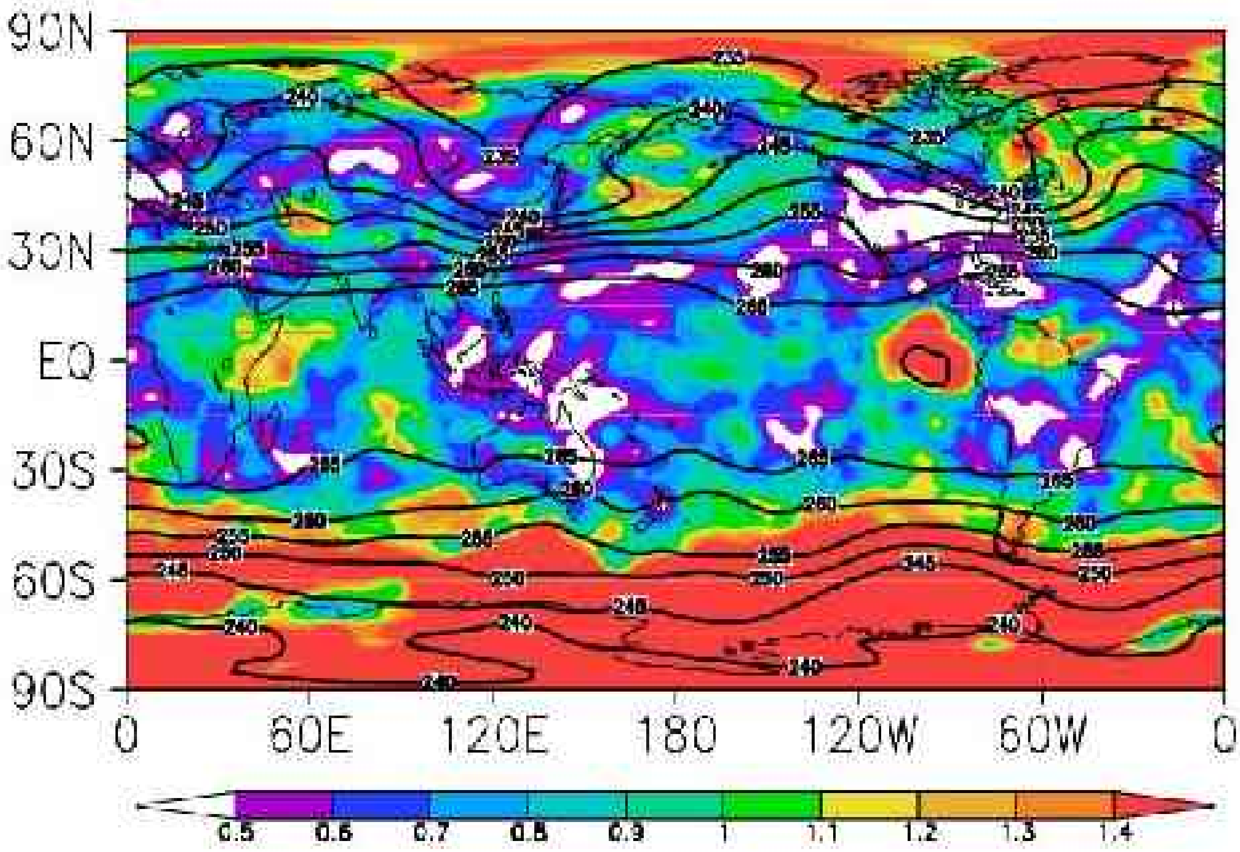
Results with real observations

- Observations are assimilated from a 3-hour window centered at analysis time (no time interpolation)
- All observations are assimilated with except for satellite radiances (~250,000 observations)
- 40-member ensemble, multiplicative variance inflation (25% in NH extra-tropics, 20% in tropics, and 15% in SH extra-tropics)
- April 2004 version of operational GFS
- Data are taken from January-February 2004
- Four cycles per day for 30 days

