Development of a Data Assimilation System to Estimate the State of Large Spatio-Temporally Chaotic Systems

Istvan Szunyogh

Institute for Physical Science and Technology

& Department of Atmospheric and Oceanic Science

CSCAMM Seminar October 17, 2007

Outline

- The data assimilation problem
- The **L**ocal **E**nsemble **T**ransform **K**alman **F**ilter
- Some comments on predictability
- The future

The System

Given are

- A large physical system (e.g., terrestrial atmosphere, ocean, planetary atmosphere, laboratory fluid system)
- A numerical model based on the spatial and temporal discretization of the partial differential equations that serve as the mathematical model of the system
- Noisy observations of the system

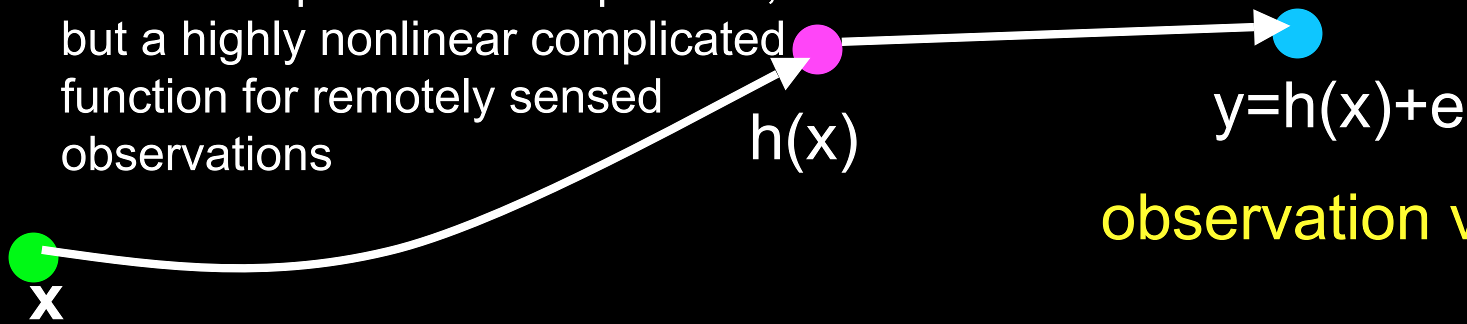
Observations

h: observation operator

often a simple linear interpolation, but a highly nonlinear complicated function for remotely sensed observations

e: observation error

Normally distributed with a “known” covariance matrix R



observation vector

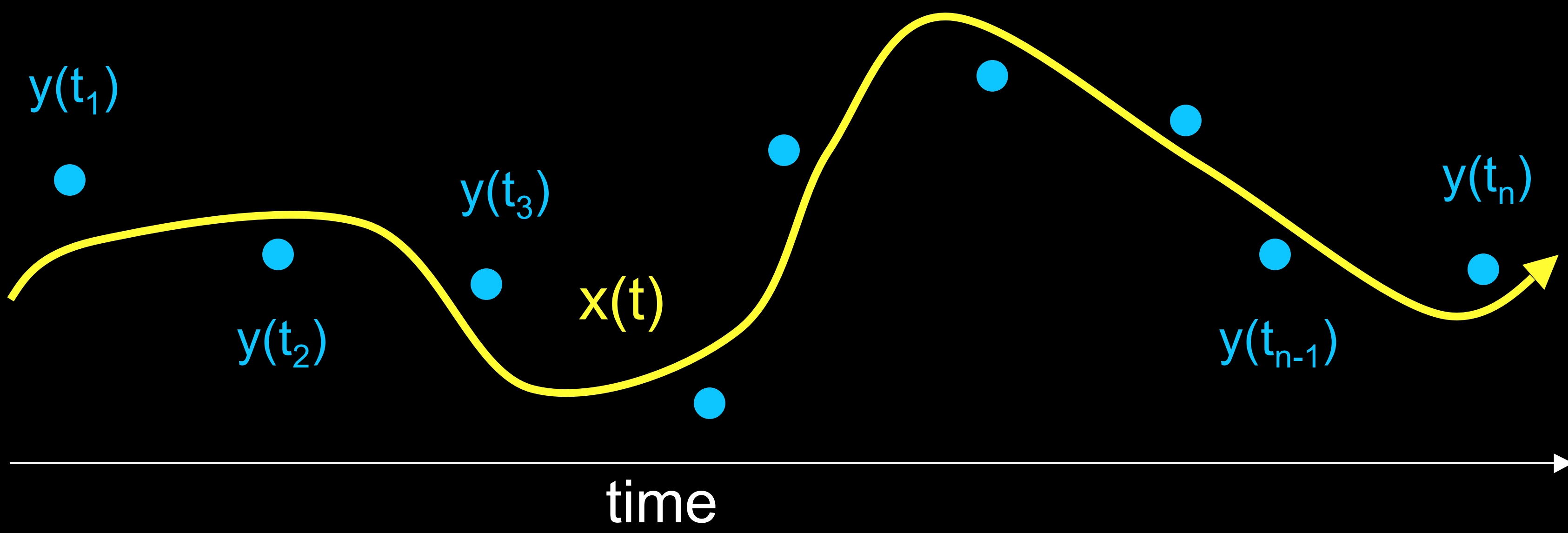
State vector

The components are the different state variables (e.g. components of the velocity vector, temperature, surface pressure, humidity concentration, etc.) at the model grid points

The Goal

To find the model trajectory x(t) that best fits a time series of observation vectors: $y(t_1), y(t_2), y(t_3), \ldots, y(t_n)$

The state estimate (analysis) at a given time t_n is $x(t_n)$



The Least-Square Problem

The likelihood of a trajectory x(t) is proportional to

$$\prod_{j=1}^{n} \exp\left(-\frac{1}{2}[\mathbf{y}_j - H_j(\mathbf{x}(t_j))]^T \mathbf{R}_j^{-1} [\mathbf{y}_j - H_j(\mathbf{x}(t_j))]\right)$$

The most likely trajectory is the one that minimizes

$$J(\{\mathbf{x}(t)\}) = \sum_{j=1}^{n} [\mathbf{y}_j - H_j(\mathbf{x}(t_j))]^T \mathbf{R}_j^{-1} [\mathbf{y}_j - H_j(\mathbf{x}(t_j))]$$

Sequential approach:

$$\begin{aligned} J[\mathbf{x}(t_n)] &= \\ & [\mathbf{x} - \mathbf{x}_n^b]^T (\mathbf{P}_n^b)^{-1} [\mathbf{x} - \mathbf{x}_n^b] \\ +& [\mathbf{y}_n - \mathbf{H}_n \mathbf{x}]^T \mathbf{R}_n^{-1} [\mathbf{y}_n - \mathbf{H}_n \mathbf{x}] \end{aligned}$$

Sequential Estimation of the State

- The background x^b is a short-term model forecast from the analysis at the previous time.
  - It reflects the combined effect of all past observations, filling up gaps in the observing network
  - Model dynamics do the filtering and build realistic dynamical "balance" between the observed and unobserved variables

An Example for the Dimensionality of the Problem:

Current global circulation model of NCEP/NOAA

- Dimension of the state vector: about 385 million
- Number of assimilated observations: 7-8 million observations per day (about two orders of magnitude less then the # of variables)
- Number of observations received: 1.43 billion observations per day not all assimilated due to (i) time constraint (total time available for data processing and analysis is 35 minutes), (ii) quality problems, (iii) lack of observation operator, (iv) redundancy
- Most of the observations are remotely sensed

