The Future of Data Assimilation:
4D-Var or Ensemble Kalman Filter?

Eugenia Kalnay
Department of Meteorology and Chaos Group
University of Maryland

Chaos/Weather group at the University of Maryland:
Profs. **Hunt, Szunyogh**, Kostelich, Ott, Sauer, Yorke, Kalnay
Former students: Drs. Patil, Corazza, Zimin, Gyarmati, Oczkowski, Shu-Chih Yang, Miyoshi
Current students: Klein, Li, Liu, Danforth, Merkova, Harlim, Kuhl, Baek

References and thanks: **(www.atmos.umd.edu/~ekalnay)**

Ott, Hunt, Szunyogh, Zimin, Kostelich, Corazza, Kalnay, Patil, Yorke, 2004: Local Ensemble Kalman Filtering, Tellus, 56A,415–428.

Hunt, Kalnay, Kostelich, Ott, Szunyogh, Patil, Yorke, Zimin, 2004: Four-dimensional ensemble Kalman filtering. Tellus 56A, 273–277.

Szunyogh, Kostelich, Gyarmati, Hunt, Ott, Zimin, Kalnay, Patil, Yorke, 2005: Assessing a local ensemble Kalman filter: Perfect model experiments with the NCEP global model. Tellus, 56A, in print.

Yang, Corazza and Kalnay, 2004: Errors of the day, bred vectors and singular vectors: implications for ensemble forecasting and data assimilation. AMS NWP Conference.

Miyoshi, 2005: Ensemble Kalman Filtering experiments with a primitive equations model. Ph.D. thesis, University of Maryland.

Hunt et al, 2005: Local Ensemble Kalman Filter: Efficient Implementation, in prep.

Patil, Hunt, Kalnay, Yorke and Ott, 2001: Local low-dimensionality of atmospheric dynamics, PRL.

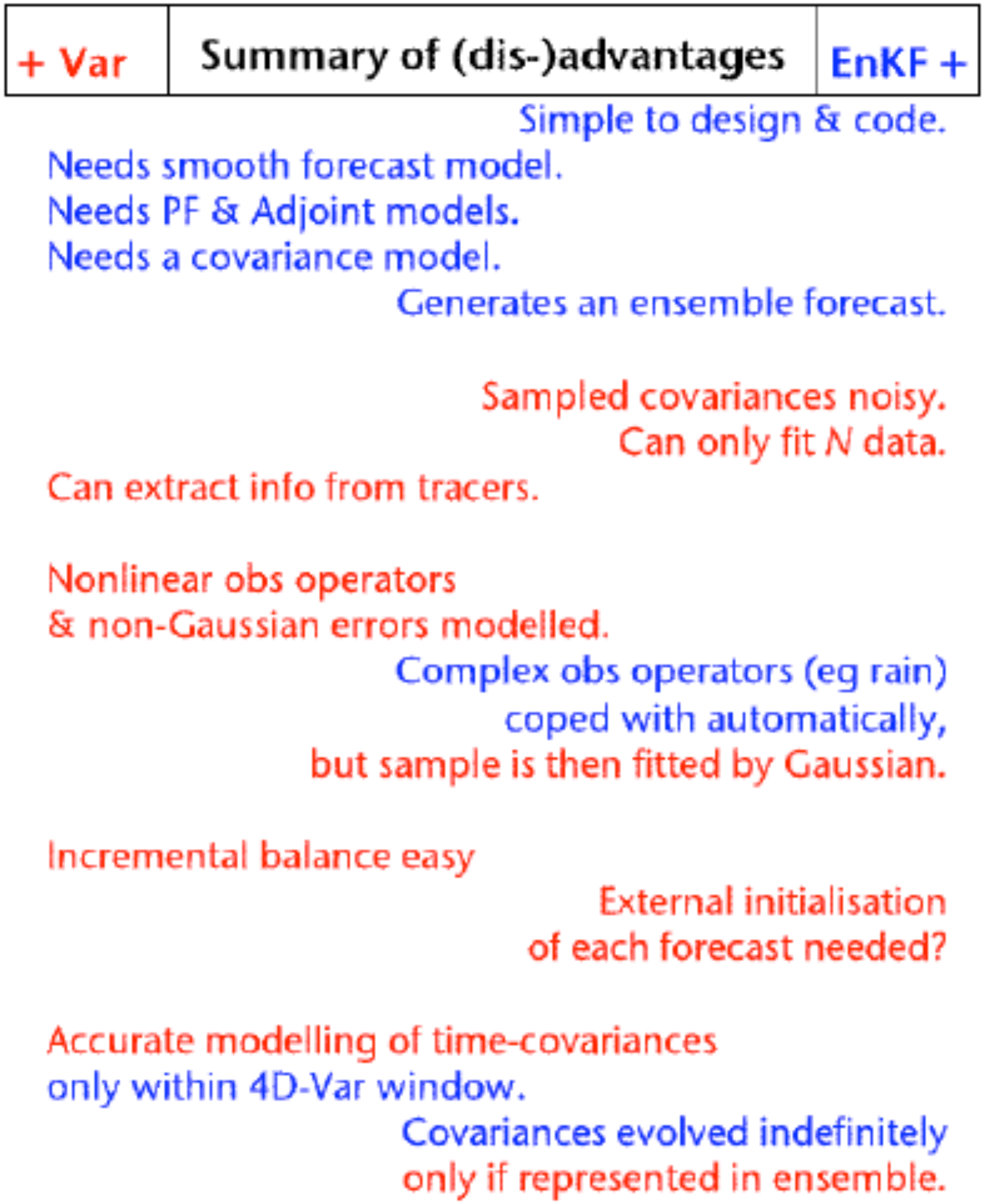
Corazza, Kalnay, Patil, Yang, Hunt, Szunyogh, Yorke, 2003: Relationship between bred vectors and the errors of the day. NPG.

Kalnay, 2003: Atmospheric modeling, data assimilation and predictability, Cambridge University Press, 341 pp. **(Chinese edition, 2005).**

Data assimilation: present and future

- Many centers (including NCEP) still use 3D-Var
- 3D-Var does not include "errors of the day"
- Several centers (ECMWF, France, UK, Japan, Canada) have switched to 4D-Var
- Kalman Filter is optimal but far too costly
- Ensemble Kalman Filter is still experimental
- In Canada, 4D-Var was clearly better than 3D-Var, but EnKF was only comparable to 3D-Var
- Who will win, 4D-Var or EnKF? How soon?

Lorenc (2004):
“Relative merits of 4DVar
and EnKF”



+ Var	Summary of (dis-)advantages	EnKF +

Simple to design & code.

Needs smooth forecast model.
Needs PF & Adjoint models.
Needs a covariance model.

Generates an ensemble forecast.

Sampled covariances noisy.
Can only fit *N* data.

Can extract info from tracers.

Nonlinear obs operators
& non-Gaussian errors modelled.

Complex obs operators (eg rain)
coped with automatically,
but sample is then fitted by Gaussian.

Incremental balance easy

External initialisation
of each forecast needed?

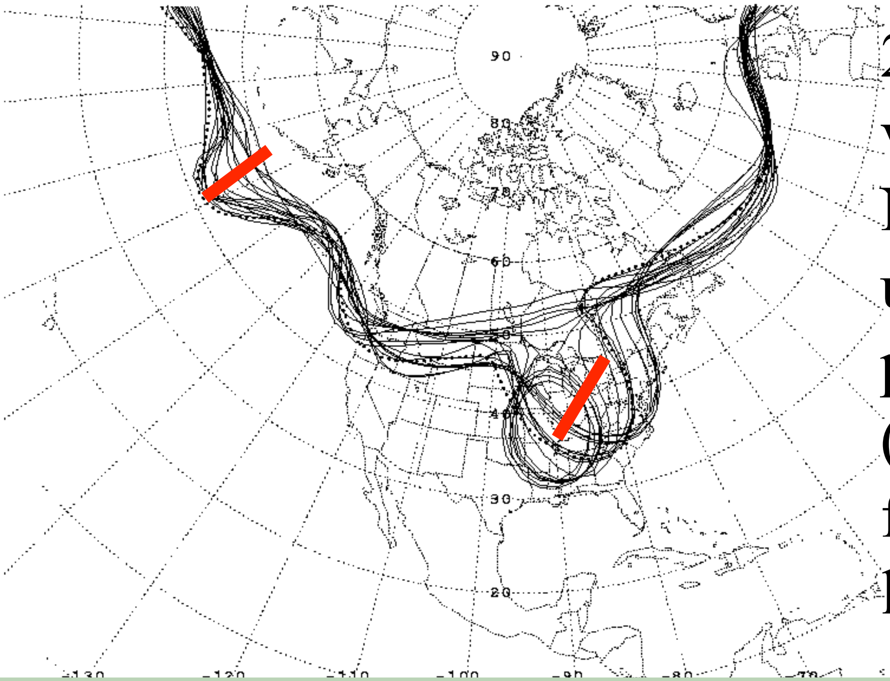
Accurate modelling of time-covariances
only within 4D-Var window.

Covariances evolved indefinitely
only if represented in ensemble.

Errors of the day

- They are instabilities of the background flow
- They dominate the analysis and forecast errors
- They are not taken into account in data assimilation except for 4D-Var and Kalman Filter (very expensive)
- Their shape can be estimated with breeding
- Their shape is frequently simple (low dimensionality, Patil et al, 2001)

“Errors of the day” grow because of instabilities of the flow. Strong instabilities have a few dominant shapes (d.o.f.)



2.5 day ensemble forecast verifying on 95/10/21. Note that where the uncertainties are large, the perturbations (difference between the forecasts) lie on a locally very low-dimensional space

It makes sense to assume that **large** errors in the analysis (initial conditions) are in similarly low-dimensional spaces that can be locally represented by a a low order (~100) EnKF

3D-Var

$$J = \min \frac{1}{2}[(\mathbf{x}^a - \mathbf{x}^b)^{\mathbf{T}} \mathbf{B}^{-1} (\mathbf{x}^a - \mathbf{x}^b) + (H\mathbf{x}^a - \mathbf{y})^{\mathbf{T}} \mathbf{R}^{-1} (H\mathbf{x}^a - \mathbf{y})]$$

Distance to forecast

Distance to observations

at the analysis time

4D-Var

$$J = \min \frac{1}{2}[(\mathbf{x}_0 - \mathbf{x}_0^b)^{\mathbf{T}} \mathbf{B}^{-1} (\mathbf{x}_0 - \mathbf{x}_0^b) + \sum_{i=1}^{s} (H\mathbf{x}_i - \mathbf{y}_i)^{\mathbf{T}} \mathbf{R}_i^{-1} (H\mathbf{x}_i - \mathbf{y}_i)]$$

Distance to background at the initial time

Distance to observations in a **time window interval t_0-t_1**

Control variable $\mathbf{x}(t_0)$

Analysis $\mathbf{x}(t_1) = M[\mathbf{x}(t_0)]$

It seems like a simple change, but it is not! (e.g., adjoint)

What is B? It should be tuned…

Extended Kalman Filter (EKF)

Forecast step:

$$\mathbf{x}_n^b = M_n\left(\mathbf{x}_{n-1}^a\right)$$

$$\mathbf{B}_n = \mathbf{M}_n\mathbf{A}_{n-1}\mathbf{M}_n^T + \mathbf{Q}_n$$

Analysis step:

$$\mathbf{x}_n^a = \mathbf{x}_n^b + \mathbf{K}_n(\mathbf{y}_n - H\mathbf{x}_n^b)$$

where the optimal weight matrix is given by

$$\mathbf{K}_n = \mathbf{B}_n(\mathbf{R} + \mathbf{H}\mathbf{B}_n\mathbf{H}^T)^{-1}$$

and the new analysis error covariance by

$$\mathbf{A}_n = (\mathbf{I} - \mathbf{K}_n\mathbf{H})_n$$

Ensemble Kalman Filter (EnKF)

Forecast step:

$$\mathbf{x}^{b}_{n,k} = M_n\left(\mathbf{x}^{a}_{n-1,k}\right)$$

$$\mathbf{B}_n = \frac{1}{K-1}\mathbf{E}^b_n\mathbf{E}^{bT}_n, \; where \; \mathbf{E}^b_n = \left[\mathbf{x}^b_{n,1} - \overline{\mathbf{x}}^b_n; \ldots, \mathbf{x}^b_{n,K} - \overline{\mathbf{x}}^b_n\right]$$

Analysis step:

$$\mathbf{x}^a_n = \mathbf{x}^b_n + \mathbf{K}_n(\mathbf{y}_n - H\mathbf{x}^b_n) \qquad \hat{\mathbf{B}}_n = \mathbf{I}$$

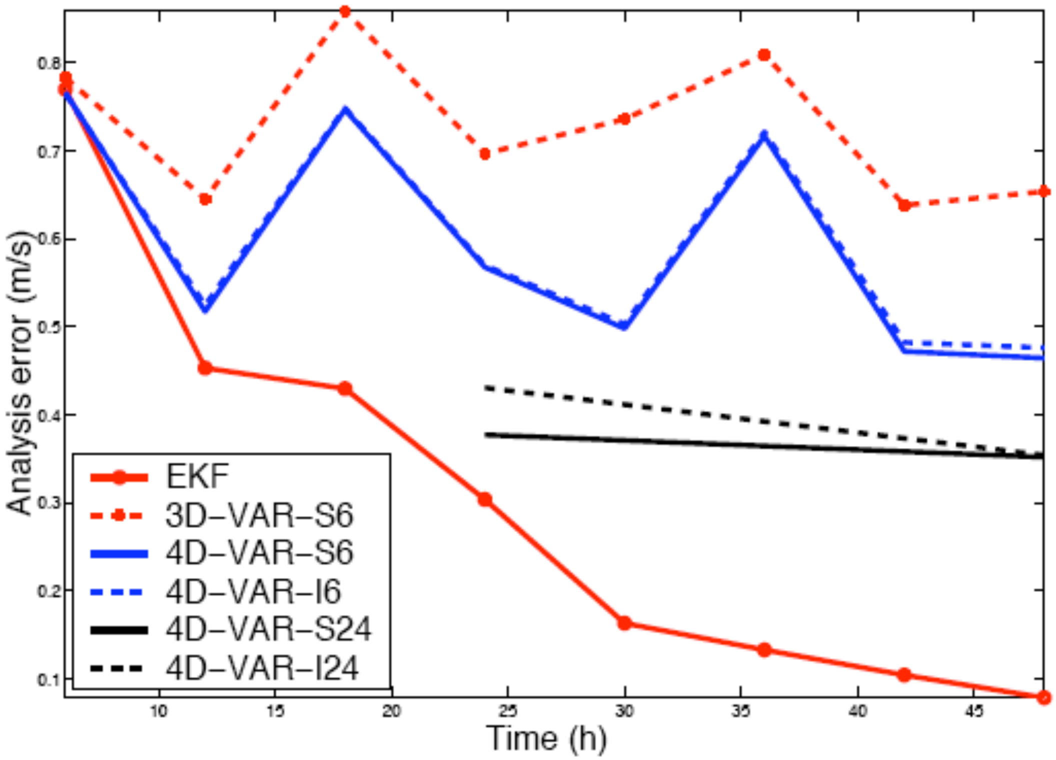
The new analysis error covariance in the ensemble space is (Hunt,2005)

$$\hat{\mathbf{A}}_n = \left[\left(K-1\right)\mathbf{I} + (\mathbf{HE}^b_n)^T\mathbf{R}^{-1}(\mathbf{HE}^b_n)\right]^{-1}$$

And the new ensemble perturbations are given by

$$\mathbf{E}^a_n = \mathbf{E}^b_n\left[\left(K-1\right)\hat{\mathbf{A}}_n\right]^{1/2}$$

From F. Rabier & Z. Liu (2003): 3D-Var, 4D-Var and Extended Kalman Filter



- 4D-Var better than 3D-Var
- Longer 4D-Var window (24h) better than shorter (6h)
- Extended Kalman Filter (at full resolution) is the best because it updates the analysis error covariance

The question is: can Ensemble Kalman Filter (with ~100 d.o.f.) do as well as Extended Kalman Filter or 4D-Var (with 10^7 d.o.f)?