Comparisons with NCEP analyses

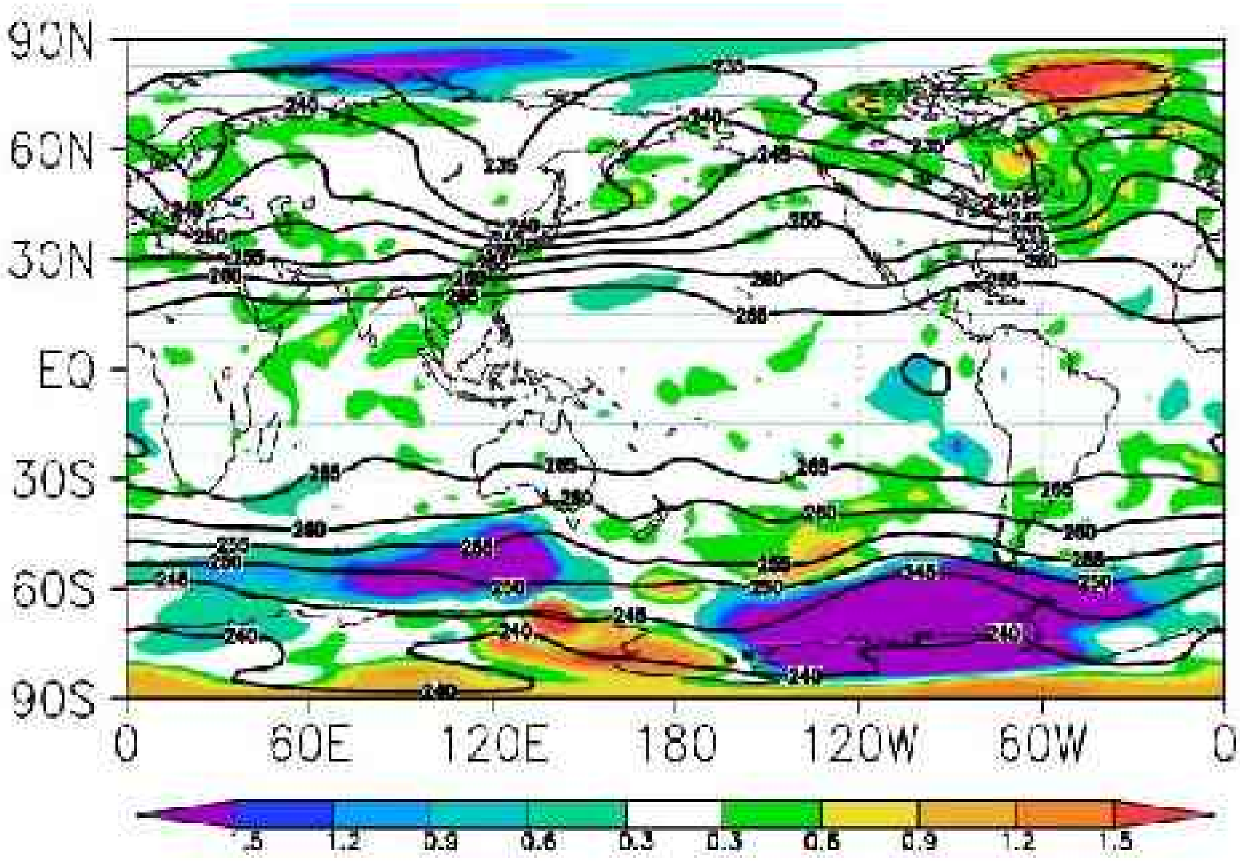
- “Benchmark” analysis: NCEP analysis prepared with the same dataset (no satellite data) with T62 version of the model
- “Operational” analysis: high-resolution (T254) model, includes satellite data
- Compute |LETKF–Operational| and |LETKF–Operational| – |Benchmark–Operational|

