Random Matrix Applications in Environmental Data Assimilation

D. McLaughlin, A. Ahanin, D. Entekhabi – MIT, Cambridge, MA, USA

Data assimilation in the earth sciences:

Characterize ocean, atmosphere, land surface by combining **diverse data sources** and **numerical model predictions** → **Bayesian estimation**

Notable features:

- **Problem size**

  Wide range of time/space scales → Large, high resolution model grids

  Increasing amounts of data at higher rates & resolution → Large estimation problems

- **Nonlinearity**

  Process & instrument nonlinearities are common, often essential for understanding dynamics

- **Uncertainty**

  Model & measurement uncertainties are significant but difficult to characterize

Methods:

Monte Carlo approximations to Bayesian estimators → Sampling problems, random matrices

Motivating Problem: Chaotic Ocean Flows

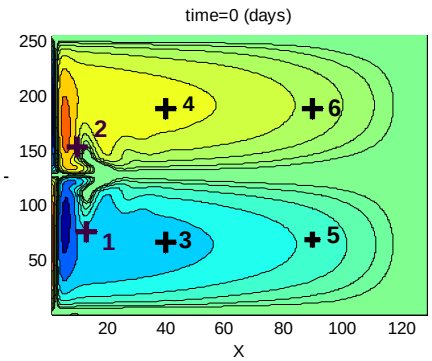
Simulated quasi-geostrophic **double gyre** along an eastern coastline

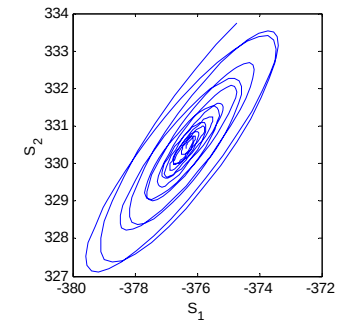
Vorticity contours

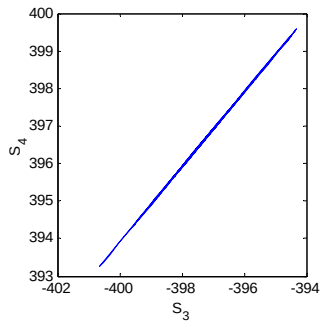
Phase plots

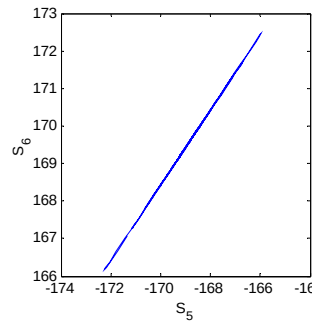
1-2 3-4 5-6

High viscosity

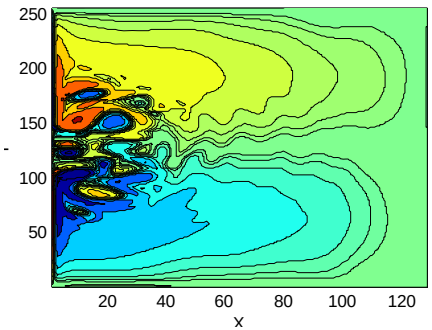


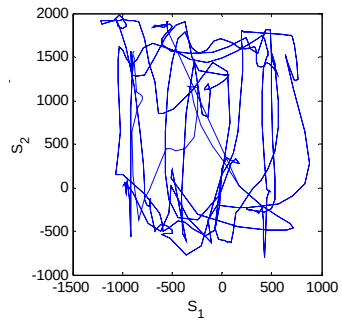


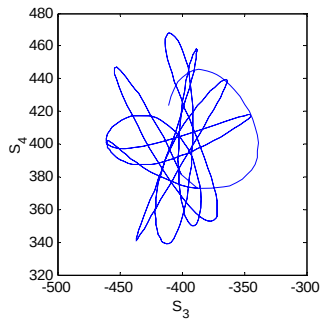


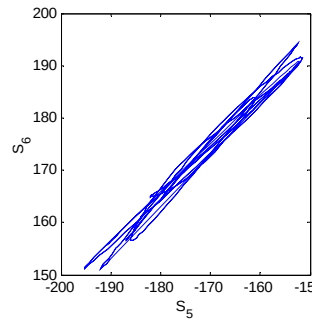


Low viscosity









Characterize chaotic ocean state by using remotely sensed and *in situ* measurements to update imperfect model predictions.