The solution to the cost of EKF: Ensemble Kalman Filter (EnKF)

1) **Perturbed observations:** ensembles of data assimilation

- Evensen, 1994
- Houtekamer and Mitchell, 1998, 2005 (Canada)

Perturbing the obs. introduces <u>sampling errors</u>

2) **Square root filter**, no need for perturbed observations:

- Tippett, Anderson, Bishop, Hamill, Whitaker, 2003
- Anderson, 2001
- Whitaker and Hamill, 2002, **2005 (AMS) results**
- Bishop, Etherton and Majumdar, 2001

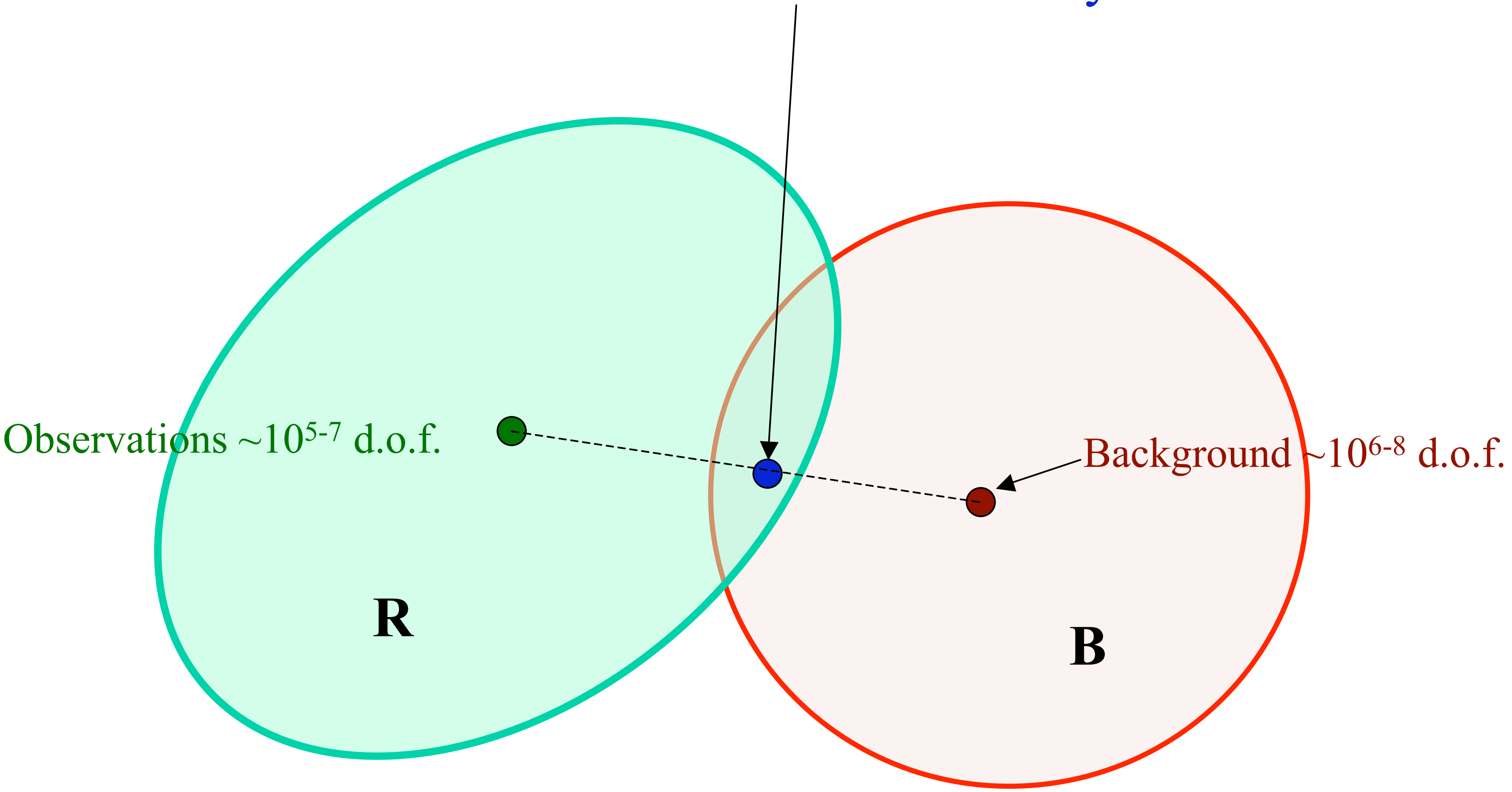
One obs. at a time (sequential): <u>inefficient for many obs.</u>

3) **Local Ensemble Kalman Filtering**: done in local patches

- Ott et al, 2002, 2004, Szunyogh et al 2005.
- Hunt et al, 2004, extended it to 4DEnKF
- Hunt, 2005, LETKF: 5x more efficient (no SVD)

Suppose we have a 6hr forecast (background) and new observations

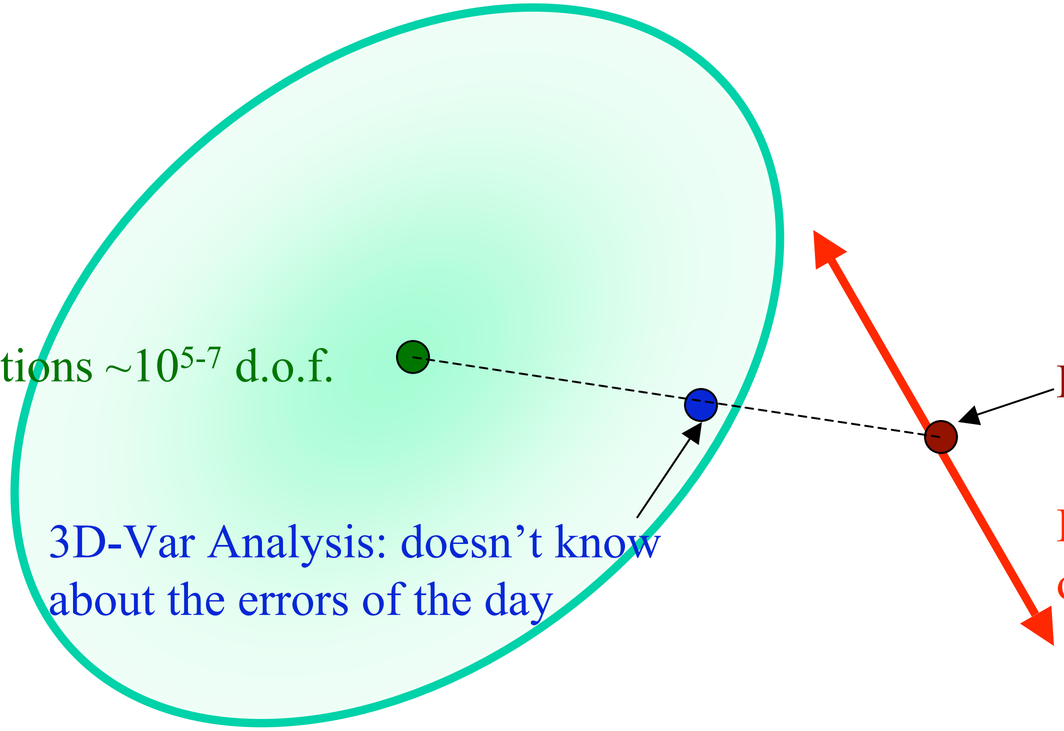
The 3D-Var Analysis doesn’t know about the errors of the day



Observations ~10^{5-7} d.o.f.

Background ~10^{6-8} d.o.f.

With Ensemble Kalman Filtering we get perturbations pointing to the directions of the “errors of the day”



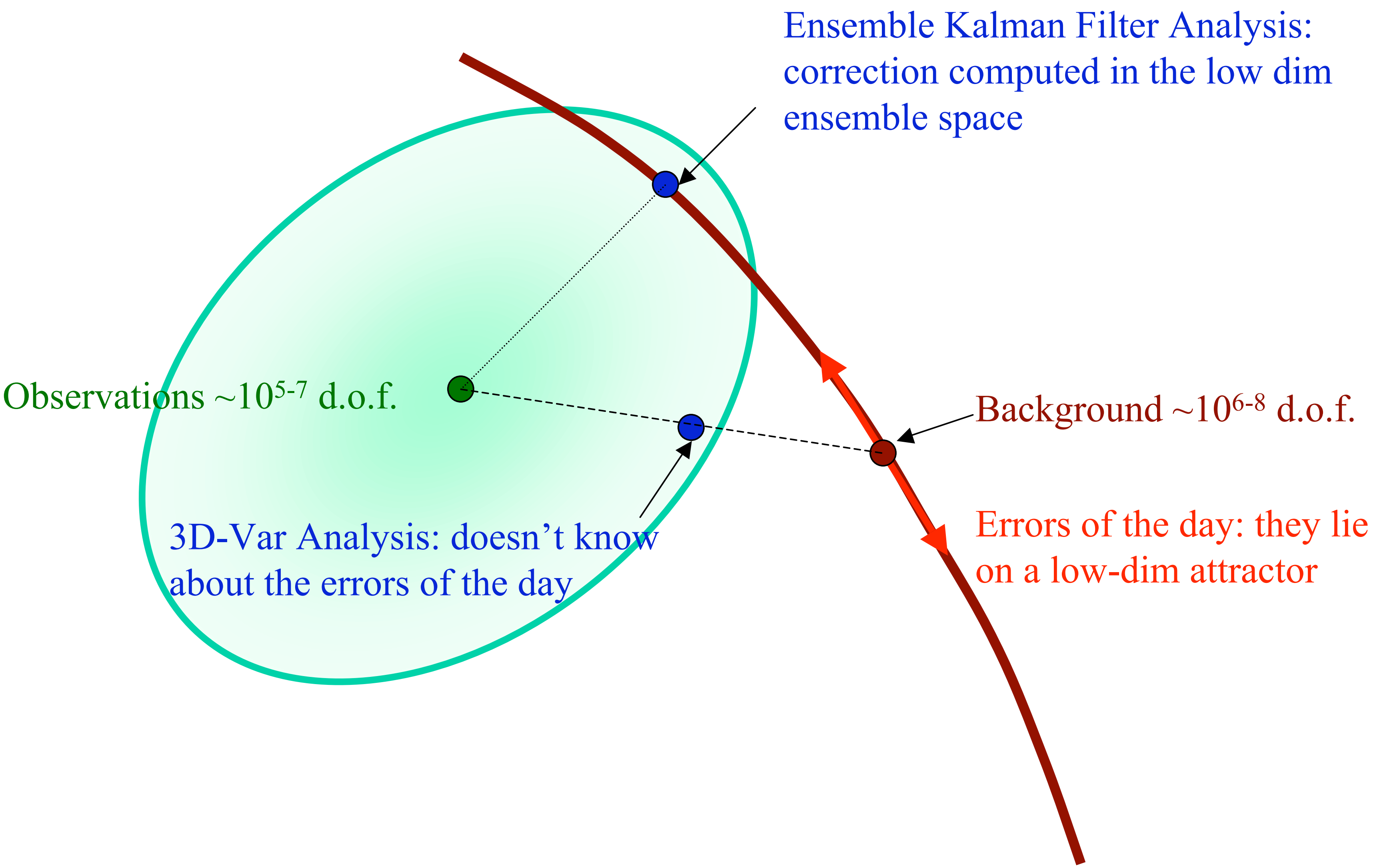
Observations ~10^{5-7} d.o.f.

Background ~10^{6-8} d.o.f.

3D-Var Analysis: doesn’t know about the errors of the day

Errors of the day: they lie on a low-dim attractor

Ensemble Kalman Filtering is efficient because matrix operations are performed in the low-dimensional space of the ensemble perturbations



Ensemble Kalman Filter Analysis: correction computed in the low dim ensemble space

Observations ~10^{5-7} d.o.f.

Background ~10^{6-8} d.o.f.