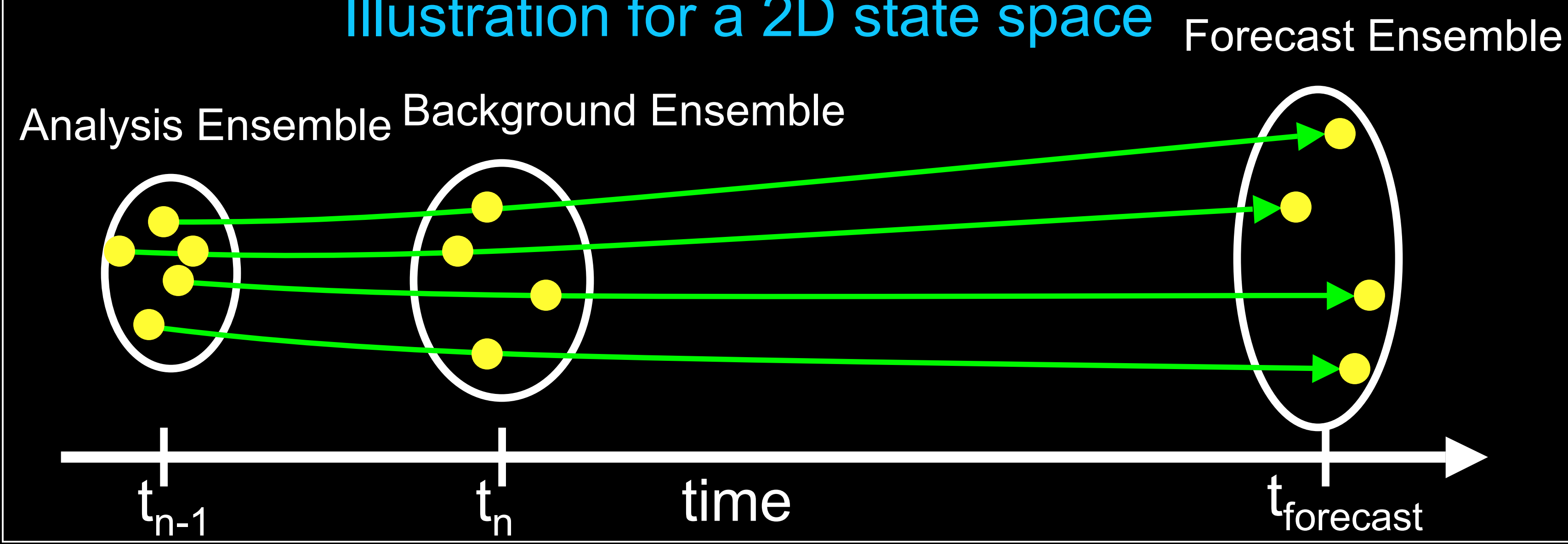
Extended Kalman Filter:
the four main components at time t_n

1. $\mathbf{x}^b = \mathcal{M}\mathbf{x}^a(t_{n-1})$: Obtaining the backgound
   - $\mathcal{M}$: Nonlinear model from time t_{n-1} to t_n
2. $\mathbf{P}^b = M\mathbf{P}^a(t_{n-1})M^T$: Obtaining the background error covariance matrix
   - M: Linearization of $\mathcal{M}$ around $\mathbf{x}^a(t_{n-1})$
   - Prohibitively expensive computationally
   - Issues of linearization
3. $\mathbf{x}^a = \mathbf{x}^b + \mathbf{K}[\mathbf{h}(\mathbf{x}^b) - \mathbf{y}]$: Update Equation
   - $\mathbf{K} = \mathbf{P}^b\mathbf{H}^T(\mathbf{H}\mathbf{P}^b\mathbf{H}^T + \mathbf{R})^{-1}$: Kalman Gain Matrix
   - H: $\mathbf{h}(\mathbf{x}_b)$ linearized around $\mathbf{x}_b$
4. $\mathbf{P}^a = (\mathbf{I} - \mathbf{KH})\mathbf{P}^b$: Analysis Error Covariance Matrix

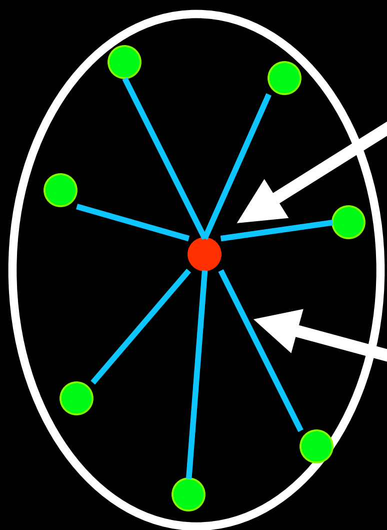
Ensemble-based Estimation of the forecast uncertainties

- The model state is considered to be a probabilistic variable: The probability distribution is evolved by a representative ensemble of model states

Illustration for a 2D state space



Ensemble Representation of the Background



The ensemble mean is the background

The background error covariance matrix is defined by the ensemble of background perturbations

Background Ensemble

Ensemble-based Kalman Filter

data assimilation schemes

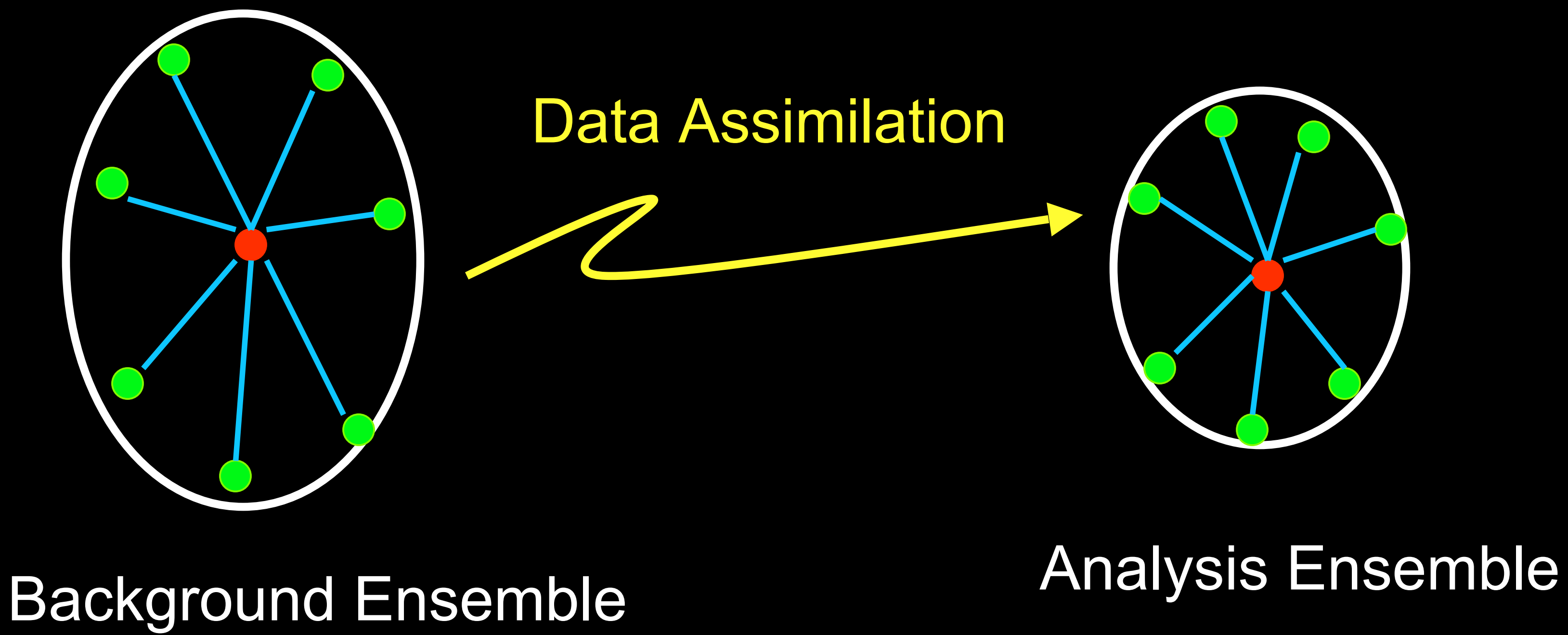
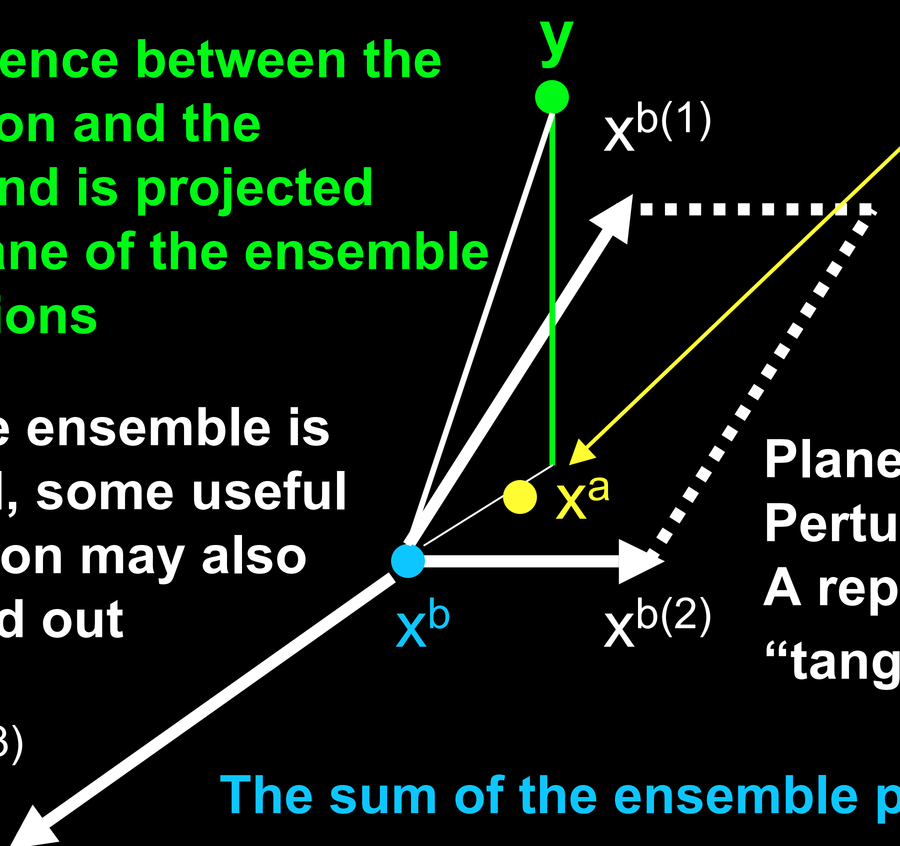