Bayesian Estimation for Nonlinear Dynamic Problems

Bayesian characterization of system state x_t (e.g.vorticity), given measurements y_t :

$p[x_t|y_{0:T}]$ Conditional PDF

System model: $x_t = f_t(x_{t-1}, u_t)$

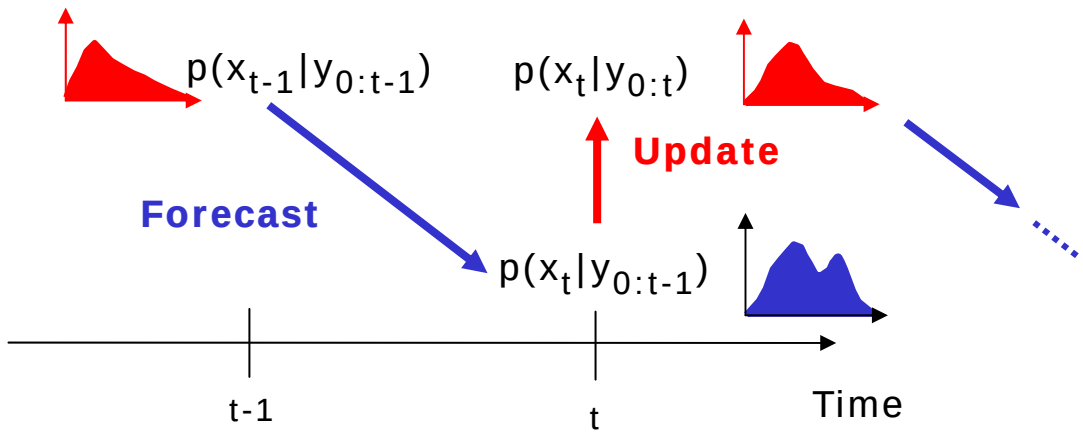
Measurement model: $y_t = h_t(x_t) + e_t$ Meas. at t=0,...T

Consider filtering problems (estimates desired at t=T):

Forecast: $p(x_{t-1}|y_{0:t-1}) \xrightarrow{f_t(x_{t-1},u_t)} p(x_t|y_{0:t-1})$ Fokker-Planck Eq.

Update: $p(x_t|y_{0:t-1}) \xrightarrow{h_t(x_t)} p(x_t|y_{0:t})$ Bayes Rule

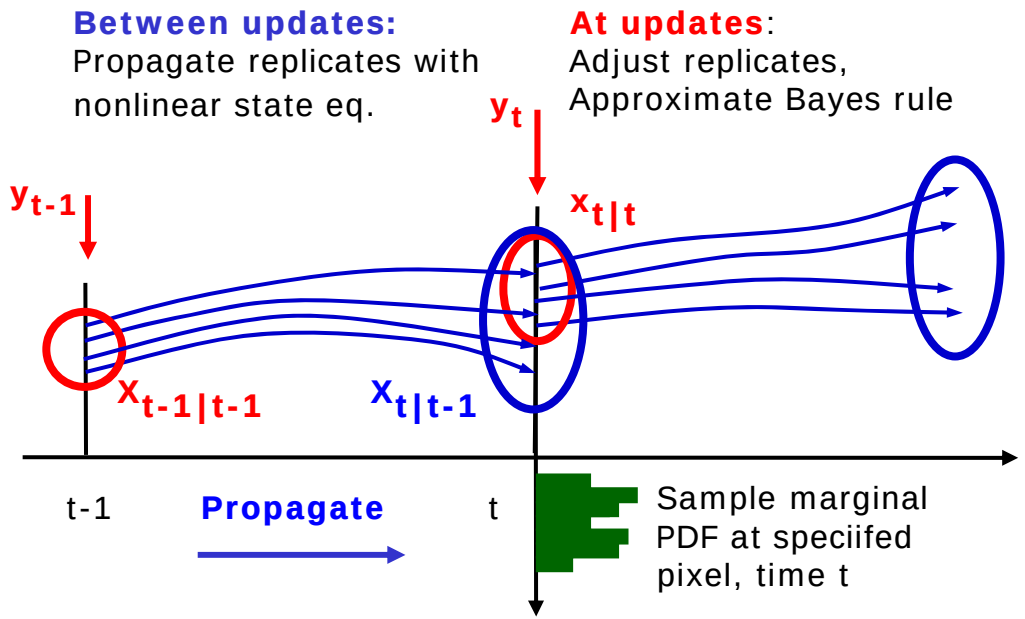
} Sequential



Ensemble Approximations

Solve Bayesian estimation problem with **ensemble methods**

Generate random samples $x^{j}_{t|t-1}$, $x^{j}_{t|t}$ from conditional densities ($j=1, \ldots, N_{rep}$).