Difference Between the LETKF and Operational NCEP Temperature Analyses at 600 hPa



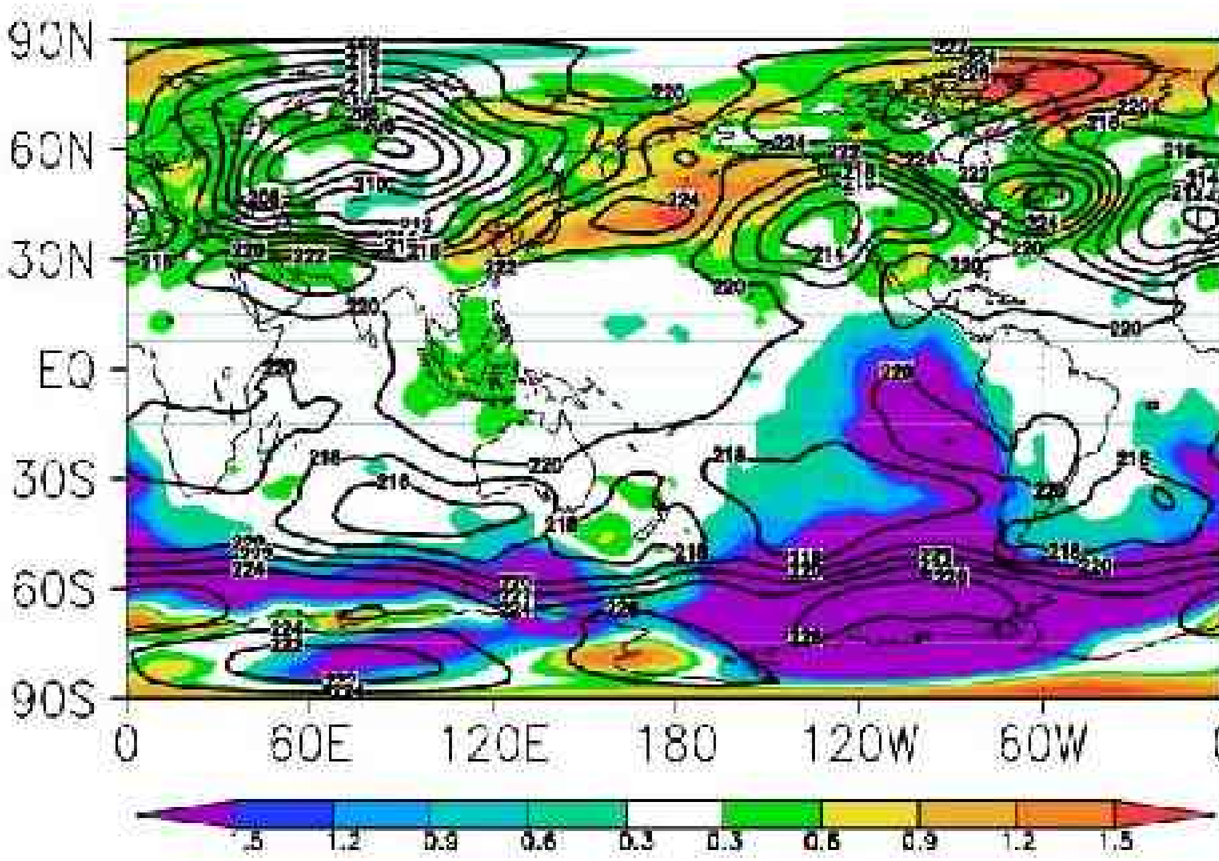
The rms difference is calculated over 84 analysis cycles