Illustration in State Space

3d state space, 3-member ensemble

The difference between the observation and the background is projected on the plane of the ensemble perturbations



x^b-x^a is obtained in the plane of the ensemble perturbations: potentially an efficient filter of observational noise

When the ensemble is too small, some useful information may also be filtered out

Plane of the ensemble Perturbations:
A representation of the "tangent space" at x^b

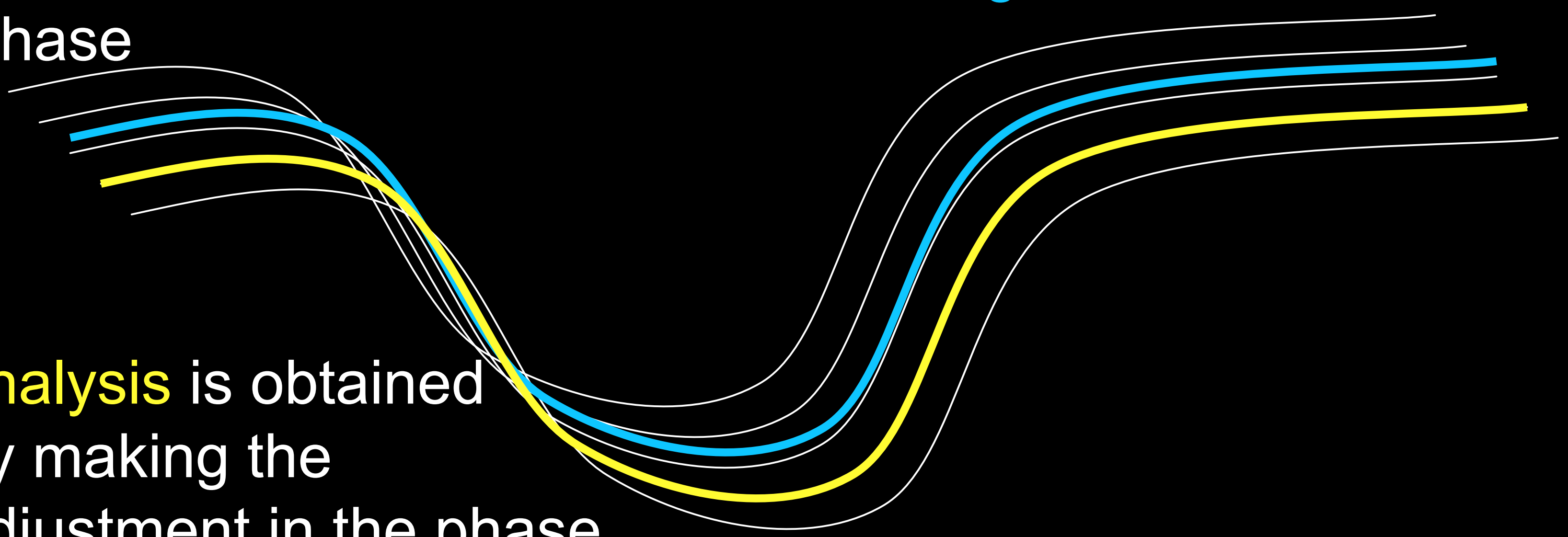
The sum of the ensemble perturbations is zero

Illustration in physical space

uncertainty in the phasing of a wave

background ensemble
indicates uncertainty in the
phase

background

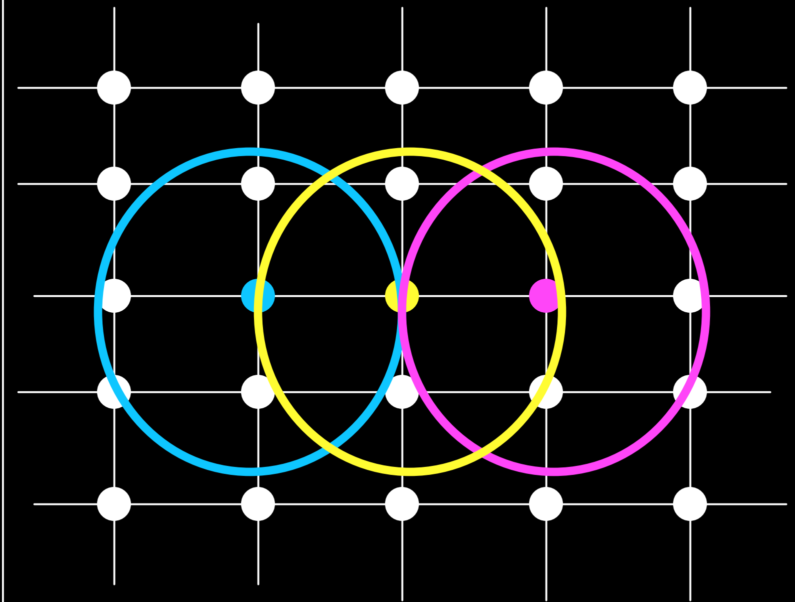


analysis is obtained
by making the
adjustment in the phase
based on the observations

Motivations for our approach of the development were

- In 2001, it was yet to be seen whether an ensemble-based Kalman filter coupled with a state-of-the-art forecast model can be used to assimilate observations of the real atmosphere. The major concerns were
  - An estimate of the background error covariance matrix based on a reasonably small ensemble would be hopelessly rank-deficient
  - An ensemble-based Kalman filter would be computationally hopelessly expensive
  - Some scientists also argued that model errors were hopelessly large for an indefinitely long cycling of an ensemble base Kalman filter
- Our goal was to design a scheme to address these concerns and a series of experiments to separate real challenges from assumed difficulties
- We wanted to design a scheme for parallel computers

Illustration of the Local Approach
for a 2D model grid



Local state vector: components of the global state vector in the local region

• A local region is assigned to each grid point
• The state at the center point is estimated based on information from the local region

• This approach provides a high-rank estimate of P^b with a small ensemble
• Since the grid points can be processed independently, they can be processed in parallel