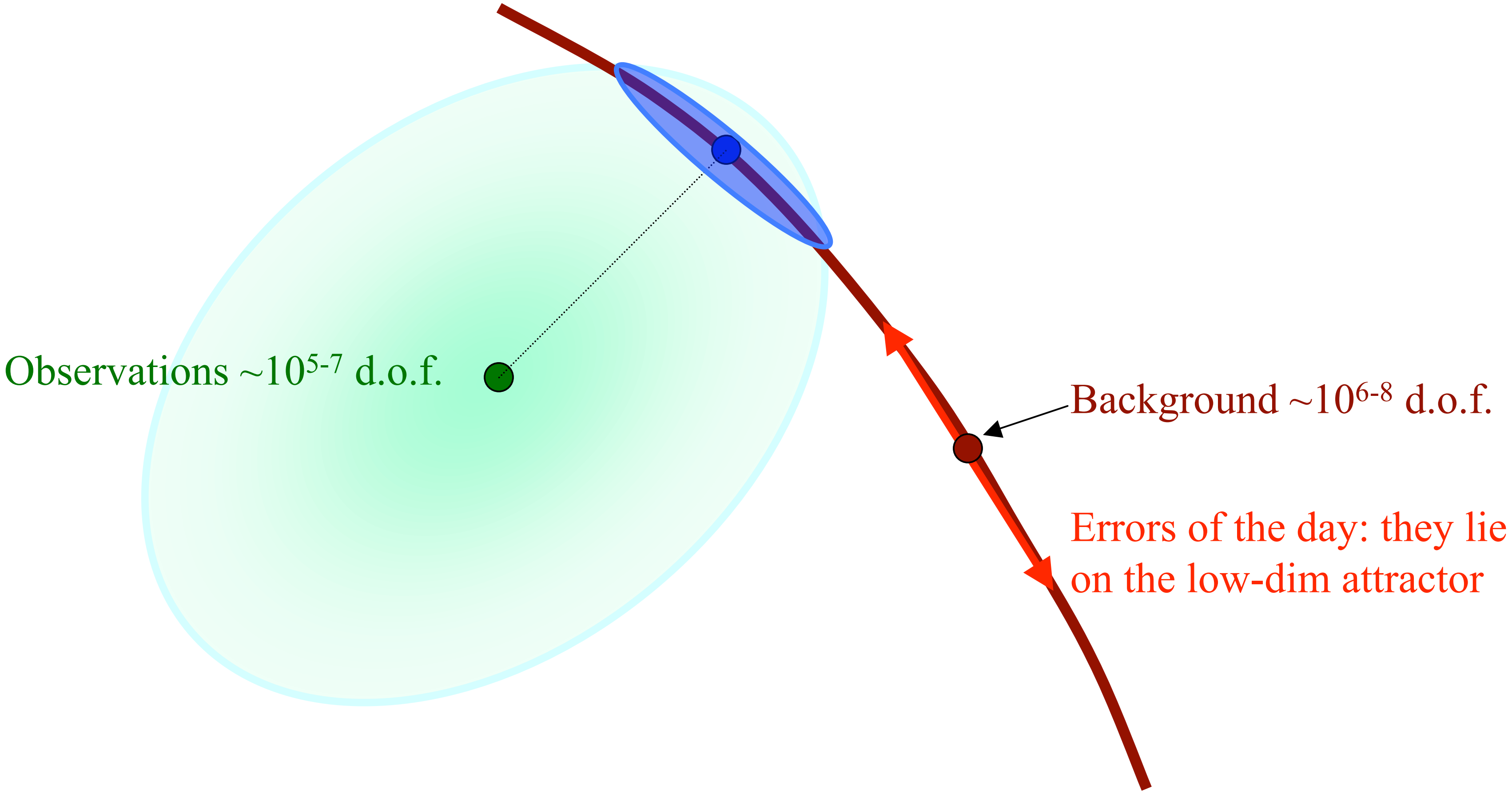
3D-Var Analysis: doesn’t know about the errors of the day

Errors of the day: they lie on a low-dim attractor

After the EnKF computes the analysis and the analysis error covariance $\mathbf{A}$, the new ensemble initial perturbations $\delta \mathbf{a}_i$ are computed:

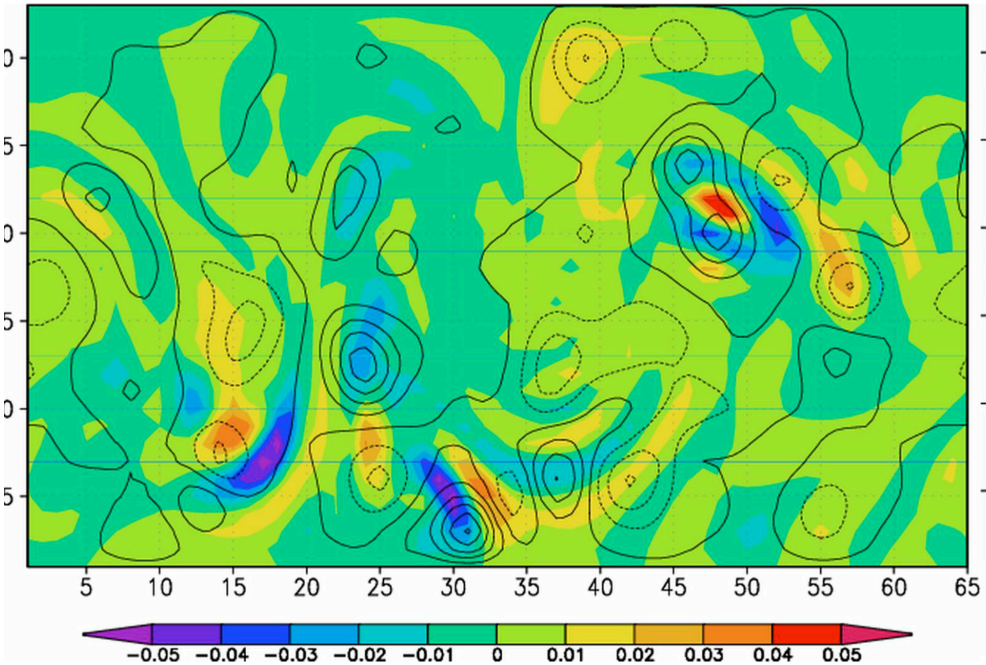
$$\sum_{i=1}^{k+1} \delta \mathbf{a}_i \delta \mathbf{a}_i^T = \mathbf{A}$$

These perturbations represent the analysis error covariance and are used as **initial perturbations** for the next ensemble forecast

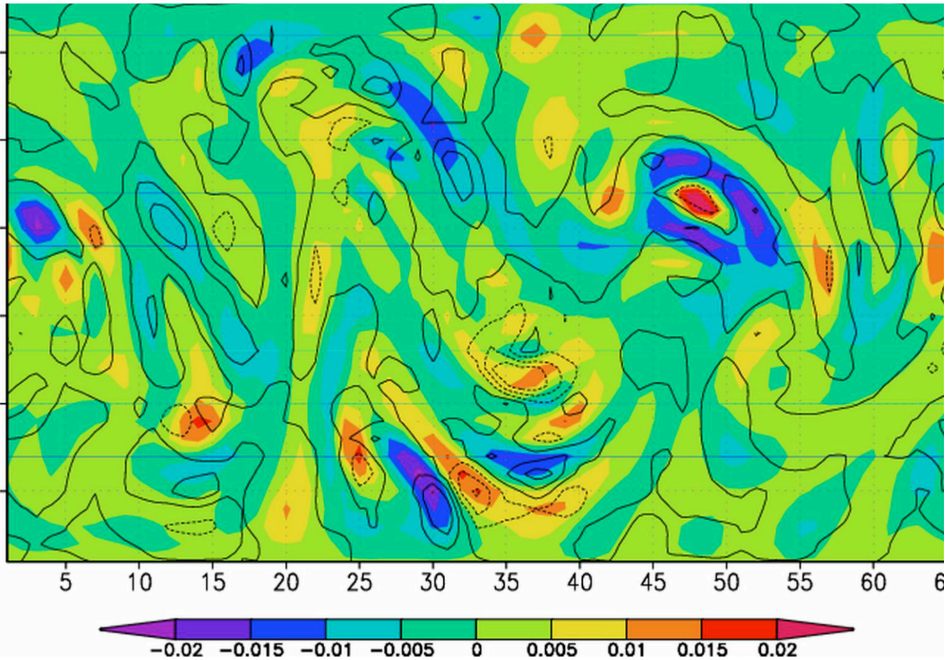

Observations ~10^{5-7} d.o.f.
Background ~10^{6-8} d.o.f.
Errors of the day: they lie on the low-dim attractor

From a QG simulation (Corazza et al, 2003)
Background errors and analysis increments

3D-Var



EnKF



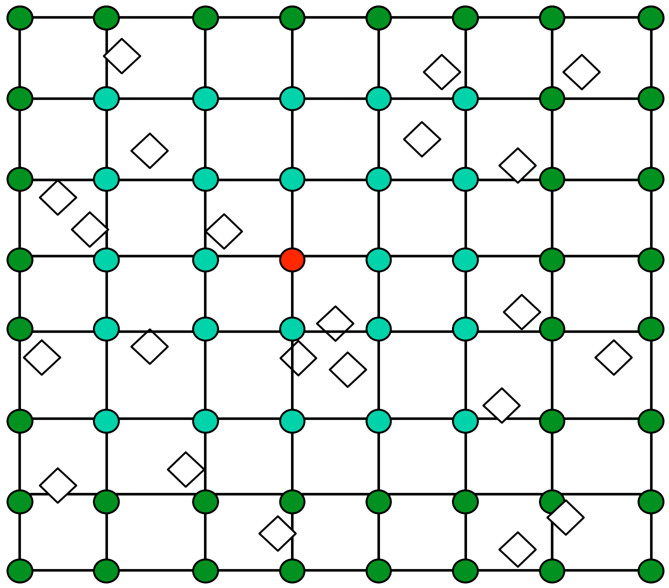
3D-Var does not capture the errors of the day

The EnKF ensemble **B** knows about the errors of the day, and uses the observations more effectively

Local Ensemble Transform Kalman Filter

Perform Data Assimilation in local patch (3D-window)

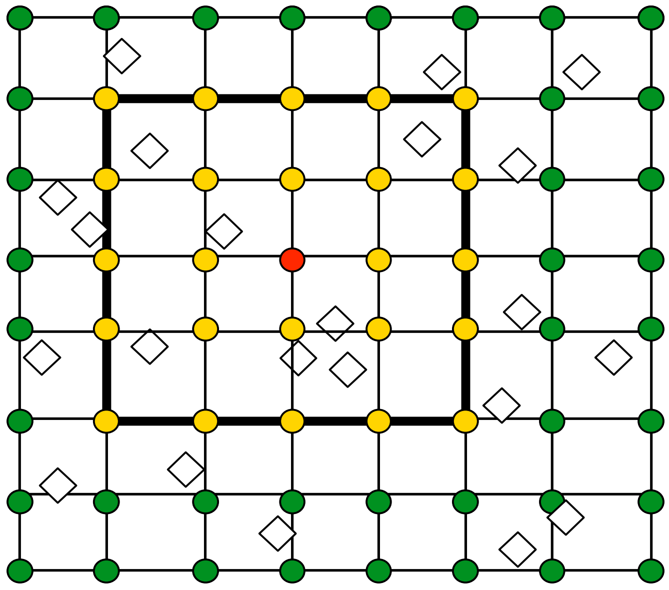
- The state estimate is updated at the central grid red dot



Local Ensemble Transform Kalman Filter

Perform Data Assimilation in local patch (3D-window)

- The state estimate is updated at the central grid red dot

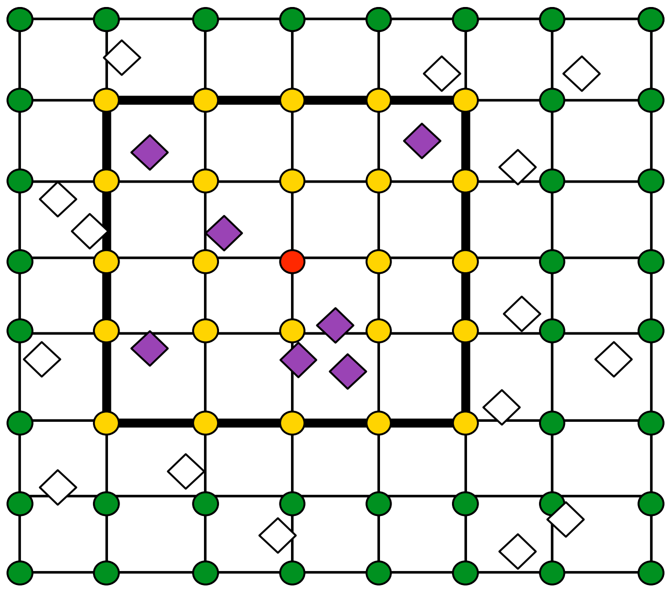


Local Ensemble Transform Kalman Filter

Perform Data Assimilation in local patch (3D-window)

- The state estimate is updated at the central grid red dot

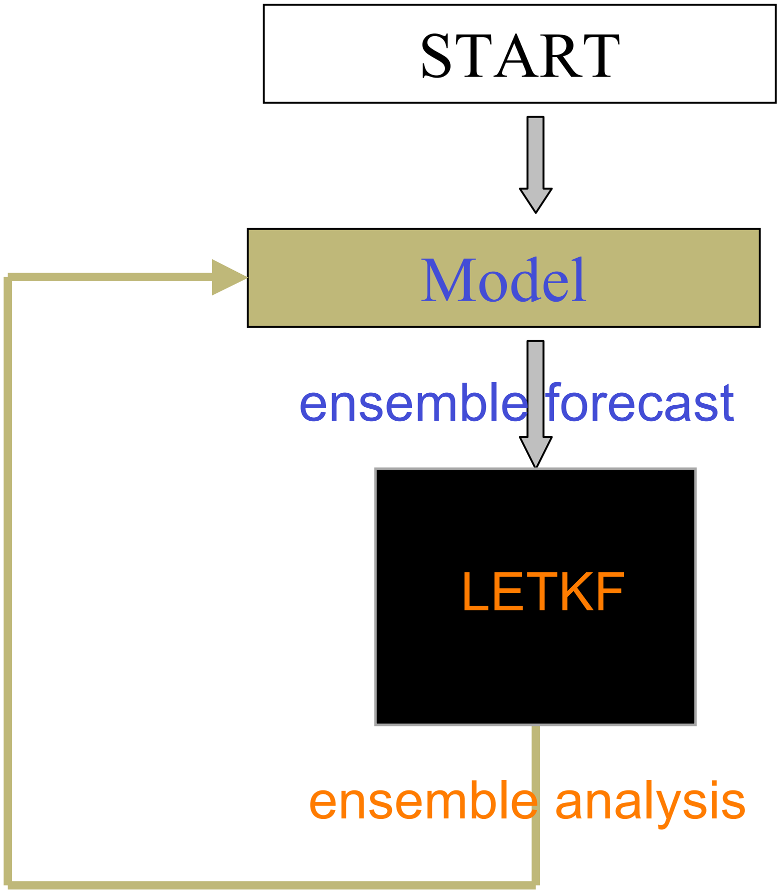
- All observations (purple diamonds) within the local region are assimilated simultaneously



Why use a “local” ensemble approach?

- In the Local Ensemble Kalman Filter we compute the generalized “bred vectors” globally but use them locally (3D cubes around each grid point of ~700km x 700km x 3 layers).
- The ensemble within the local cubes provides the local shape of the “errors of the day”.
- At the end of the local analysis we create a new global analysis and initial perturbations from the solutions obtained at each grid point.
- This reduces the number of ensemble members needed.
- It allows to use **all** the observations in a cube simultaneously.
- **It allows to compute the KF analysis independently at each grid point (“embarrassingly parallel”)**.