|LETKF–Operational| – |Benchmark–Operational| 600 hPa Temperature



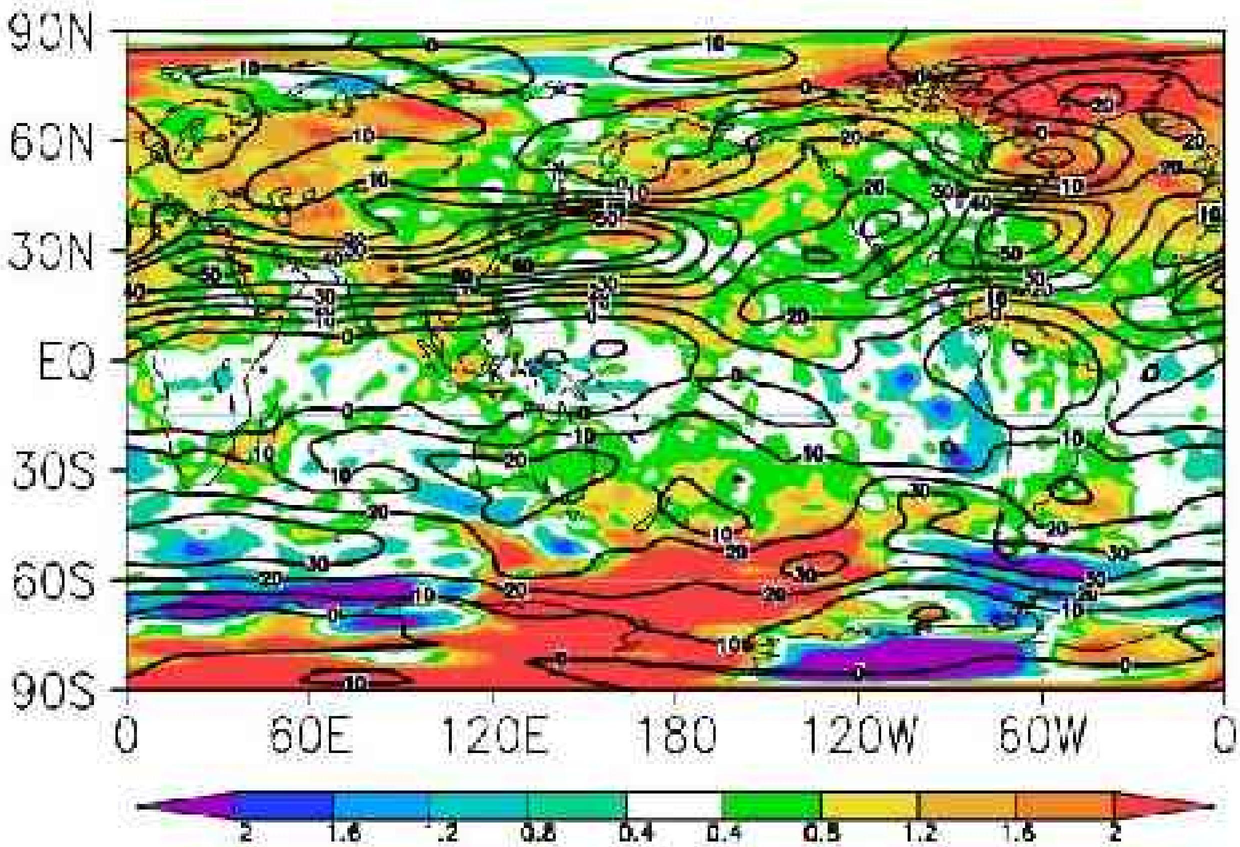
Negative values indicate that the LETKF analysis is more similar to the operational analysis than the benchmark