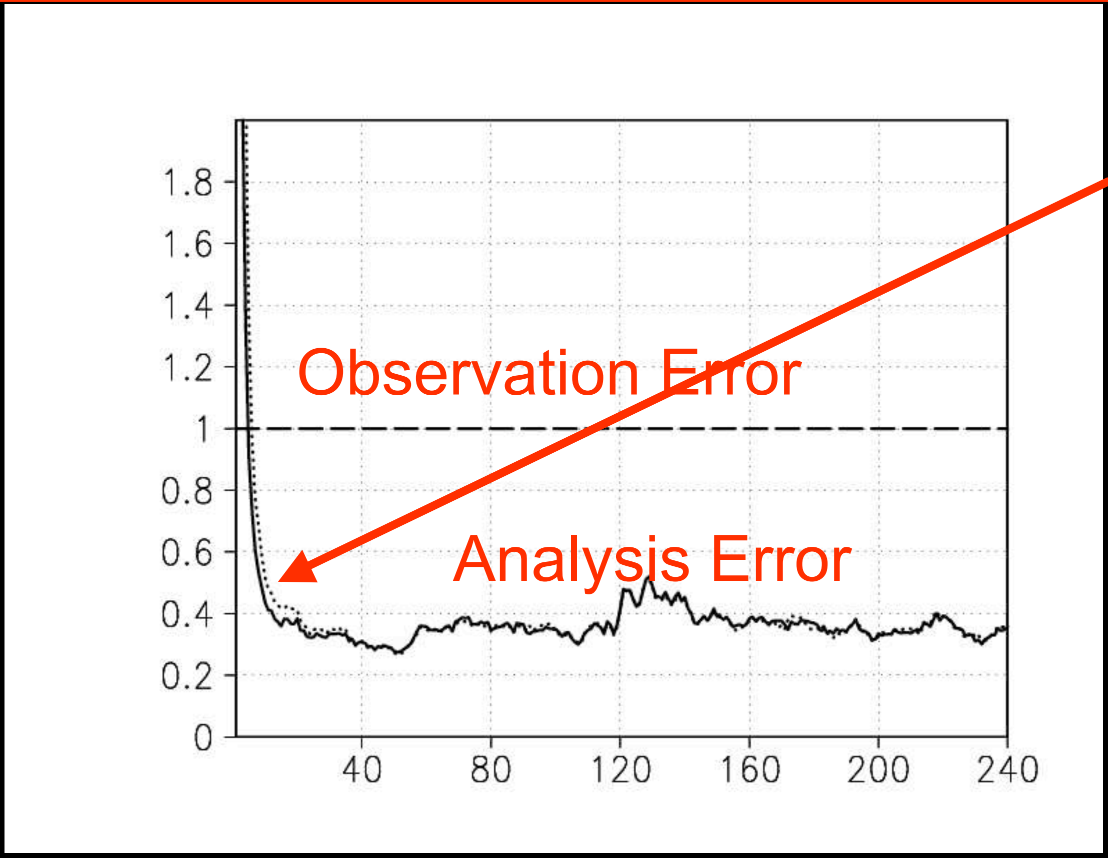
Local Ensemble Kalman Filter (LEKF)

- First Formulation: Ott, Hunt, Szunyogh et al. 2004, *Tellus* **A**
  - Investigated the conditions under which the local approach provided a smooth global analysis
  - Scheme was tested on the Lorenz-96 model (40-120 variables)
- First experiments with the NCEP GFS were designed to address the following issue
  - Is it possible to track the state of the model with a small (40-80-member) ensemble under the perfect model scenario? A necessary condition for an ensemble-based Kalman filter to work
  - Results were reported in Szunyogh, Kostelich, Gyarmati et al., 2005, *Tellus* ***A***

Experimental design
of Szunyogh et al. 2005

- Observations: Noisy observations of a time series of true states (generated by a long model integration), full vertical soundings are located at randomly selected model grid point location (10% coverage for the results shown here, but the scheme is still stable at 2.5% coverage)
- Data Assimilation: LETKF with 40 ensemble members
- Model: NCEP GFS at resolution T62 (about 150 km) and 28-levels
- Error Statistic collected for 45 days (January-February)

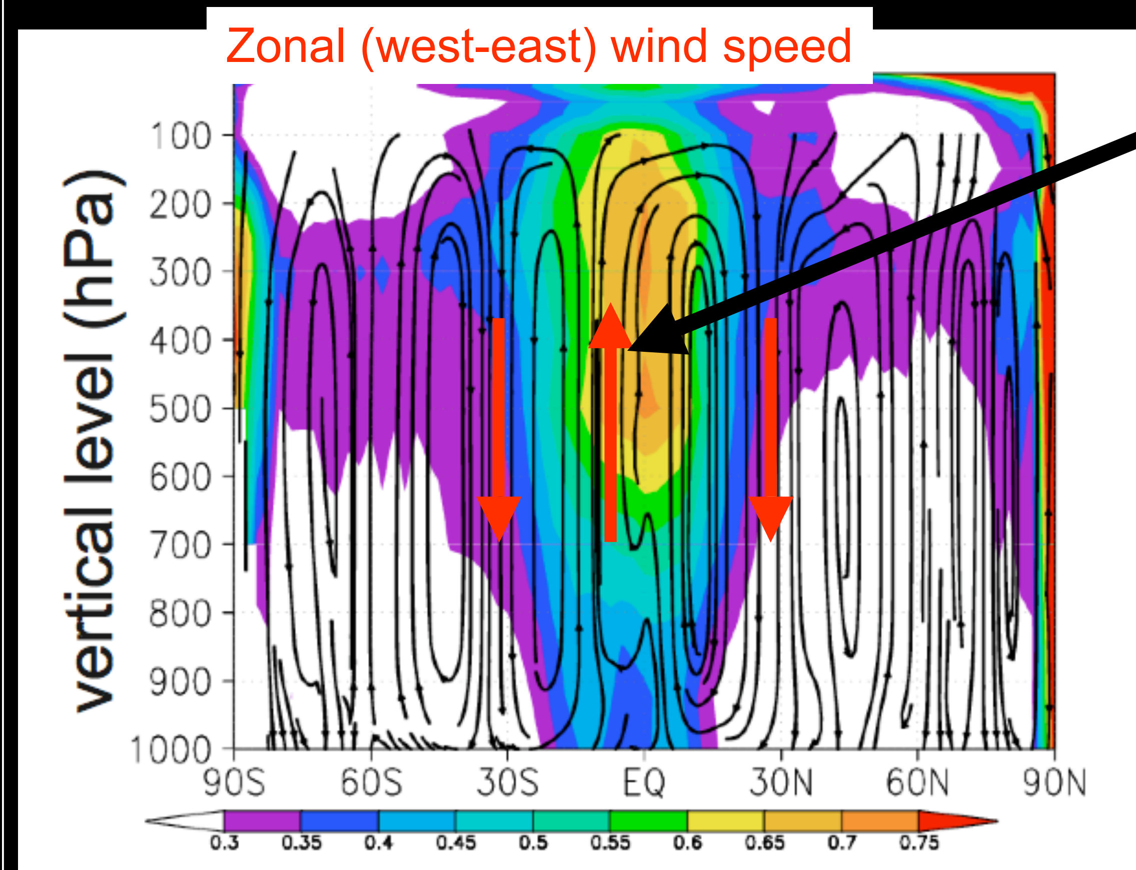
Time Evolution of RMS Error in Surface Pressure Analysis



- The analysis error Settles in a few steps
- The analysis error is much smaller than the observation error
- The results are similar for the other model variables

Vertical Distribution of RMS Error

averaged over time and along latitudes



The error is the largest in the region of upward motions in the Tropics (parameterized deep convection)

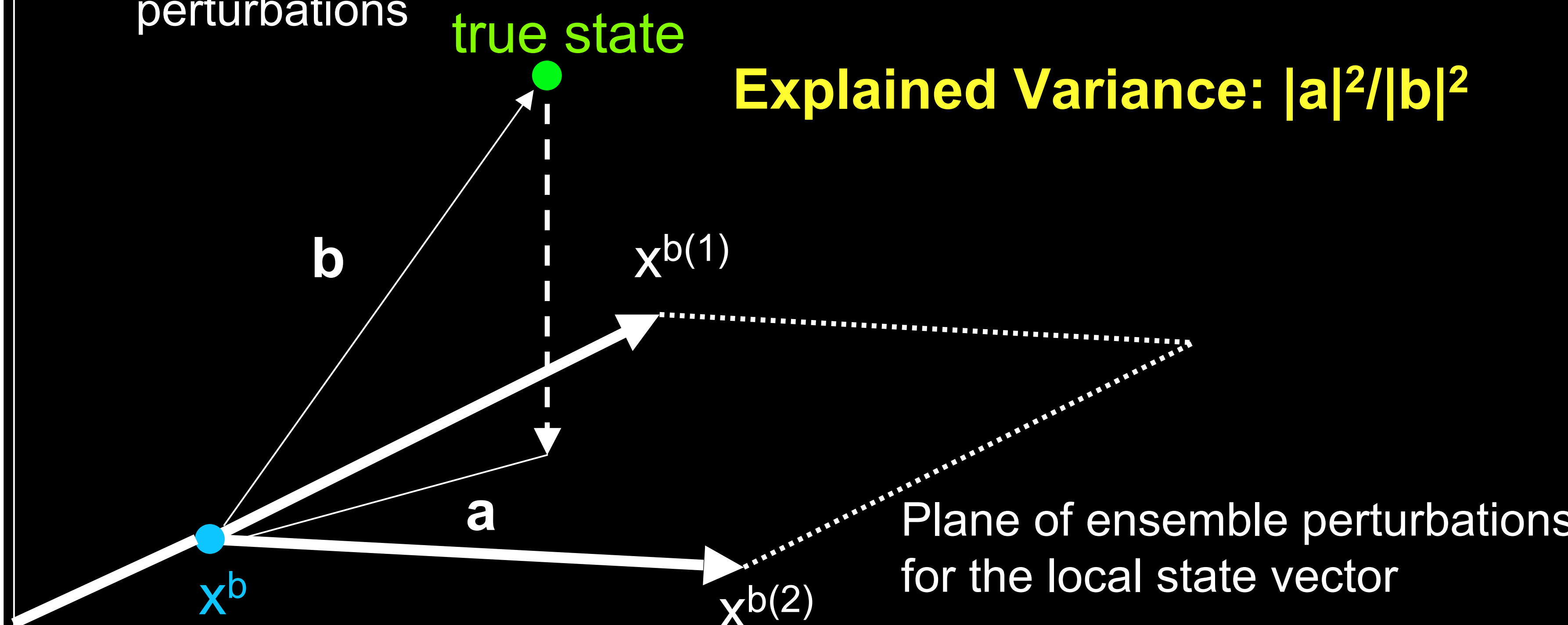
Reminder: the model is perfect, observation coverage homogeneous!!!