Local Ensemble Transform Kalman Filter
(Ott et al, 2004, Hunt et al, 2004, 2005)

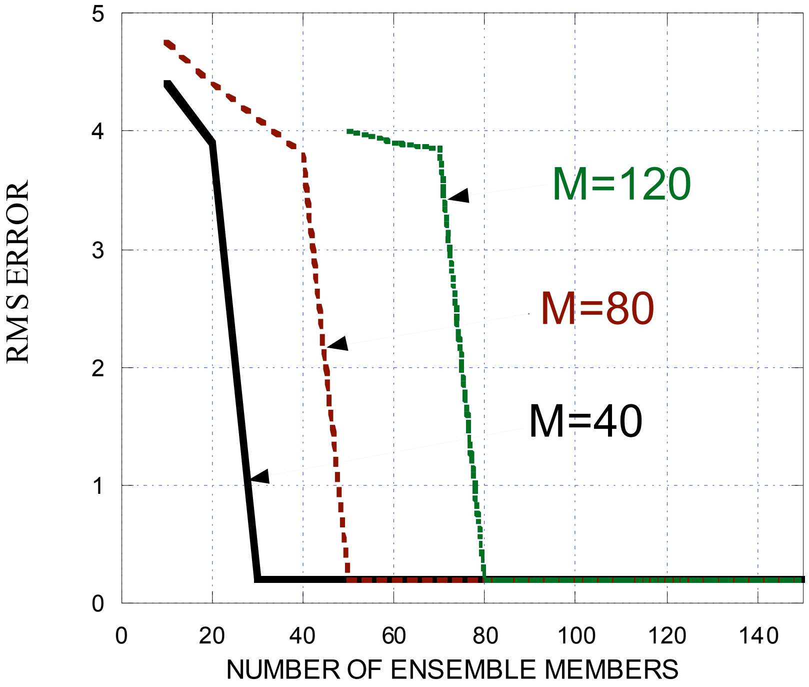


- Model independent
- 100% parallel
- Simultaneous data assim.
- 4D LETKF extension

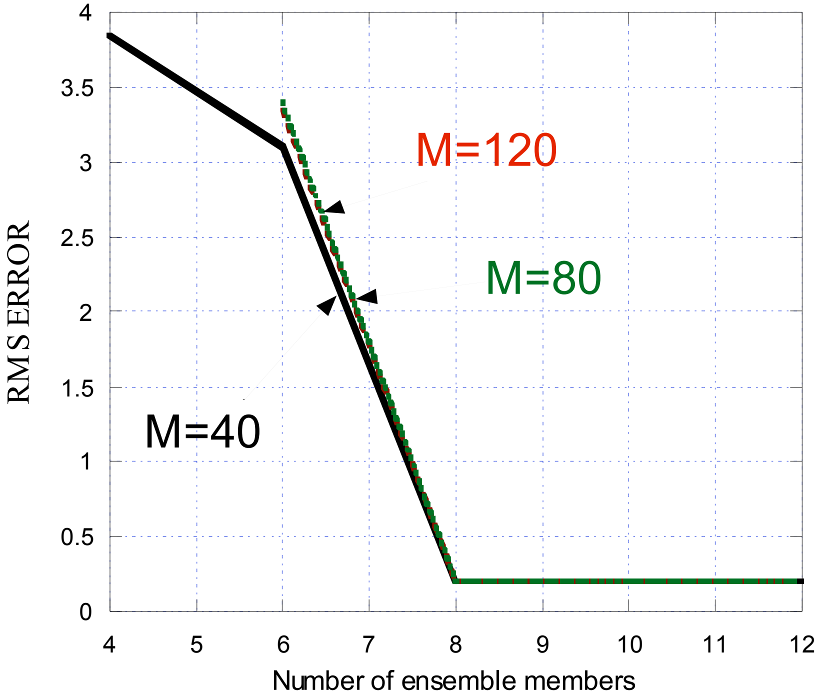
Results with Lorenz 40 variable model (Ott et al. 2004)

- A very large global ensemble Kalman Filter converges to an "optimal" analysis rms error=0.20
- This "optimal" rms error is achieved by the LEKF for a range of small ensemble members
- We performed experiments for different size models: M=40 (original), M=80 and M=120, and compared a global KF with the LEKF

FULL ENSEMBLE KALMAN FILTER ANALYSIS ERROR AS A FUNCTION OF THE NUMBER OF ENSEMBLE MEMBERS



LEKF ANALYSIS ERROR AS A FUNCTION OF THE NUMBER OF ENSEMBLE MEMBERS



With the global EnKF approach, the number of ensemble members needed for convergence increases with the size of the domain M

With the local approach the number of ensemble members remains small (from Ott et al, 2004)

Comparison of EnKF, EKF, 4D-Var with Lorenz (1963)

x,y,z observed every 8 steps (easy), 25 steps (hard)

a) Observations and analysis every 8 time steps			
EnKF, 3 members	EnKF, 6 members	EKF from Yang et al (2005)	4D-Var (W=Window)
0.31 ($\Delta=0.08$)	0.28 ($\Delta=0.04$)	0.32 (μ=0.02, Δ=0)	0.31 (W=48)
b) Observations and analysis every 25 time steps			
EnKF, 3 members	EnKF, 6 members	EKF from Yang et al (2005)	4D-Var (W=Window)
0.71 ($\Delta=0.7$) 0.61 (hybrid + Δ=0.2)	0.59 ($\Delta=0.3$) 0.56 (hybrid, + Δ=0.02)	0.63 (μ=0.1, $\Delta=0.1$)	0.53 (W=75)

To get these results we had to carefully optimize 4D-Var:

Use the exact background in every Runge-Kutta substep in the adjoint
Optimize the background error covariance
Optimize the window length starting with short windows

Experiments with a QG channel model (Corazza, Yang, Carrasi, Miyoshi)

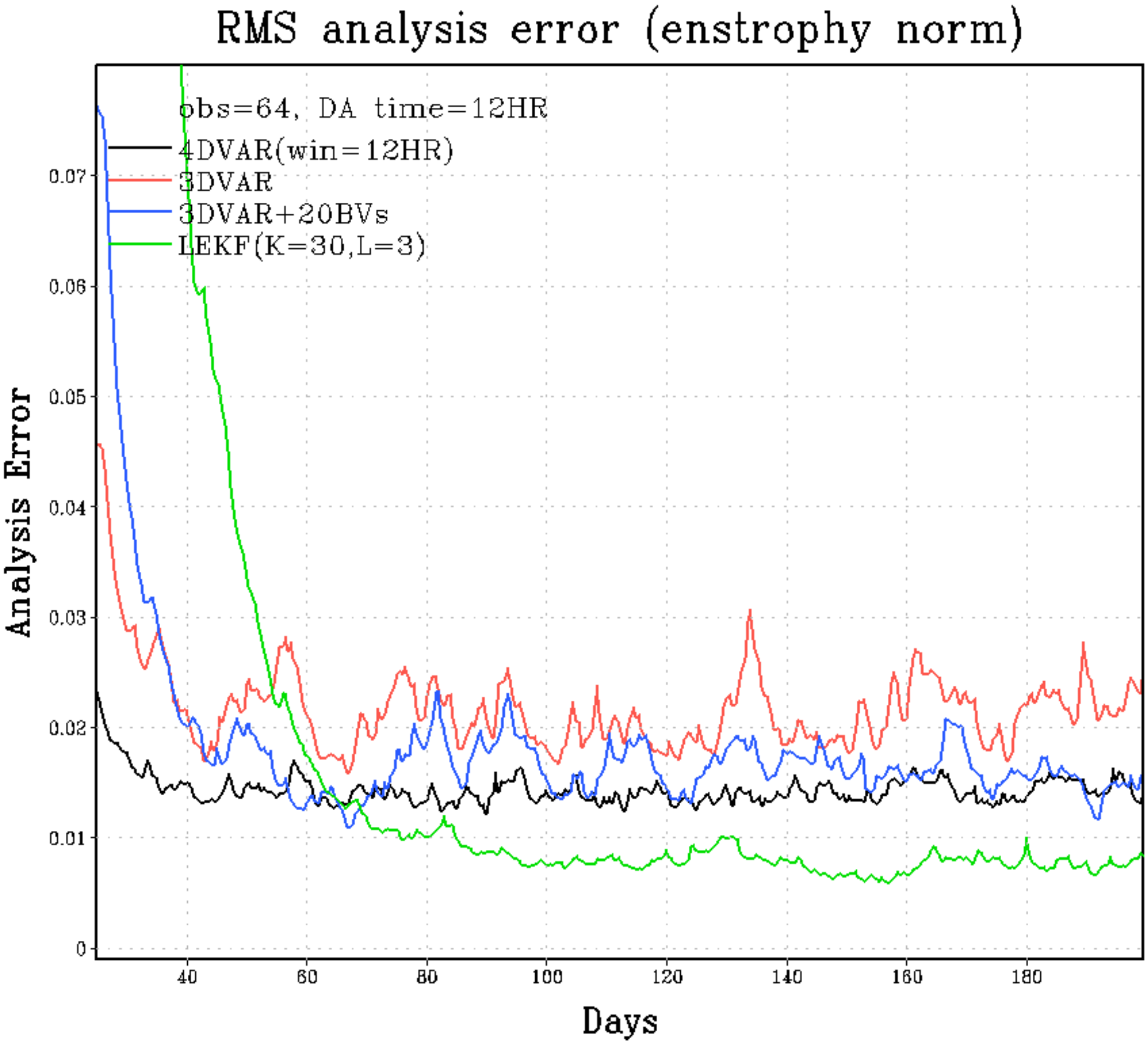
- 3D-Var
- 3D-Var/bred vectors hybrid (almost free);