|LETKF–Operational| – |Benchmark–Operational|
200 hPa Temperature



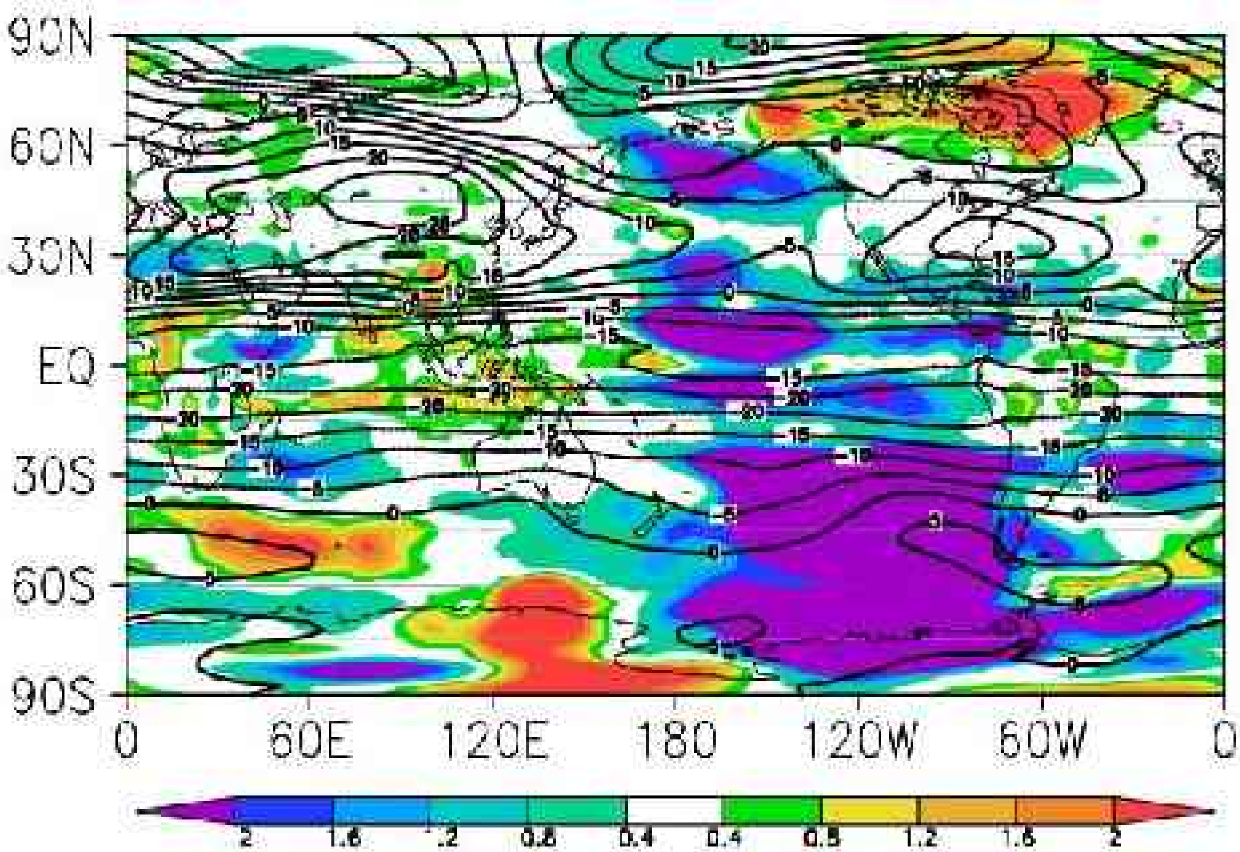
Negative values indicate that the LETKF analysis is more similar to the operational analysis than the benchmark

|LETKF–Operational| – |Benchmark – Operational|
200 hPa u-wind



Negative values indicate that the LETKF analysis is more similar to the operational analysis than the benchmark

|LETKF–Operational| – |Benchmark–Operational|
50 hPa u-wind



Negative values indicate that the LETKF analysis is more similar to the operational analysis than the benchmark

Conclusions

- The LETKF with a 40-member ensemble provides a stable analysis cycle for real observations
- In areas of high observational density, the LETKF analysis is *very similar* to the operational NCEP analysis
- The LETKF efficiently propagates information from data-dense to data-sparse regions
- Work in progress:
  - Time interpolation (“4d”) implementation and tuning
  - Verification of short term forecasts against observations
  - Implementation of bias correction