Differences are due to differences in the dynamics

Why such large differences?

Explained Variance: a measure of ensemble performance in the local region

- b: True error
- a: Projection of the true error on the space of the ensemble perturbations



E-dimension: a measure of complexity in the local region

- E-dimension: A measure of the steepness of the spectrum of the ensemble-based error covariance matrix in the local region
- The smaller the E-dimension the steeper the spectrum (introduced in Patil et al. 2001, *PRL*; discussed in details an illustrated on complex meteorological examples in Oczkowski, Szunyogh, and Patil, 2005, *JAS*)

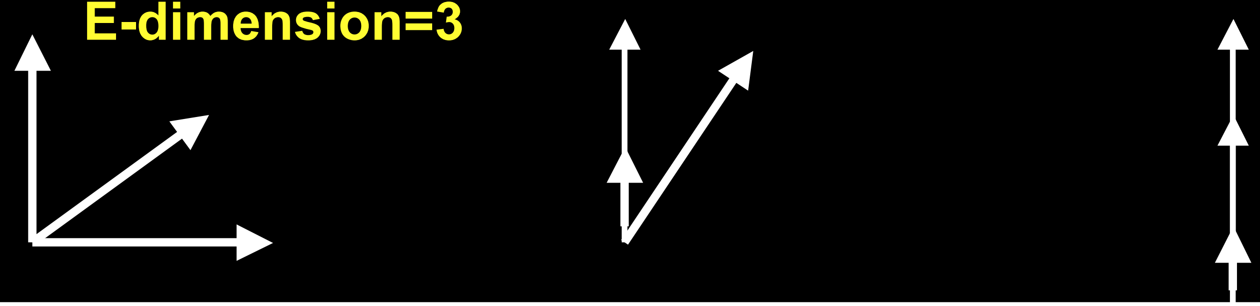
Three orthogonal perturbations

All three perturbations in one plane

E-dimension=3

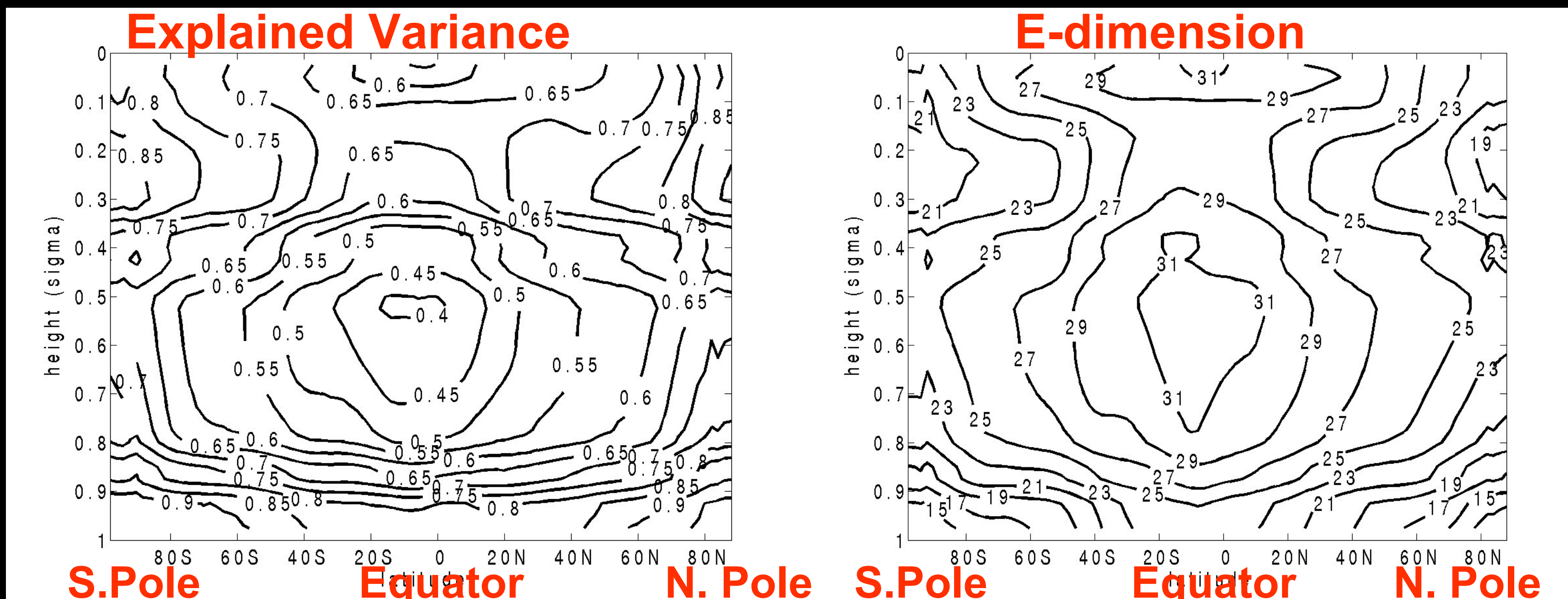
1<E-dimension<2

E-dimension=1



Relationship Between Explained Variance and E-dimension: Correlation:-0.93

averaged in time and along latitudes



When # of ensemble members >20, the explained variance changes little in time and the filter remains stable ("unstable" manifold is well captured), beyond 40, the improvement is small

Main Conclusion of the Study

Lower E-dimension ← **Fast Error Growth** is typically confined to few phase space directions

↓

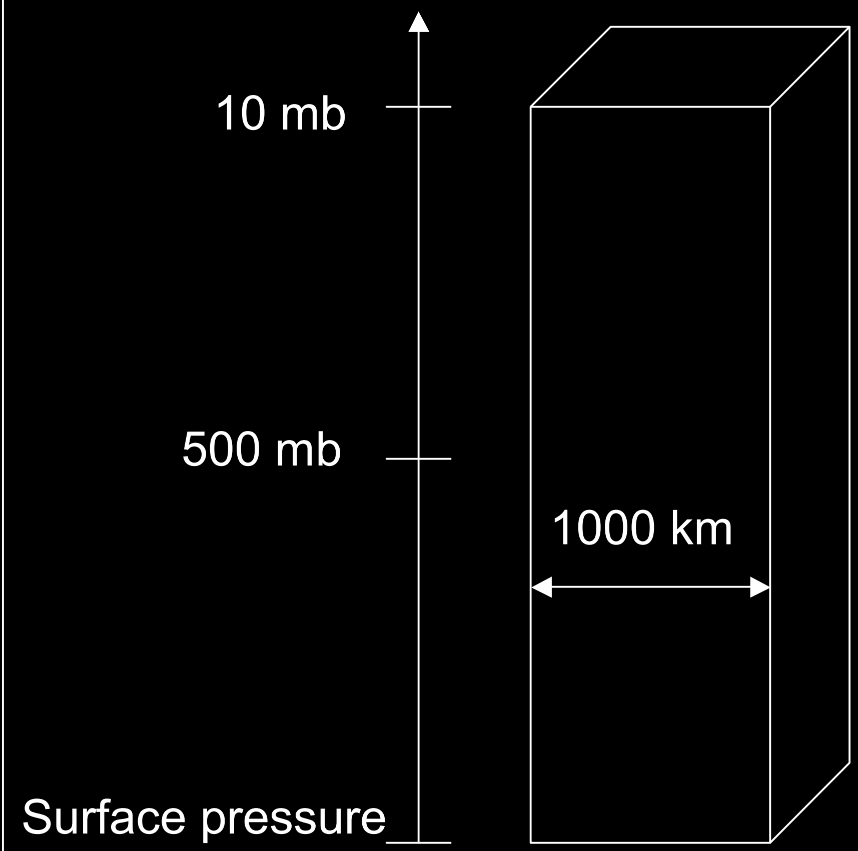
Higher Explained Variance

↓ Analysis expects the right background errors and few observations can make a big correction