BV are refreshed with random perturbations

- 4D-Var
- LEKF

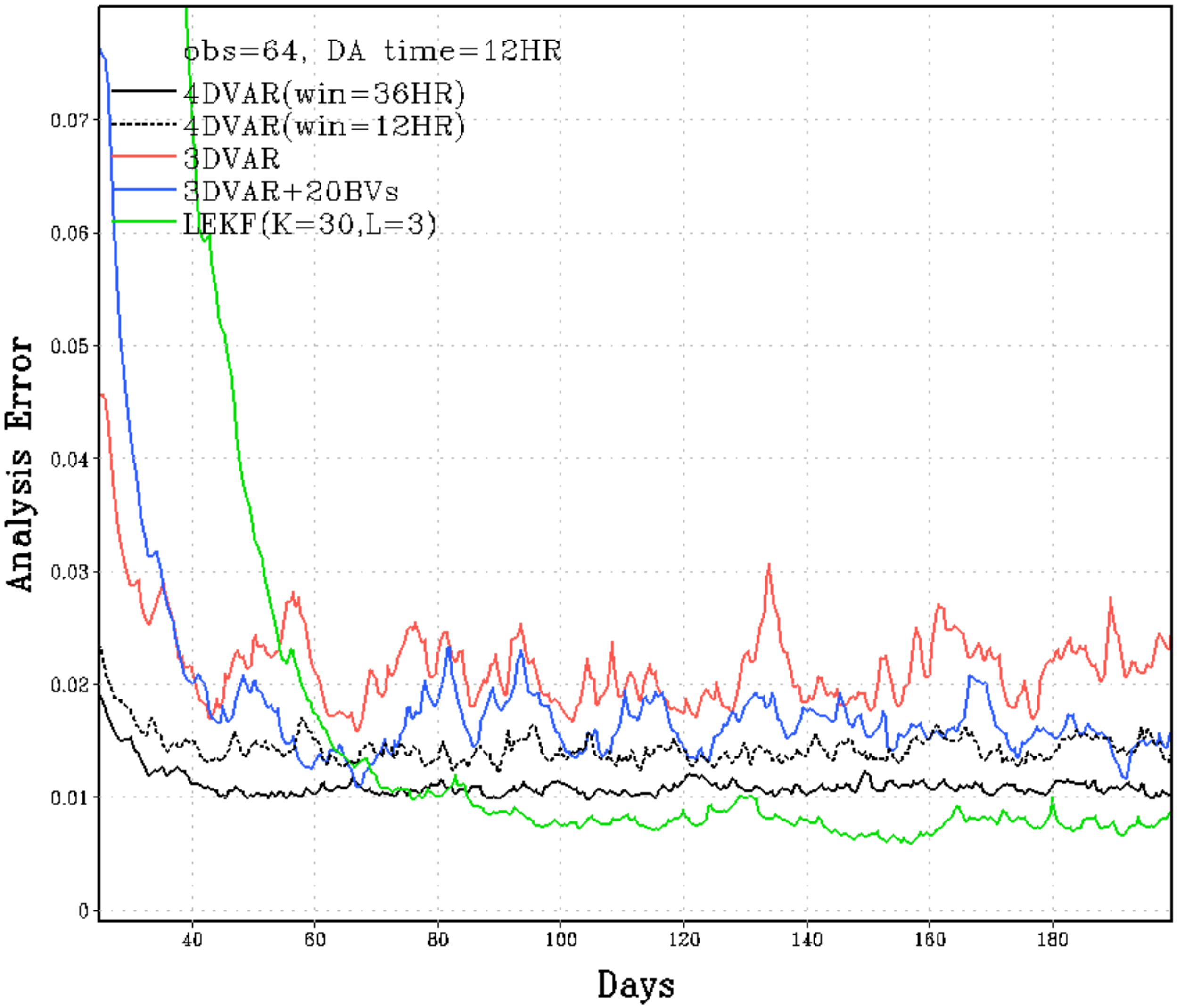
All optimized

3D-Var, 3D-Var augmented with BV, 4D-Var and LEKF



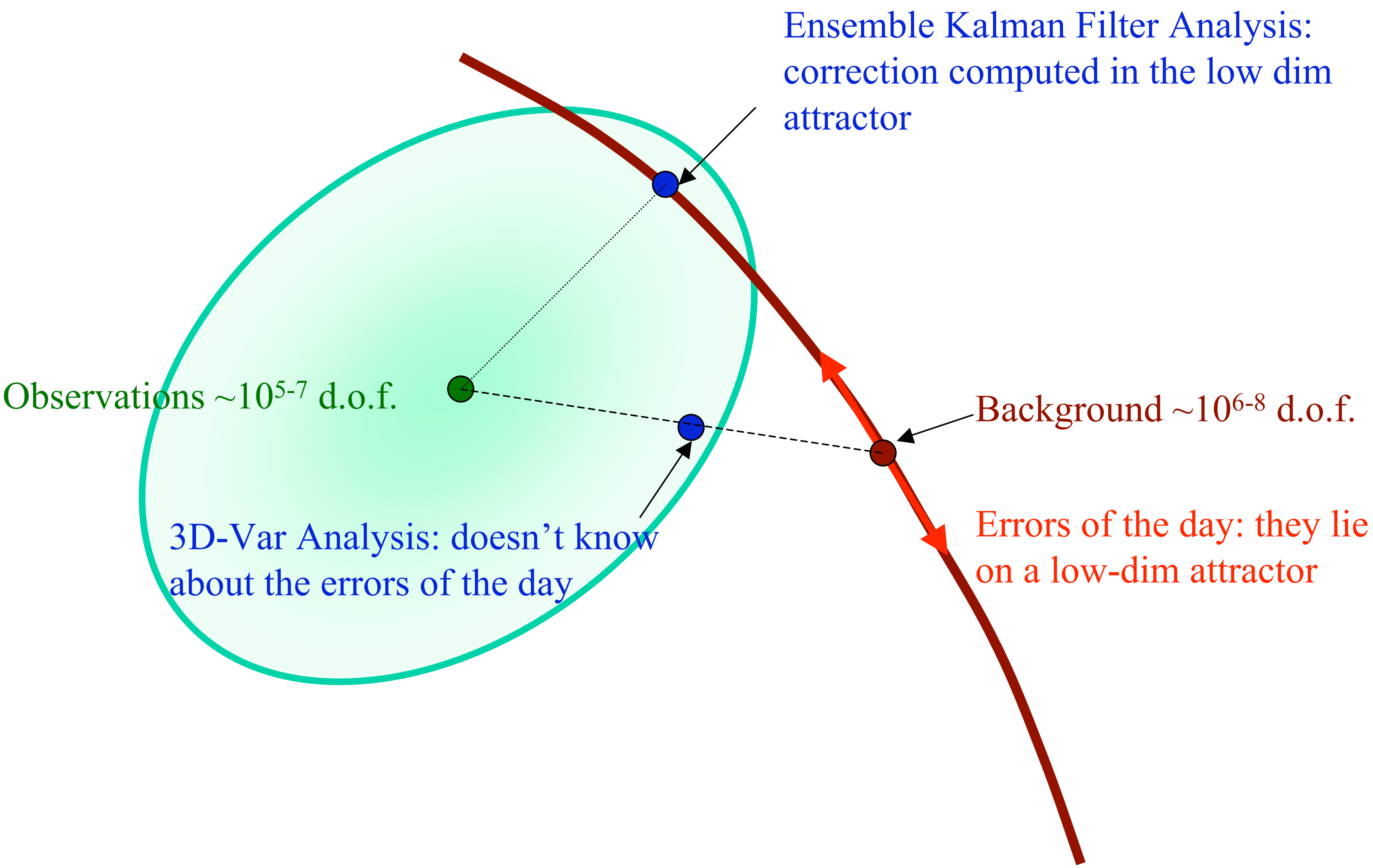
Timings:
3D-Var: 1
3D-Var+BV: 1.4
4D-Var: 36
LEKF: ?

Impact of enlarging the window of 4D-Var



Increasing the 4D-Var data assimilation window from 12 to 36 hr (optimal) improves the results but **increases the cost**

Ensemble Kalman Filtering is prone to slow spin-up if the initial ensemble perturbations do not point in the right direction



Ensemble Kalman Filter Analysis: correction computed in the low dim attractor

Observations ~10^{5-7} d.o.f.

Background ~10^{6-8} d.o.f.

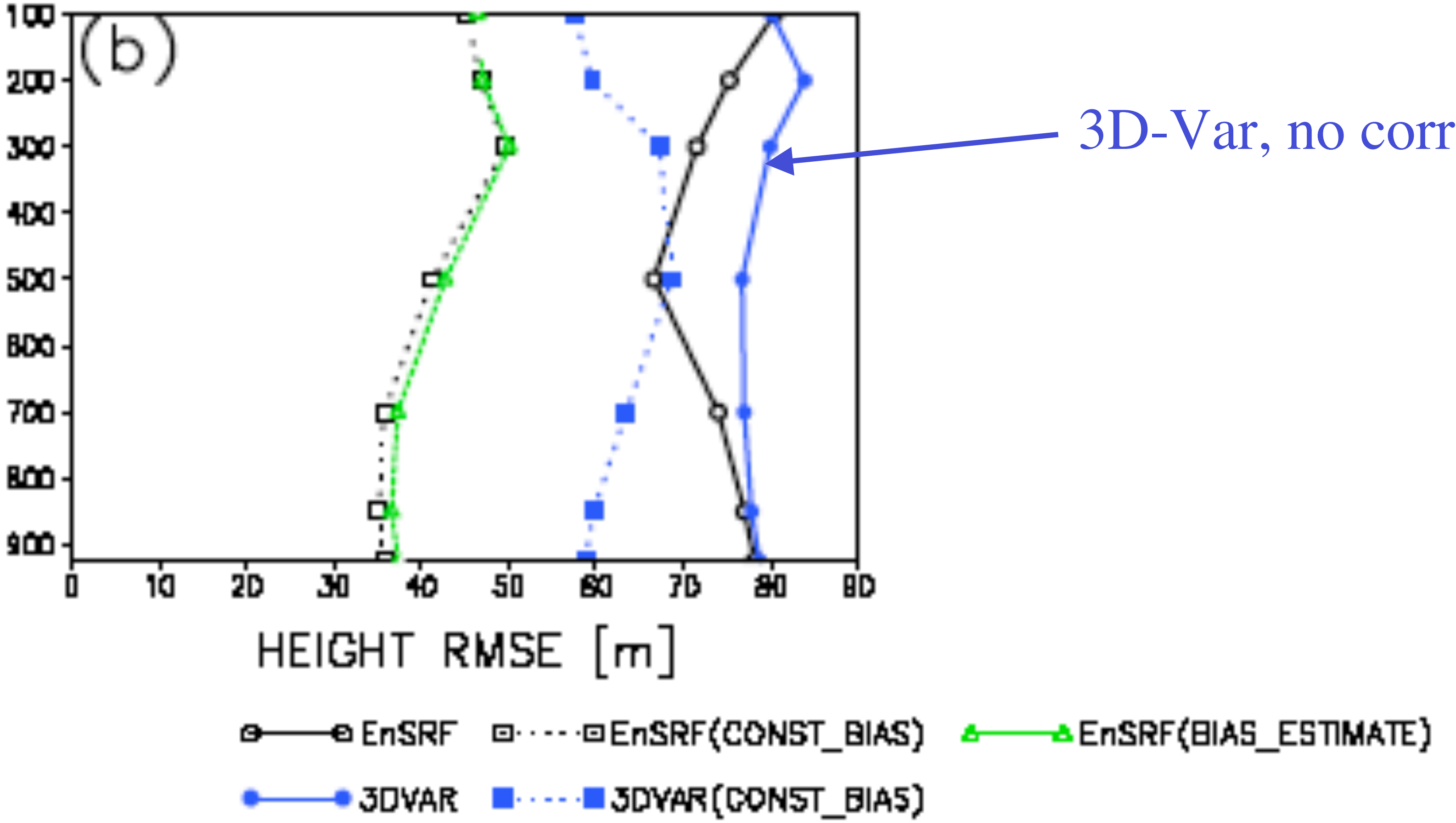
3D-Var Analysis: doesn’t know about the errors of the day

Errors of the day: they lie on a low-dim attractor

Primitive equations global models: Miyoshi (2005)

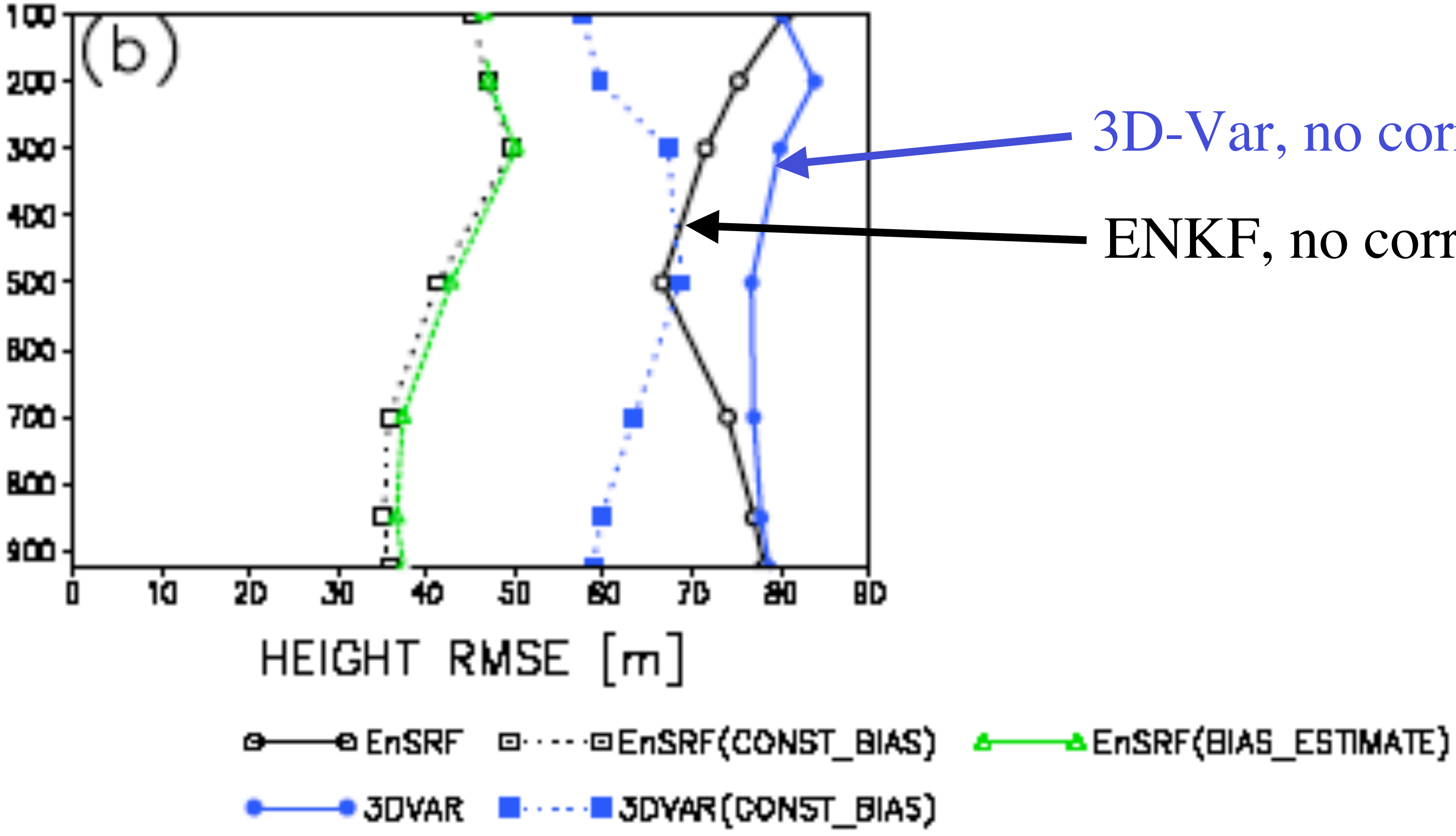
- Used the SPEEDY P.E. model of Molteni
- Both perfect model and Reanalysis observations
  - Perfect model: EnKF much better than 3D-Var
  - But, with model errors: EnKF similar to 3D-Var
  - **Model error correction in low order EOF-space:**
    - **EnKF much better corrected than 3D-Var**
  - LEKF slightly worse than EnSRF, but better if using **observation localization:**
  - A new, simple **online inflation estimation**
  - Assimilation of humidity improves results: **tracer?**

With model error, the advantage of EnKF over 3D-Var is small.



3D-Var, no correction

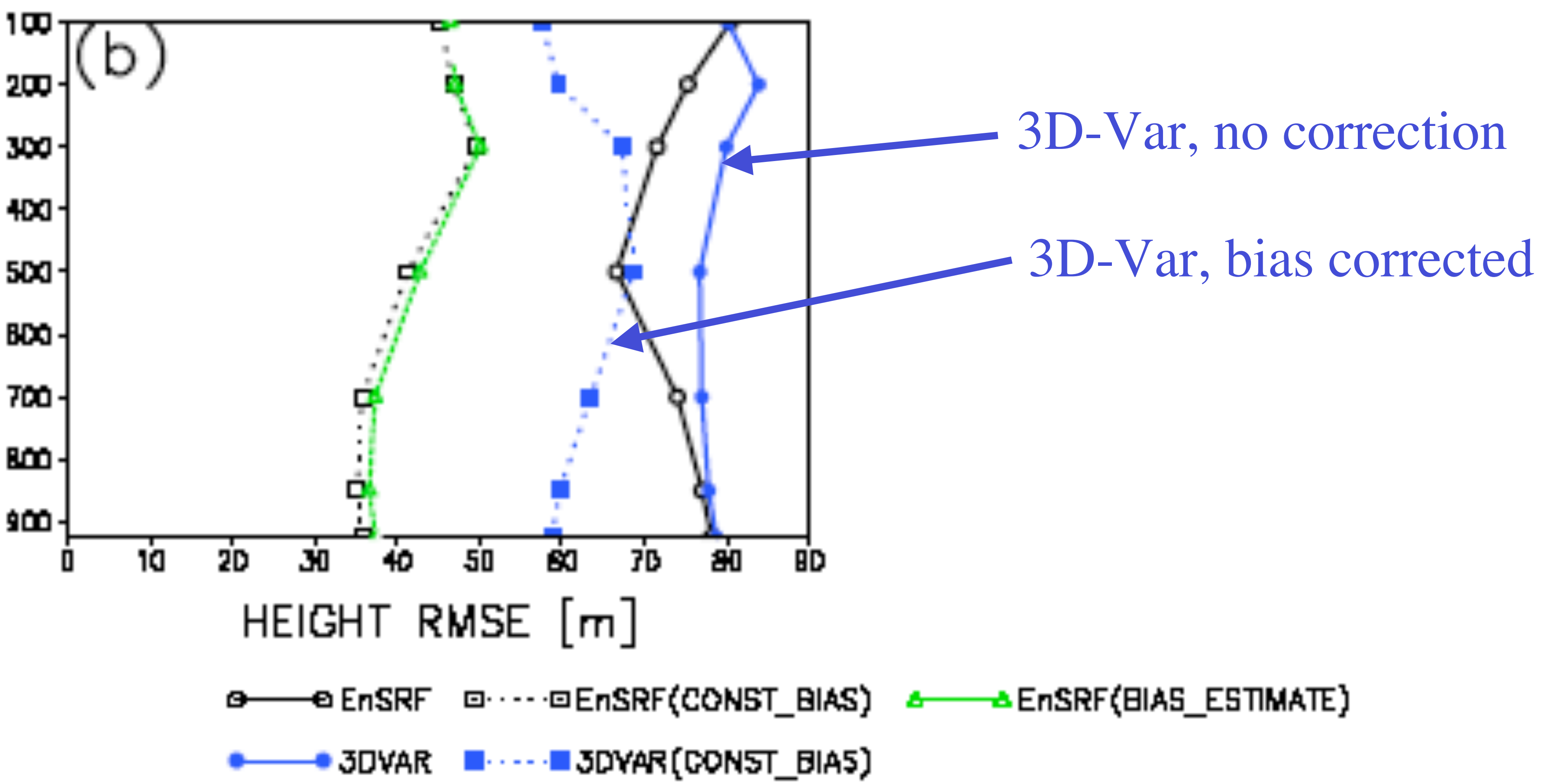
With model error, the advantage of EnKF over 3D-Var is small.