Lower analysis error

Result: smaller than average errors in extratropical storm track regions and larger than average errors in regions of deep convection

Extension to Forecast Ensembles

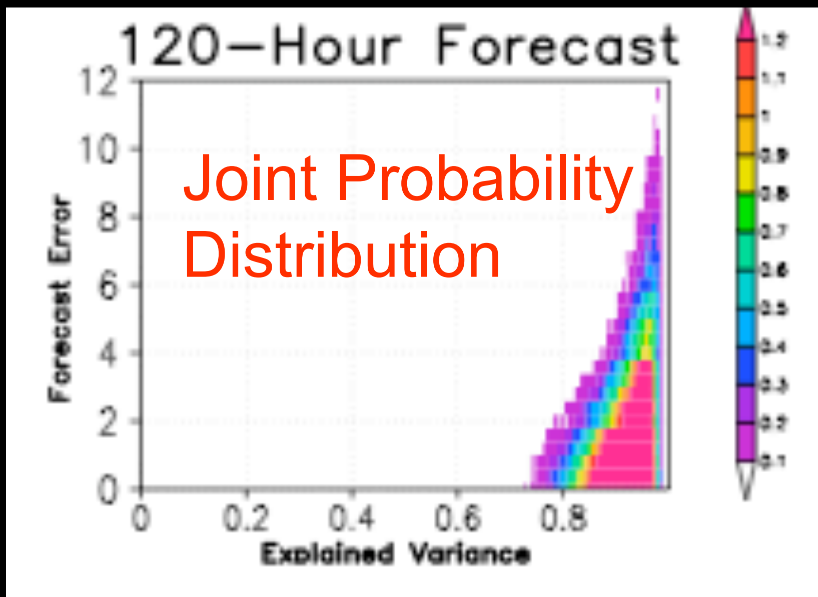


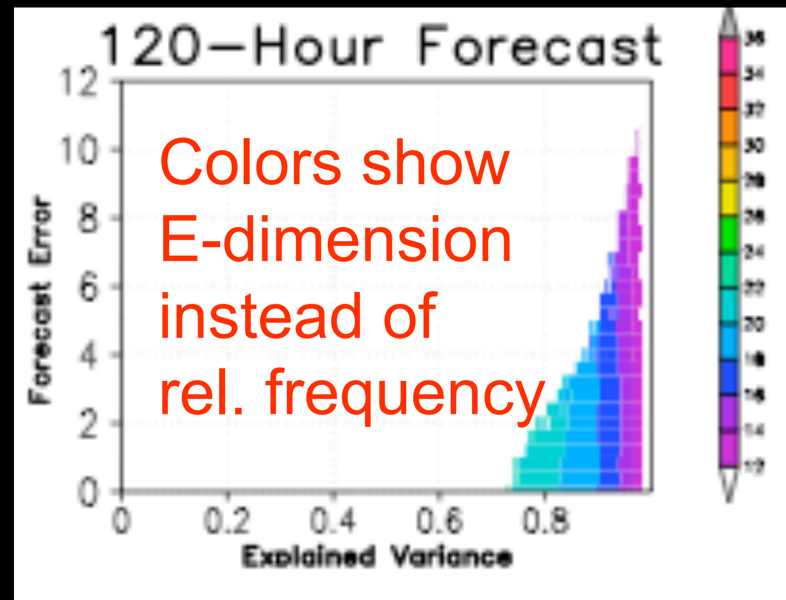
The local region is an atmospheric column (cube)

For the computation of the explained variance and E-dimension, we consider the following state vector components: grid point values of the two horizontal components of the wind, temperature and surface pressure (scaled to have dimension of square-root of energy)

Forecast error is computed for the meridional (south-north) component of the wind at 500 mb

Performance of the Forecast Ensemble Kuhl, Szunyogh, Kostelich et al., 2007, JAS

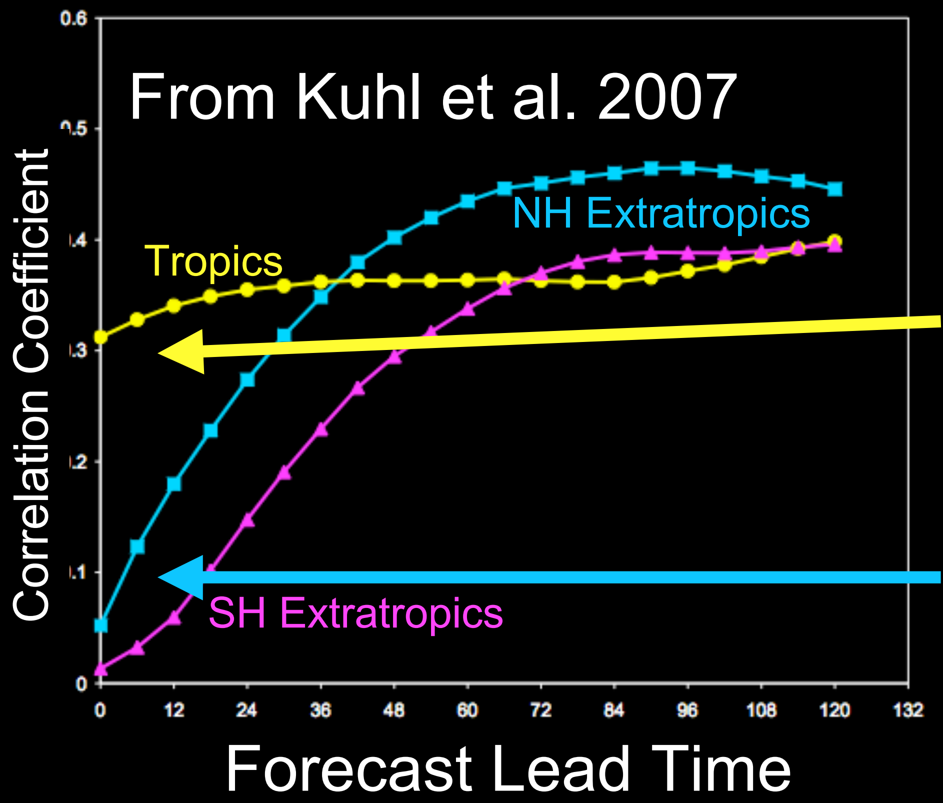




Rapid Error Growth → Low E-dimension → Good Representation of Uncertainties

Low predictability → High Predictability of the Space of Uncertainties

Skill-Spread Correlation for the randomly distributed simulated vertical soundings



Skill: Error in the ensemble mean forecast

Spread: Standard deviation of the ensemble

In the tropics the data coverage is not sufficient to remove all errors captured by the ensemble

In the two extratropics, the data coverage is sufficient to suppress the errors captured by the ensemble

The Motivation for the LETKF
Ensemble DA Comparison Project

- The LETKF was designed to be
  - Computationally the most efficient ensemble-based scheme when a large number of observations is assimilated
  - able to use a community h(x) as a black box
  - able to use different local regions for different type observations
- The results is the LETKF algorithm (Hunt, Kostelich, Szunyogh, 2007: Physica D) and its computer implementation (Szunyogh, Kostelich, Gyarmati et al., 2008: Tellus A)

The Mathematical and Computational Algorithm of the LETKF: Part I

1. Background ensemble is generated by multiple integration of the model from the analysis ensemble of the previous cycle
    1. The different ensemble members are integrated in parallel
2. The observation operator $h(x^b)$ is applied to all ensemble members (This is the only point where h is used and unlike in the earlier LEKF scheme and in the variational schemes, an h linearized around x^b is not needed, which significantly simplifies the development and the maintenance of the system)
    1. The same processors are used to compute h for a given ensemble member that were used to evolve the same ensemble member
3. Information needed to obtain the analysis at the grid points is searched for
    1. K-D tree

The Mathematical and Computational Algorithm of the LETKF: Part II

4. Grid points and related data are distributed between processors
    1. Statistics on wall cloak time/number of observations are collected
    2. Bisection strategy to balance work dynamically
5. Linear algebra is done choosing the basis such that P^b becomes the identity I for the given set of background perturbations, therefore, the expensive computation of its inverse is not required (This is similar to the global ETKF approach of Bishop et al. 2001, the naming LETKF comes from combining LEKF and ETKF)
    1. LAPACK routines
    2. In data rich regions a K^2L operations step (L: number of observations) dominates the cost of the entire algorithm
5. Global ensemble fields are assembled to obtain the analysis ensemble

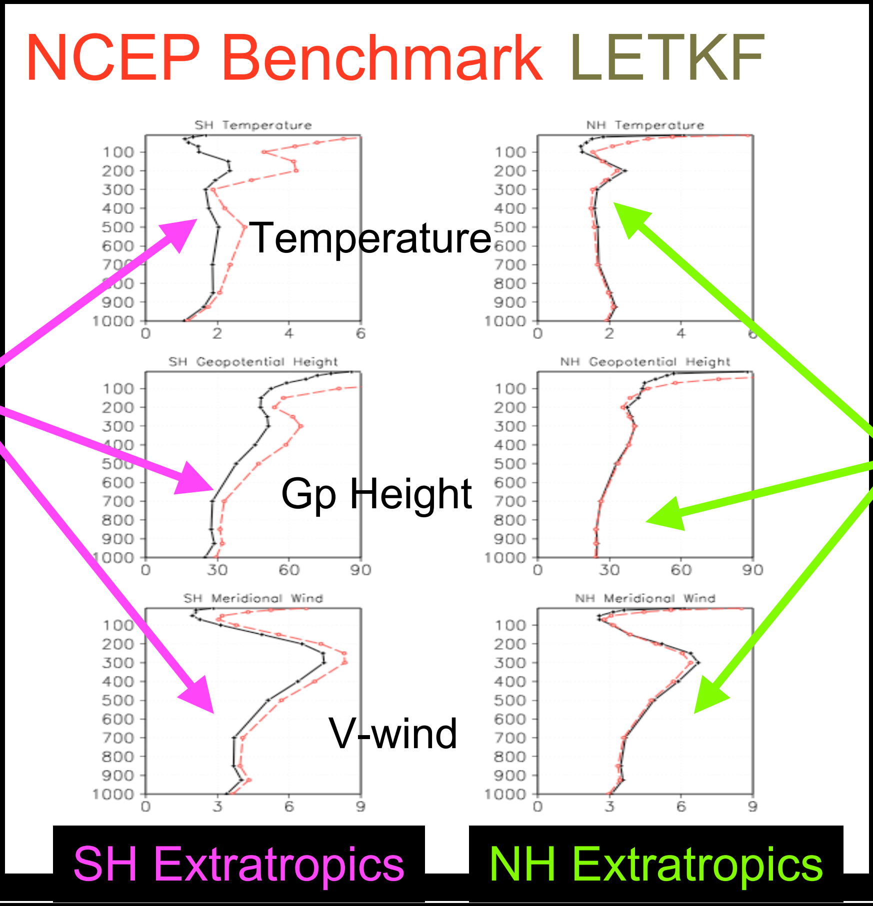
Validation Experiments with the NCEP GFS at resolution T62L28-reanalysis resolution

- Observations of the real atmosphere, except for radiances (Szunyogh, Kostelich, Gyarmati et al. 2007, Tellus, in press)
  - The LETKF and the Benchmark SSI system use different **H** operators; the one used with the LETKF is less sophisticated. This may affect the results near the surface and in areas of high observational density
  - Benchmark SSI data are provided by NCEP (Y. Song and Z. Toth)
  - 60-member ensemble

Comparison of the LETKF and the SSI

48-hour forecasts with real observations (no radiances)

In the SH XT
The LETKF
is more
accurate

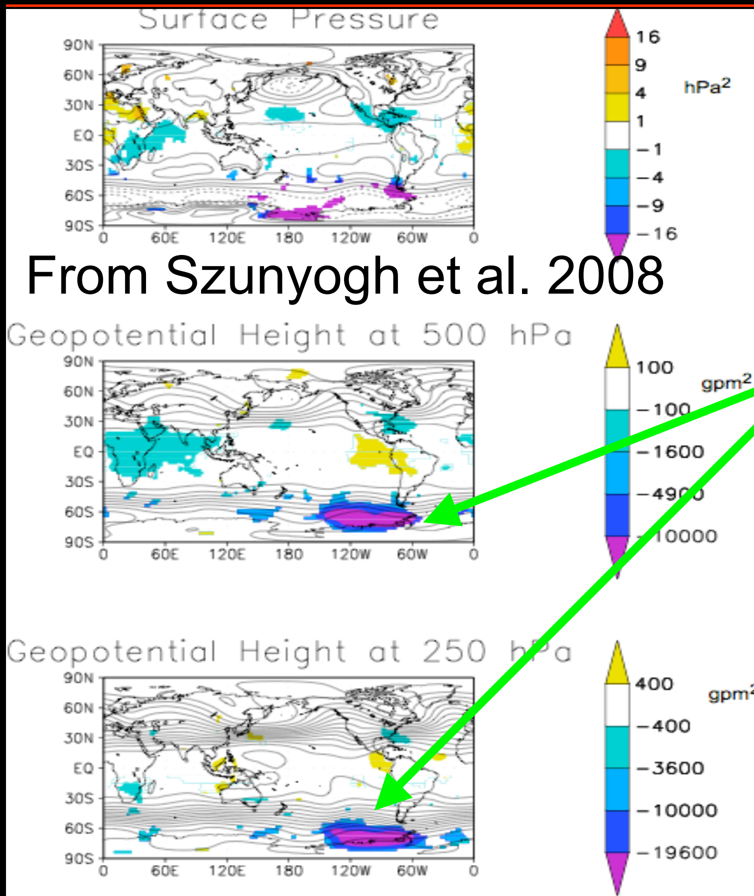


In the NH XT
the two systems
are comparable

From Szunyogh et al. 2008

Comparison of the LETKF and the SSI

48-hour forecasts with real observations (no radiances)

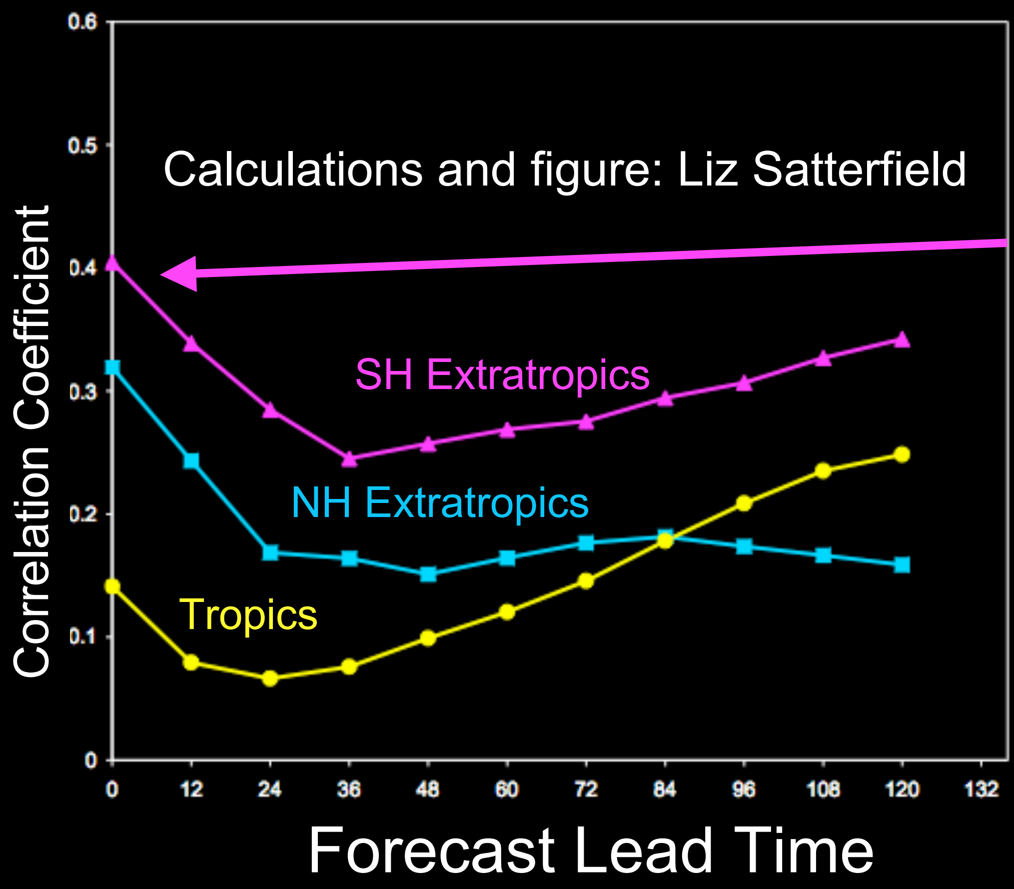


From Szunyogh et al. 2008

The advantage of the LETKF is the largest where the observation density is the lowest