3D-Var, no correction

ENKF, no correction

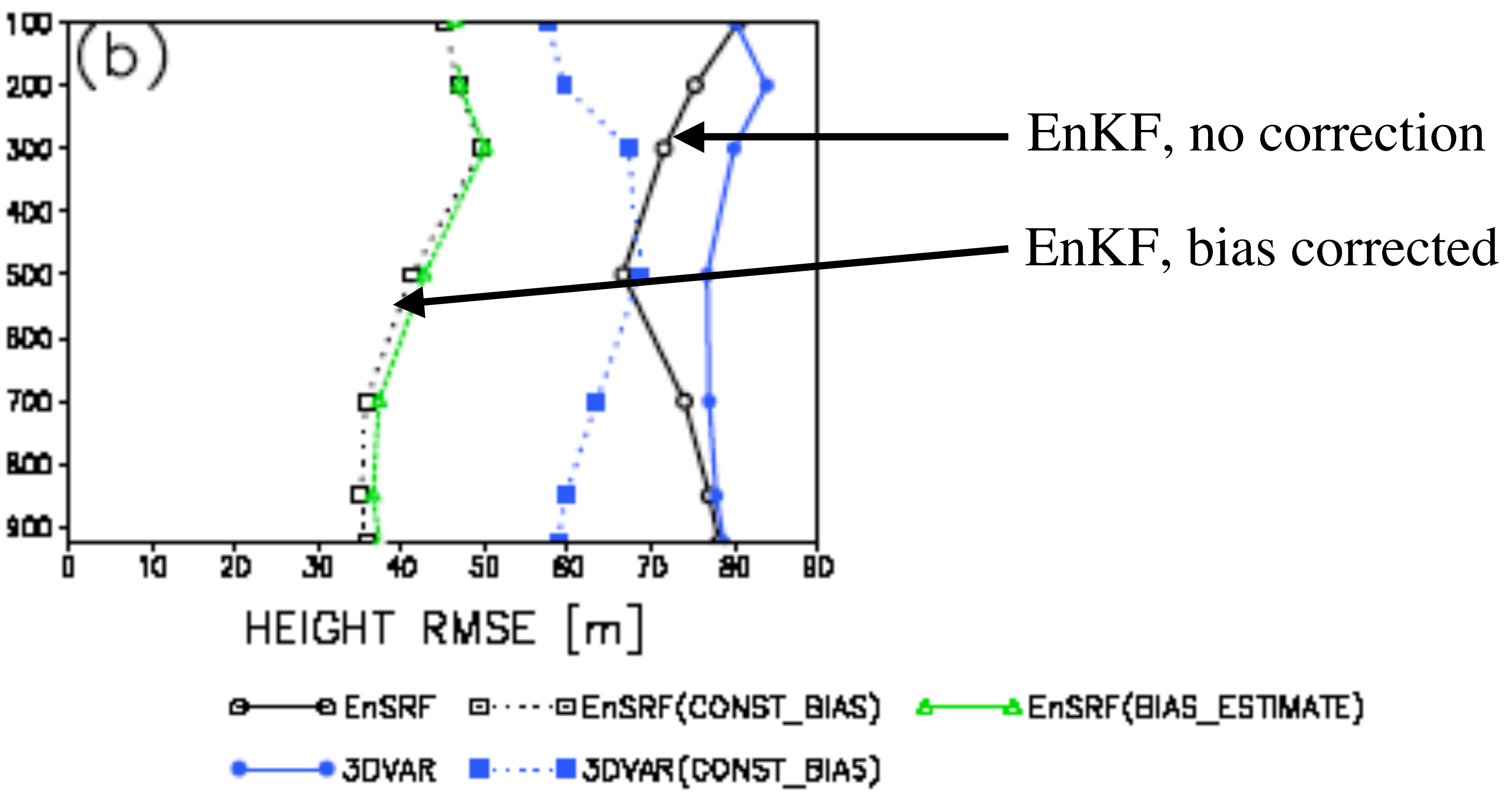
With model error correction, EnKF improves much more than 3D-Var.



3D-Var, no correction

3D-Var, bias corrected

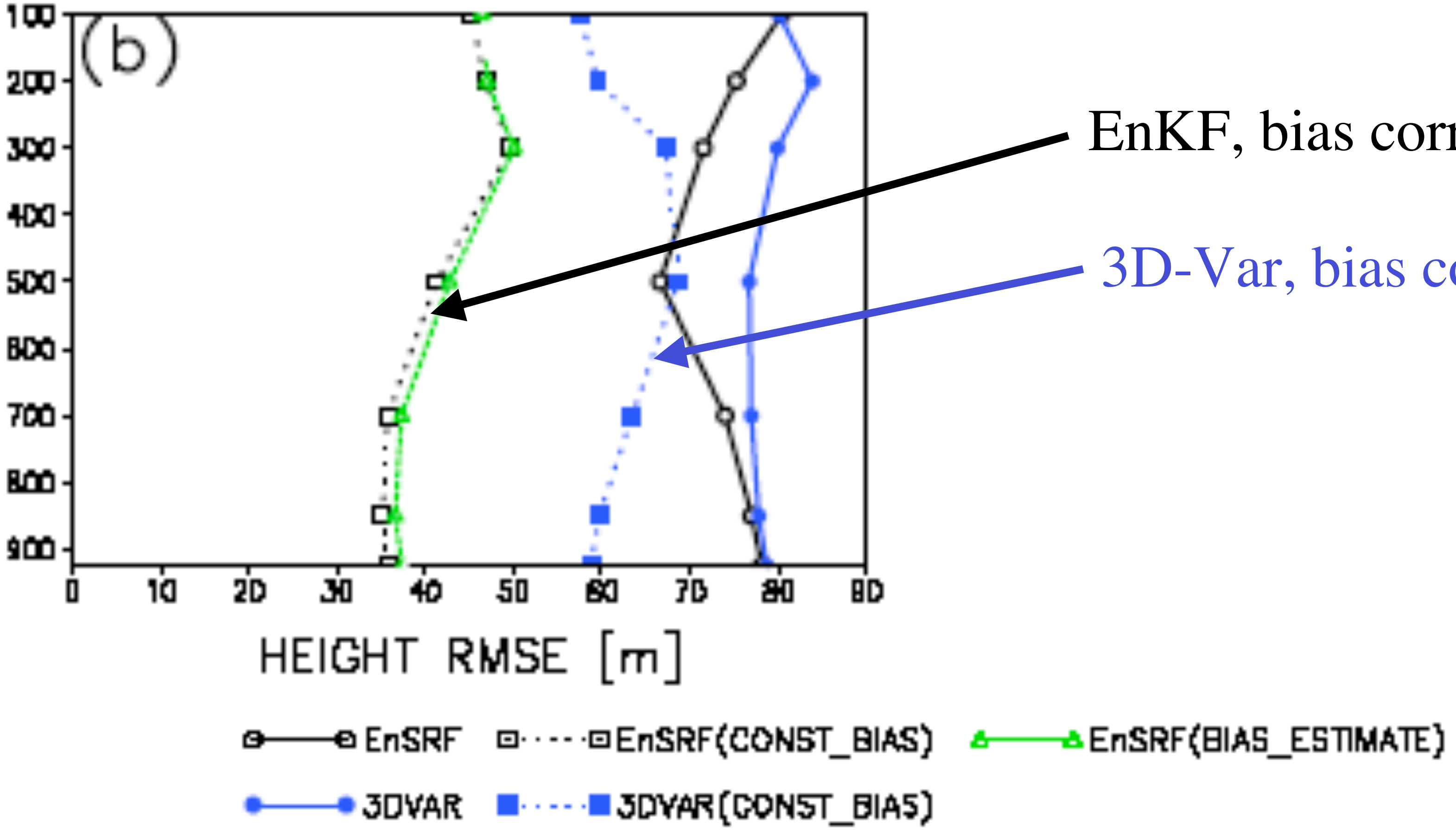
With model error correction, EnKF improves much more than 3D-Var.



EnKF, no correction

EnKF, bias corrected

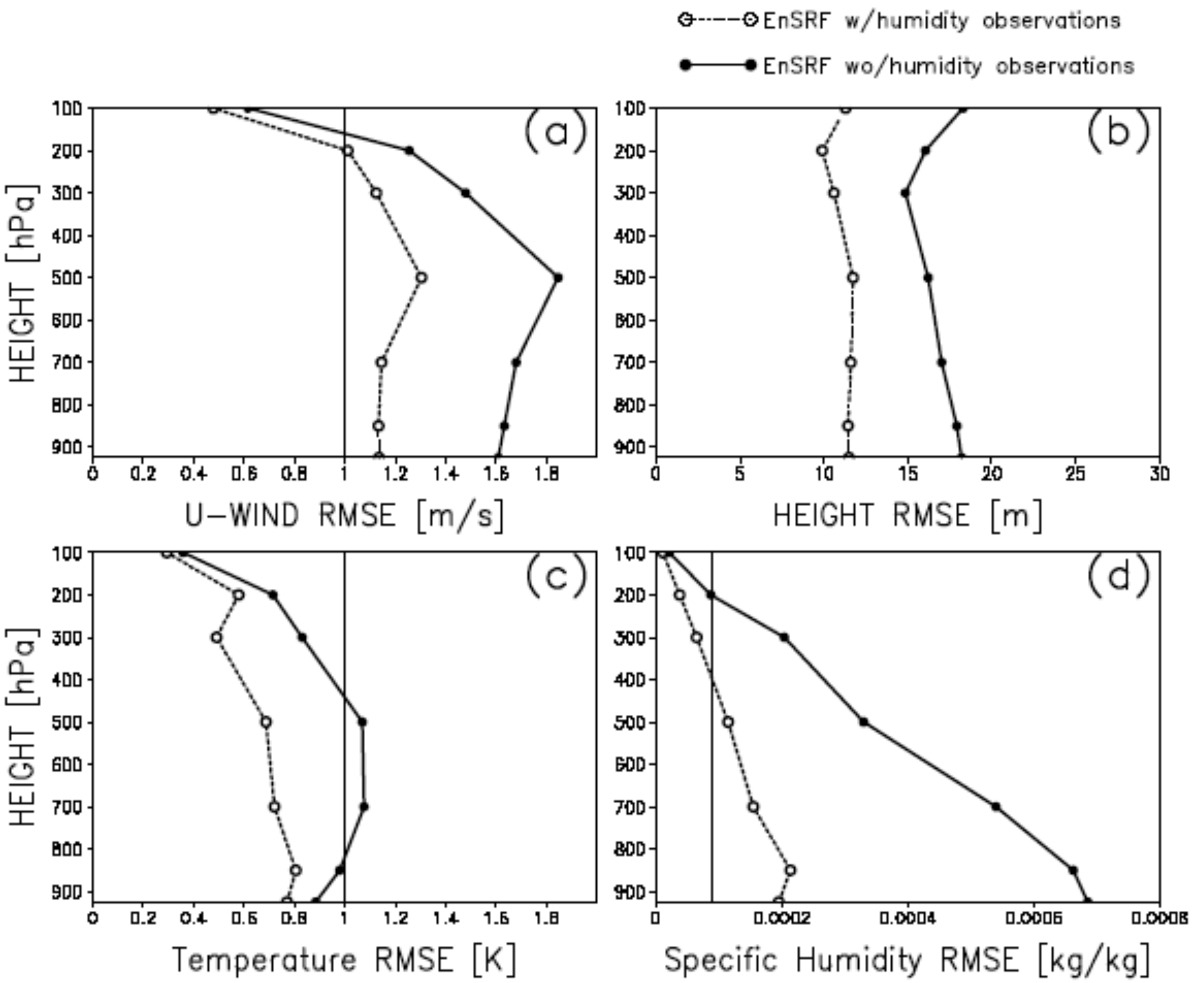
With model error correction, the advantage of EnKF over 3D-Var becomes large again.



EnKF, bias corrected

3D-Var, bias corrected

Impact of assimilating humidity observations (Miyoshi, 2005, Ph.D. thesis): Tracer info?