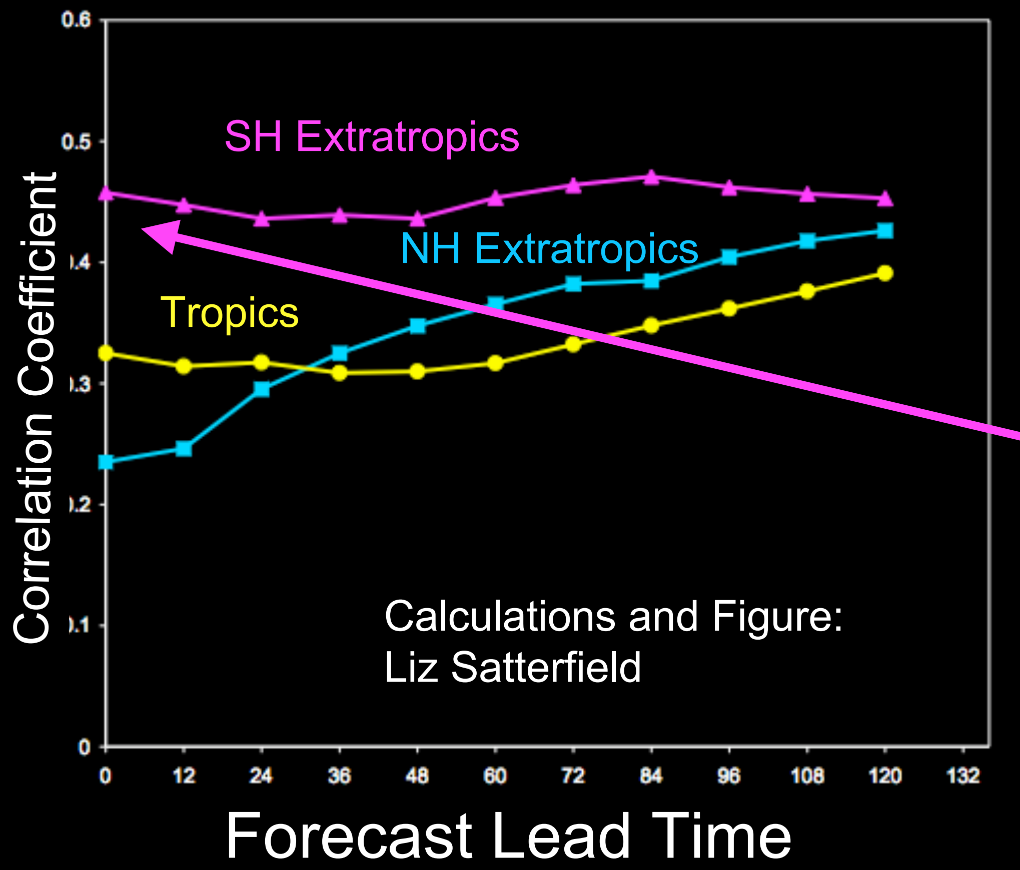
Results are shown only where
The difference is statistically
Significant at the 99% level

Skill-Spread Correlation for the real observations (radiances not included!)



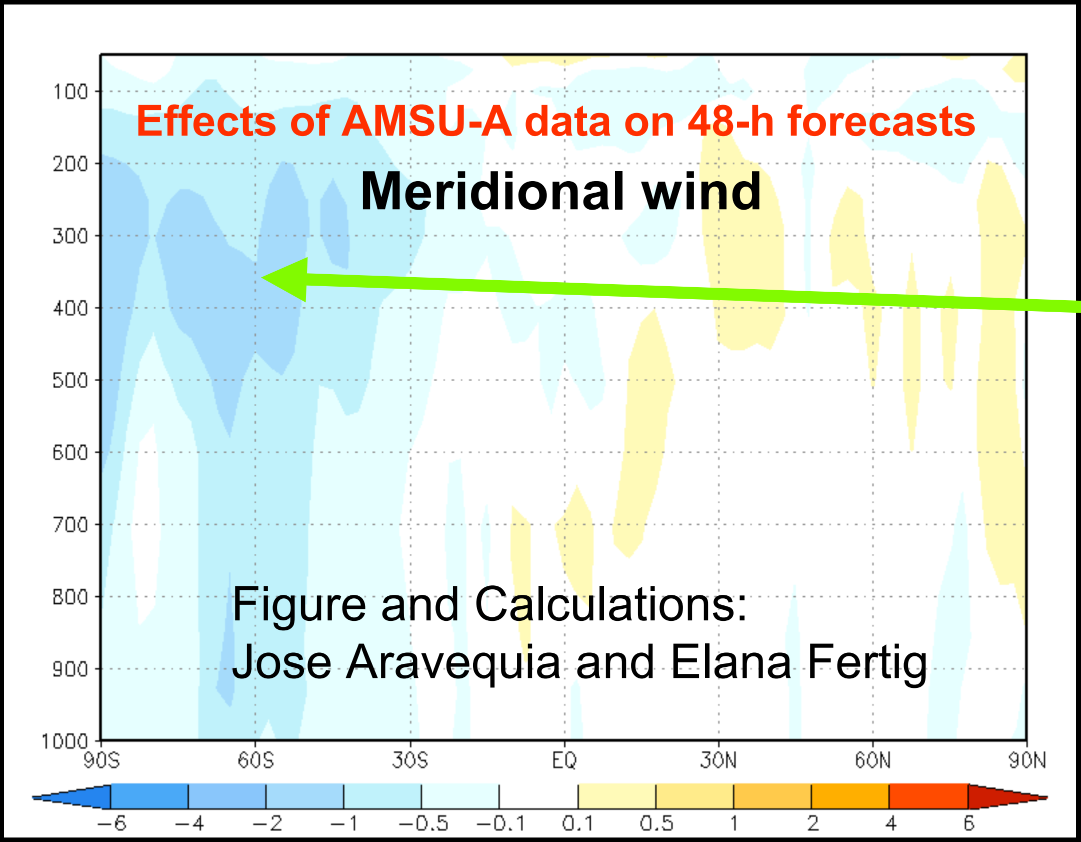
Model errors have little effect on the initially high correlations in the SH extratropics

Skill-Spread Correlation for the simulated observations at the location of the real observations



This time the correlation at analysis time is the highest in the SH XT, indicating that the data coverage is not sufficient to to remove All errors correctly Identified by the ensemble

Latest results: capability to assimilate satellite radiances



- The large improvements in the SH suggests, that there is a lot of useful information in the estimated background error covariance matrix between the temperature (most closely related to the radiances) and the wind

The Goal

is to convince others that they should use our code and/or algorithm

- Those who use our code
  - CPTEC Brazil is in the process of implementing in operations
  - Atmospheric and Environmental Research Inc. (ocean DA for Navy, Phase 2 starts in October)
  - University of Massachusetts-Dartmouth (ocean)
  - ECMWF expressed interest for research-depends on availability of funding
  - Different ocean DA effort at UMD
  - UCLA/JPL proposal to couple the LETKF with the ROMS ocean model
- Those who use our algorithm
  - Japan Meteorology Agency (See talk by Takemasa Miyoshi)
  - Jeff Whitaker (effort on NCEP computer)--at resolution T126 L40 the consensus LETKF system broke even with the GSI, the new operational DA system of NCEP

The Future
has already started

- Further development of the LETKF: estimation of model errors, balance issues, observation error estimation, observation bias, adaptation to higher model resolutions
- Further investigation of predictability with the LETKF/GFS system
- Martian Data Assimilation (2 NASA funded project will start in October--the goal is to couple the GFDL Mars model (also a community model) and the LETKF,
- Impact of wildfire emission (1 NASA funded project, Dave Kuhl)
- Carbon cyclone data assimilation (4-year DOE funded project led by Eugenia)

Reminder:

http://weatherchaos.umd.edu

- Most complete review paper available from the web: Szunyogh et al., 2007: The Local Ensemble Transform Kalman Filter and its implementation on the NCEP global model at the University of Maryland. ECMWF proceedings, in press.