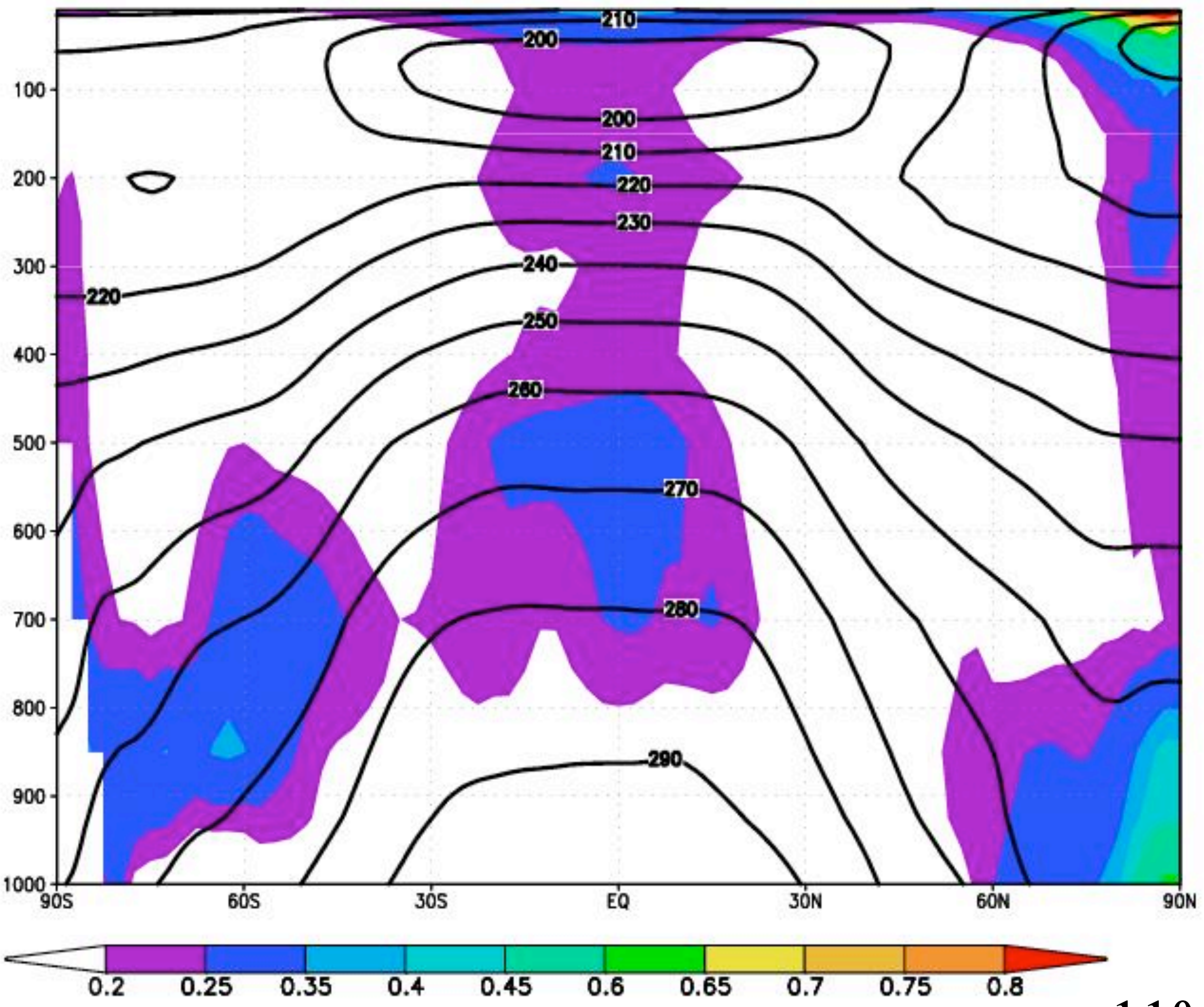


From Szunyogh, et al, 2005, Tellus

LEKF/LETKF results with NCEP's global model

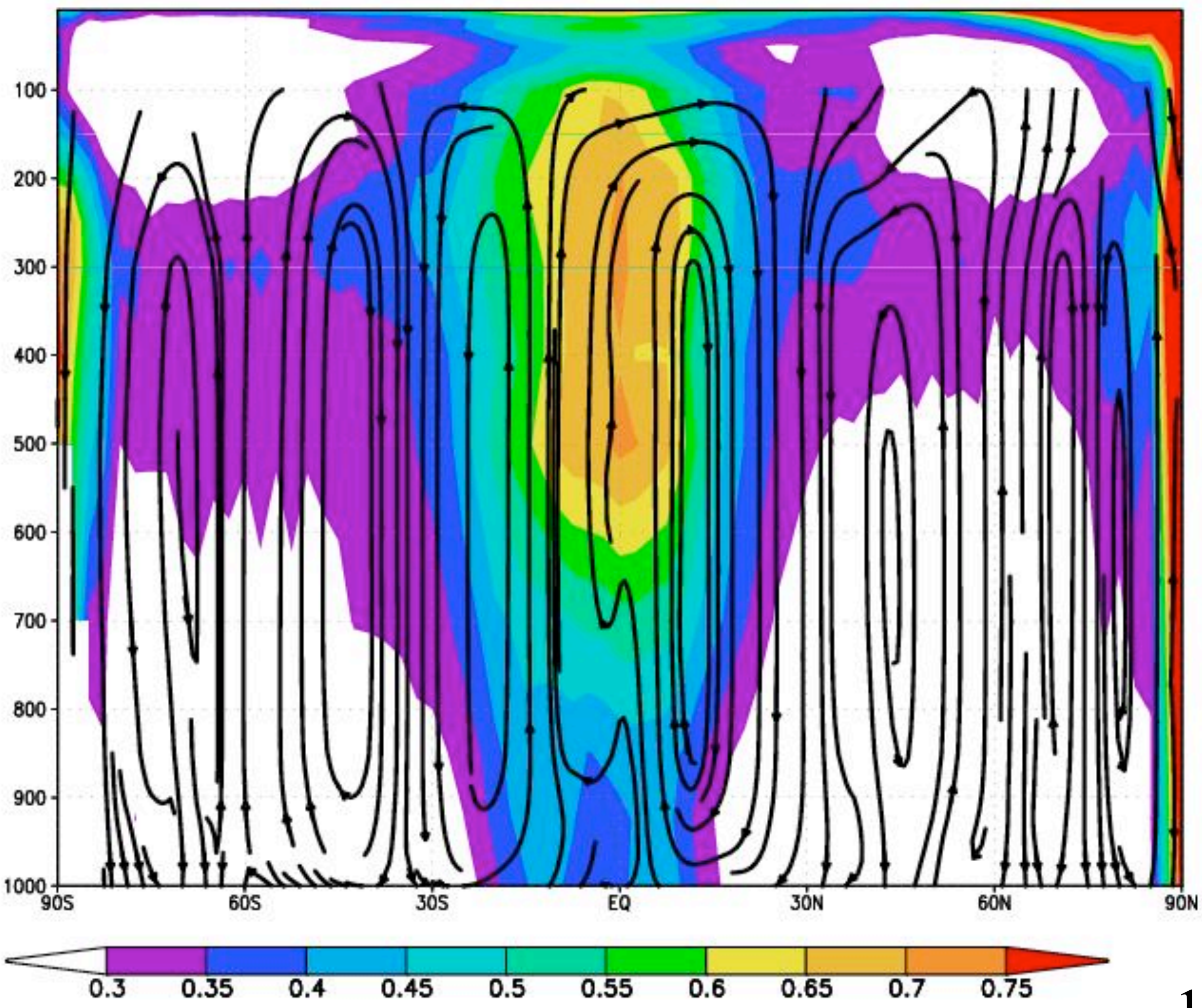
- T62, 28 levels (1.5 million d.o.f.)
- The method is model independent: the same code was used for the L40 model as for the NCEP global spectral model
- Simulation with observations at every grid point (1.5 million obs)
- Very fast! With LETKF ~3 minutes for a 40-member ensemble in a cluster of 25 PCs
- Results excellent even with low observation density

RMS temperature analysis errors



11% coverage

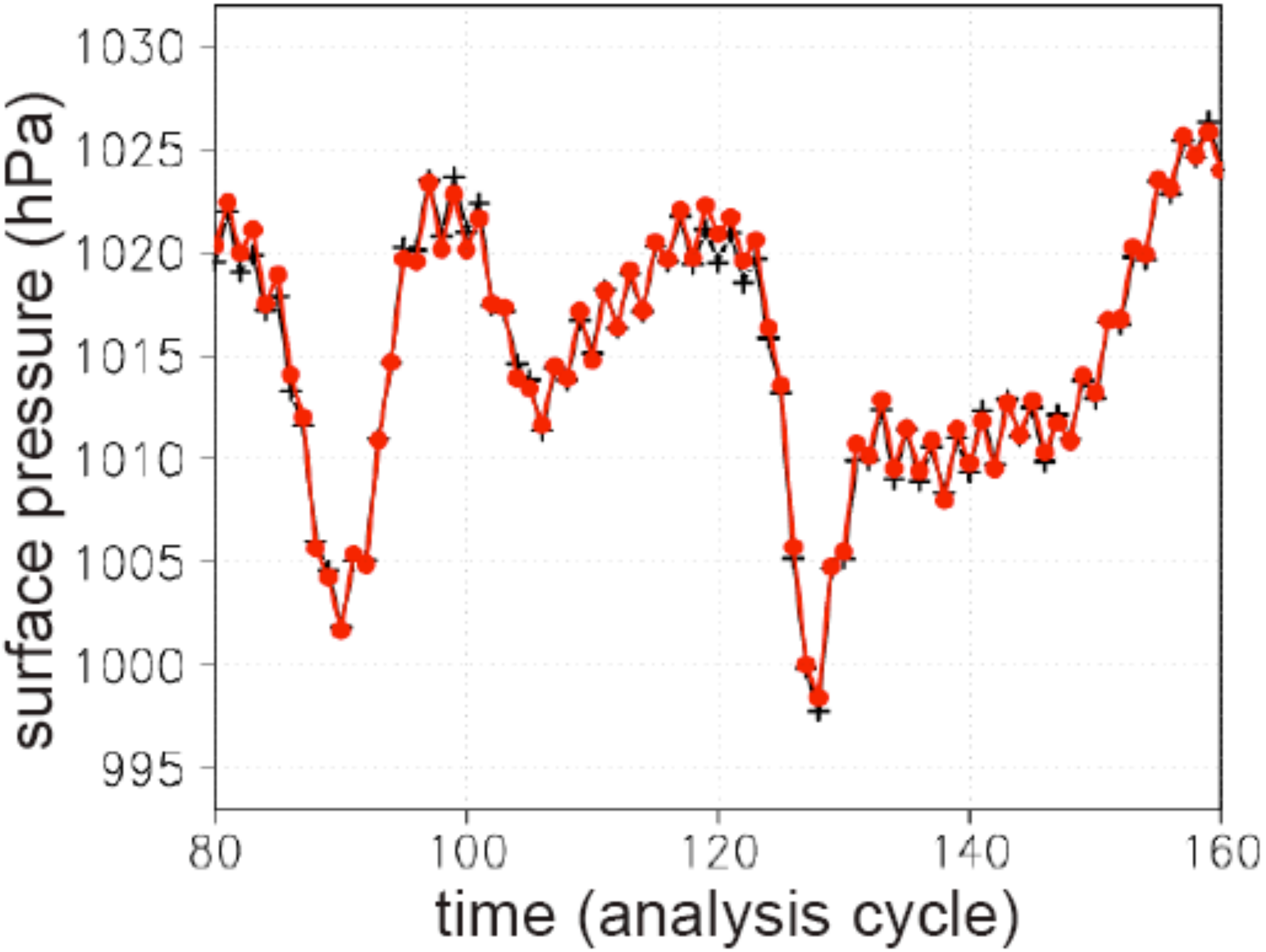
RMS zonal wind analysis errors



11% coverage

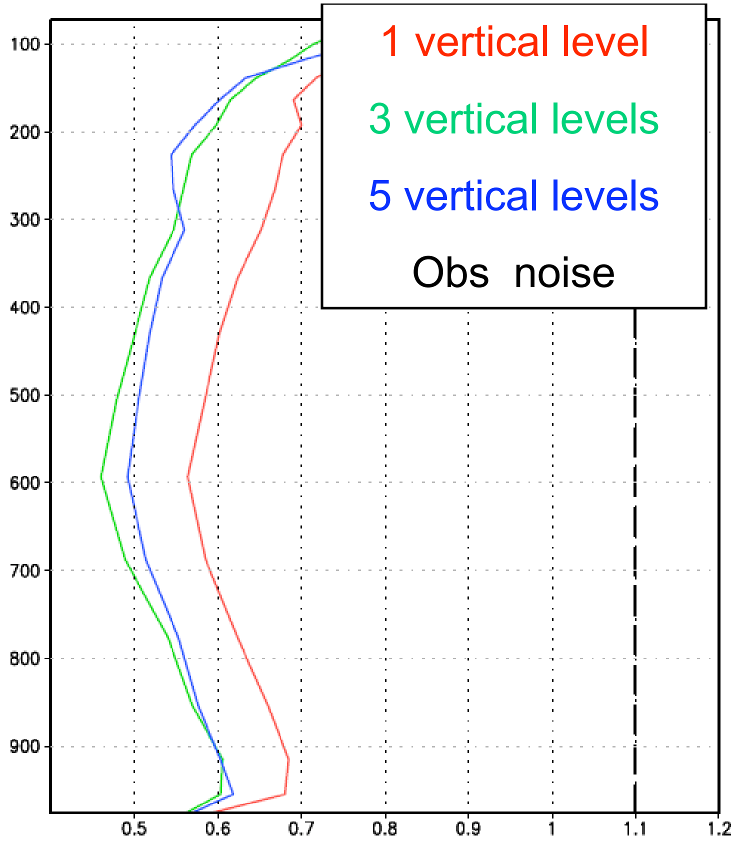
Szunyogh et al (2005)

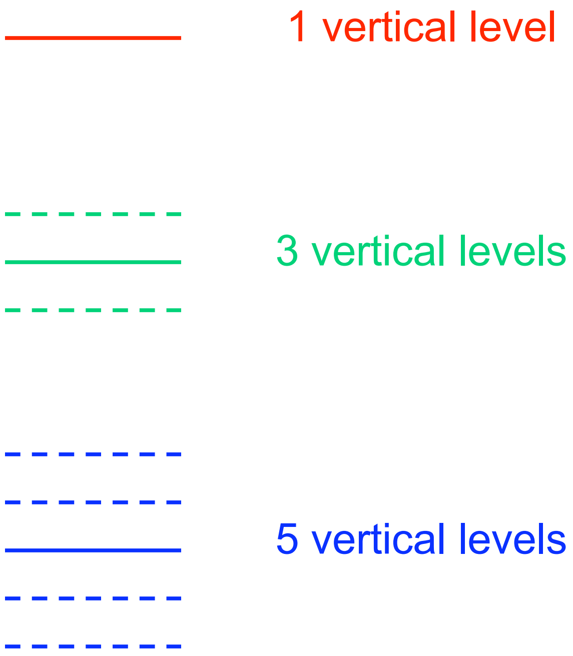
Superbalance: observed gravity wave is reproduced by analysis with just 2% observations



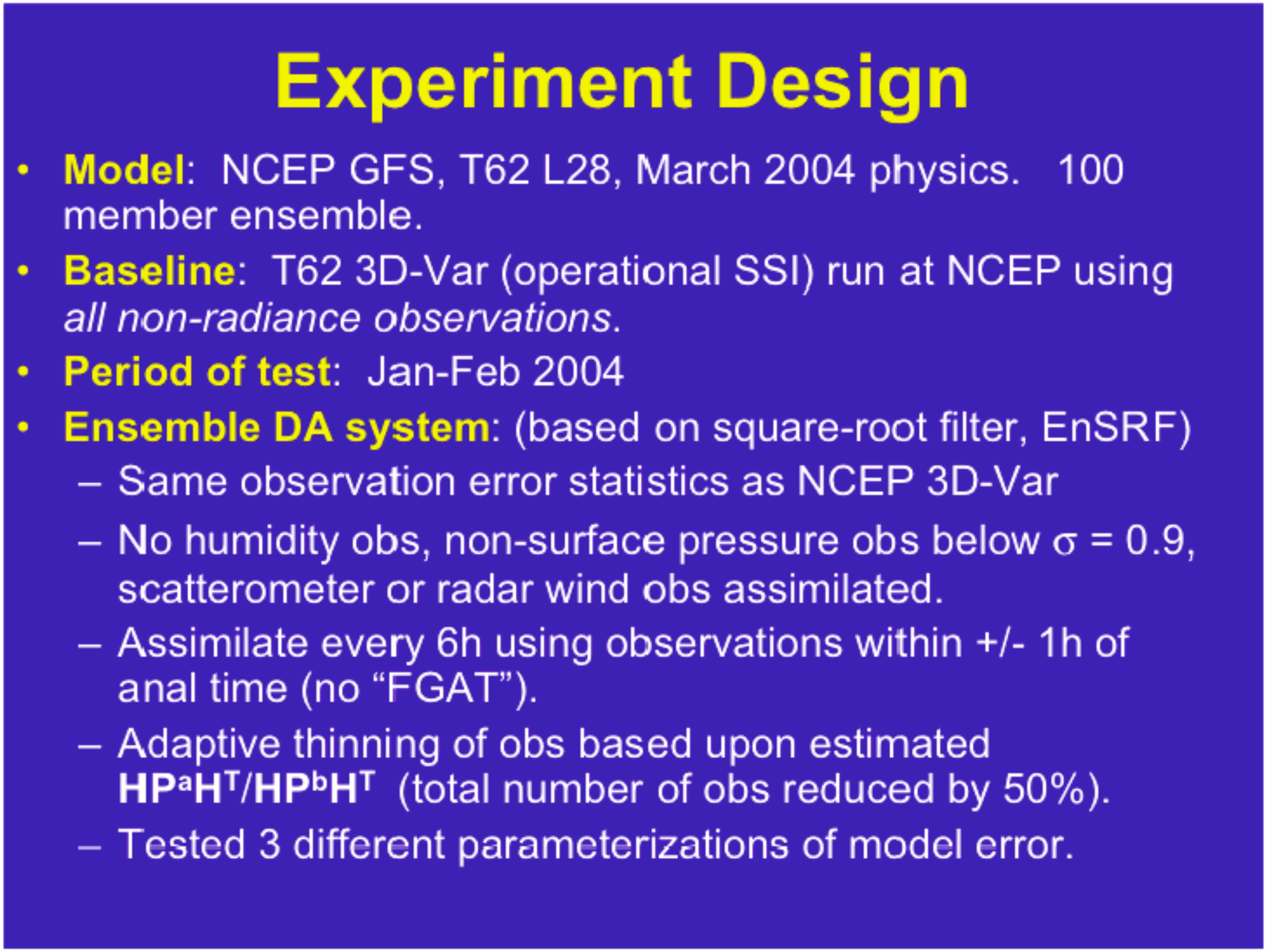
Similar good results obtained with the NASA/NOAA fvGCM (Hong Li et al, 2005)

u RMS error vs height, obs error: 1.1m/s





Whitaker and Hamill (AMS, 2005)
Serial Ensemble Square Root Filter



Experiment Design

- **Model**: NCEP GFS, T62 L28, March 2004 physics. 100 member ensemble.
- **Baseline**: T62 3D-Var (operational SSI) run at NCEP using *all non-radiance observations*.
- **Period of test**: Jan-Feb 2004
- **Ensemble DA system**: (based on square-root filter, EnSRF)
  - Same observation error statistics as NCEP 3D-Var
  - No humidity obs, non-surface pressure obs below $\sigma = 0.9$, scatterometer or radar wind obs assimilated.
  - Assimilate every 6h using observations within +/- 1h of anal time (no “FGAT”).
  - Adaptive thinning of obs based upon estimated $\mathbf{HP^aH^T/HP^bH^T}$ (total number of obs reduced by 50%).
  - Tested 3 different parameterizations of model error.

Ensemble Square-Root Filter (EnSRF; *Whitaker and Hamill, MWR ‘02*)

$$\mathbf{X}=(\mathbf{x}_1^b-\overline{\mathbf{x}^b},\ldots,\mathbf{x}_n^b-\overline{\mathbf{x}^b})$$
$$\mathbf{P}^b=\rho\circ\tfrac{1}{n-1}\mathbf{X}\mathbf{X}^{\mathrm{T}}$$

← background-error covariances estimated from ensemble, with *covariance localization*.

$$\mathbf{K}=\mathbf{P}^b\mathbf{H}^{\mathrm{T}}(\mathbf{H}\mathbf{P}^b\mathbf{H}^{\mathrm{T}}+\mathbf{R})^{-1}$$
$$\bar{\mathbf{x}}^a=\bar{\mathbf{x}}^b+\mathbf{K}\Big(\mathbf{y}-H(\bar{\mathbf{x}}^b)\Big).$$

← Mean state updated, correcting background to new observations, weighted by **K**, the Kalman gain

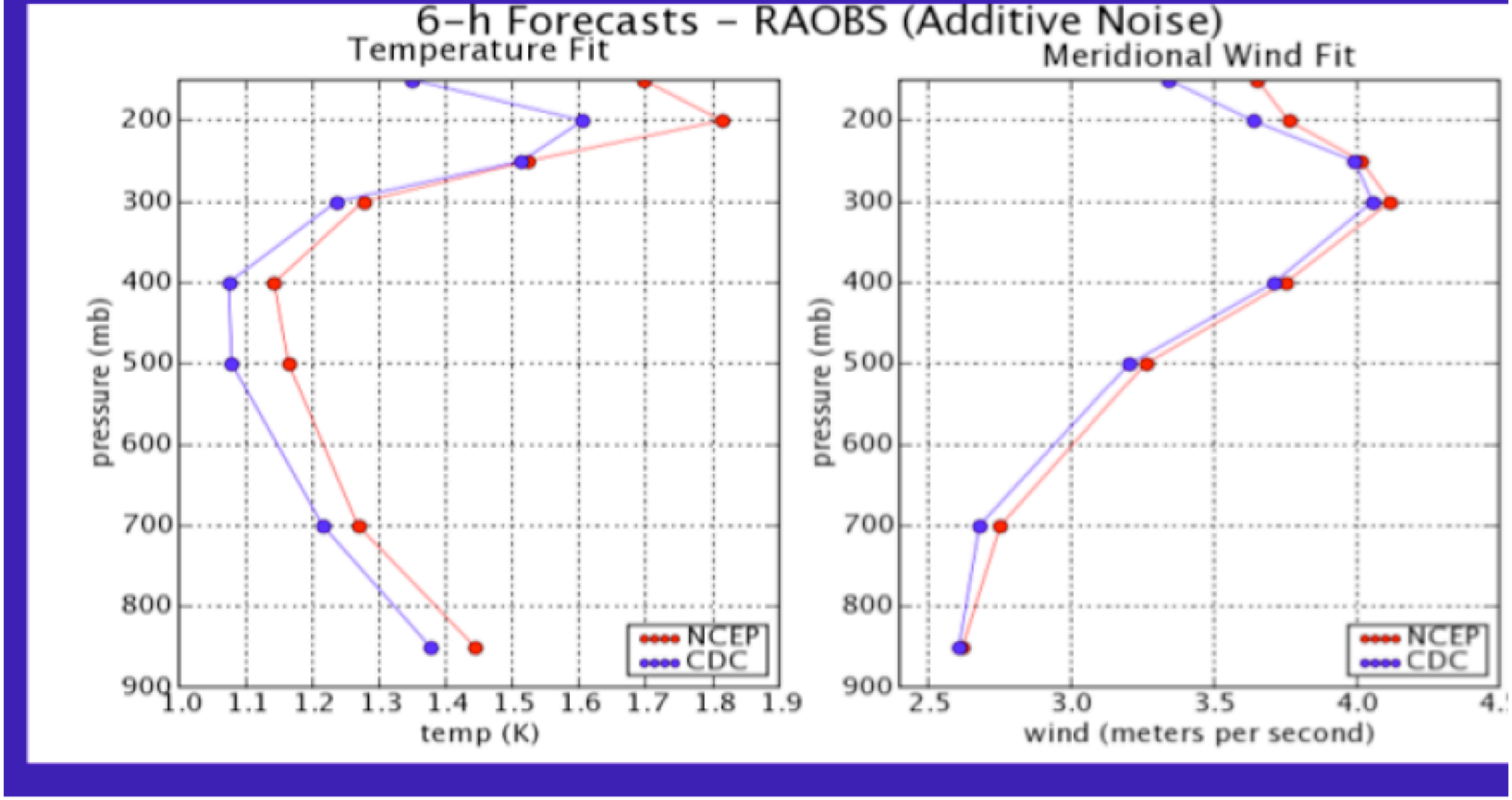
$$\widetilde{\mathbf{K}}=\left(1+\sqrt{\frac{\mathbf{R}}{\mathbf{H}\ \mathbf{P}^b\mathbf{H}^{\mathrm{T}}+\mathbf{R}}}\right)^{-1}\mathbf{K}.$$
$$\mathbf{x}_i'^a=\mathbf{x}_i'^b+\widetilde{\mathbf{K}}\Big(H(\mathbf{x}_\mathbf{i}'^\mathbf{b})\Big).$$

← “reduced” Kalman gain calculated to update perturbations around mean

$$\mathbf{x}_i^b(t+1)=M(\mathbf{x}_i^b(t))+e,\qquad e\sim N(0,\mathbf{Q})$$

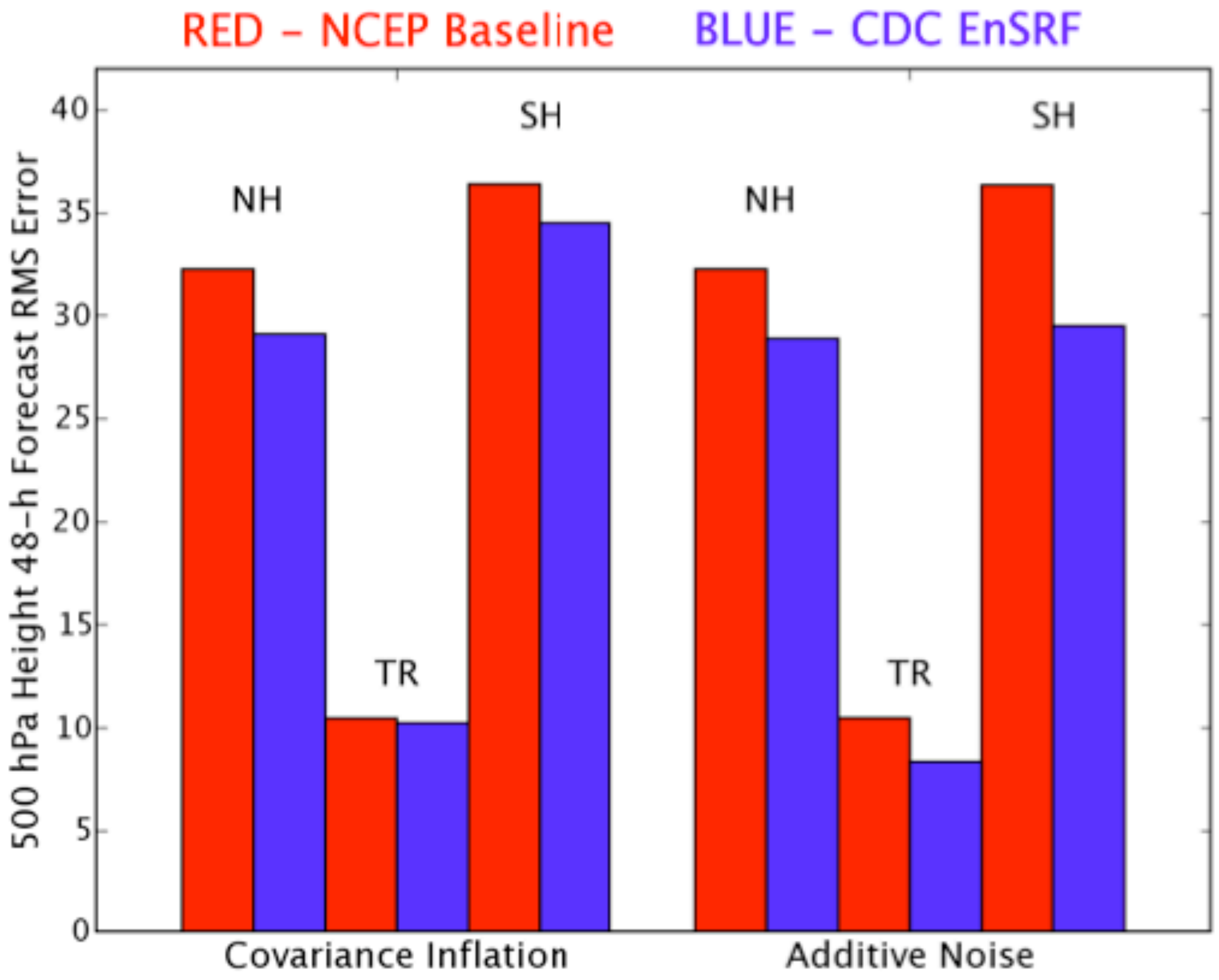
Forecast forward to the next time when data is available. Optionally, add noise or inflate to simulate model error.

Additive Noise (6-h Forecast)



Whitaker and Hamill (2005)

500 hPa Z 48-h Forecast Errors



Advantages **(**disadvantages**)** of EnKF (adapted and modified from Lorenc, 2004)

- Simple to design and code.
- Does not need a smooth forecast model.
- Does not need perturbation forecast and adjoint models.
- Generates [optimal] ensemble [perturbations].
- Complex observation operators, for example rain, coped with automatically **(**but sample is then fitted with a Gaussian**)**
- Covariances evolved indefinitely (only if represented in ensemble)
  - *Under-representation should be helped by "refreshing" the ensemble.*
- (Sampled covariance is noisy**)** and **(**can only fit N data**)**
  - *Localization reduces the problem of long-distance sampling of the "covariance of the day" and increases the ability to fit many observations.*
  - *Observation localization helps LEKF (Miyoshi)*
  - *Fast but irrelevant processes filtered automatically*
  - *Superbalance with perfect model observations*

Advantages (disadvantages) of 4D-Var

- [Can assimilate asynchronous observations]
  - *4DEnKF can also do it without the need for iterations*
- Can extract information from tracers
  - *EnKF should do it just as well*
- Nonlinear observation operators and non-Gaussian errors [can be] modeled
  - *In EnKF nonlinear observation operators (e.g., rain) are very simple*
- Incremental 4D-Var balance easy
  - *In EnKF "superbalance" is achieved without initialization*
- Accurate modeling of time-covariances (but only within the 4D-Var window)
  - *Only if the background error covariance (not provided by 4D-Var) includes the errors of the day, or if the assimilation window is long.*

Summary

- Both 4D-Var and EnKF are better than 3D-Var
- 4D-Var with long windows is competitive with EKF
- 4D-EnKF can assimilate asynchronous obs as 4D-Var
  - EnKF does not require adjoint of the NWP model (or the observation operator)
  - Free 6 hr forecasts in an ensemble operational system $\sum \delta \mathbf{x}_i^a \delta \mathbf{x}_i^{aT} = \mathbf{A}$
  - Provides optimal initial ensemble perturbations
  - LETKF: Assimilates all obs. in a local cube. 4 minutes for 40 members, 2M obs in a cluster of 25 PCs: fast enough for operations.
- Model errors can be handled with a low order approach (Miyoshi, Danforth). EnKF improves more than 3D-Var
- The results of Whitaker and Hamill (2005) show for the first time a clear advantage over 3D-Var with real observations