Common update approx:

Ensemble Kalman filter:

$$x^{1:N_{rep}}_{t|t} = x^{1:N_{rep}}_{t|t-1} A(y_t)$$

Transformation matrix $A(y_t)$ depends on meas. & **forecast sample covariance**

Basic questions:

How should we generate replicates to properly capture changing statistics, with minimal computational effort ?

How many replicates do we need?

Generating Replicates

Focus on **random initial condition** problems:

Randomly generated replicates (classical Monte Carlo sampling):

1. Draw random replicates from known initial pdf (suppose Gaussian)
2. Propagate replicates with nonlinear dynamics and construct sample statistics (marginal pdf's, covariances, etc) as needed

Computational demands $\longrightarrow$ small ensemble size (low rank sample covariance)

Alternatives: Some or all replicates aligned with dominant directions of variability

How can we identify dominant directions? Leading eigenvectors ?
Approach is difficult to implement for highly nonlinear systems.

For now, adopt classical random sampling approach.

Determine how sampling errors depend on sample size, time, meas. updates.

Numerical Experiment – Unconditional (no measurement updates)

Chaotic Lorenz 1995 model

For j=1, …, 100:

$$\frac{dx_j}{dt} = (x_{j+1} - x_{j-2})x_{j-1} - x_j + F$$

$$x_{-1} = x_{J-1}$$

$$x_0 = x_J$$

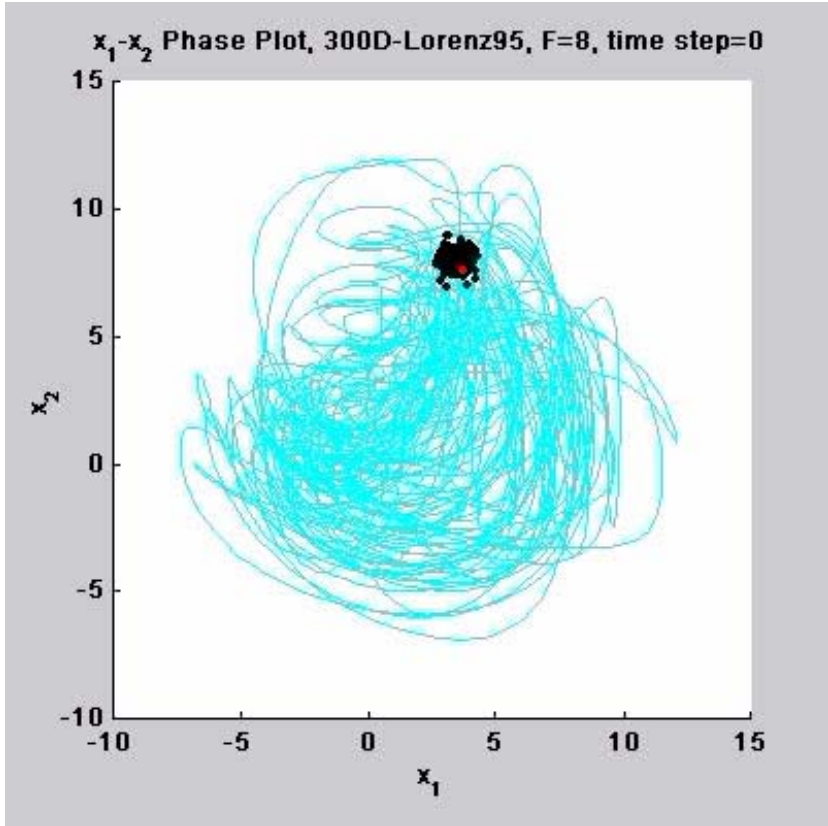
$$x_{J+1} = x_1$$

x(0) uncorrelated Gaussian, F=8

Sampling procedure

1. Follow cloud of 2000 replicates over time, evaluating “true” covariances as required.
2. Estimate covariances from 200 replicates, evaluate leading eigenvalue error. Repeat 100 times. Evaluate resulting sampling error statistics.
3. Compare sampling error statistics to results from random matrix theory

Trajectory of initial Gaussian cloud over time



Note that replicates disperse to fill entire attractor -- transition to non-Gaussian state

Numerical Experiment – Conditional (with measurement updates)

Updating procedure

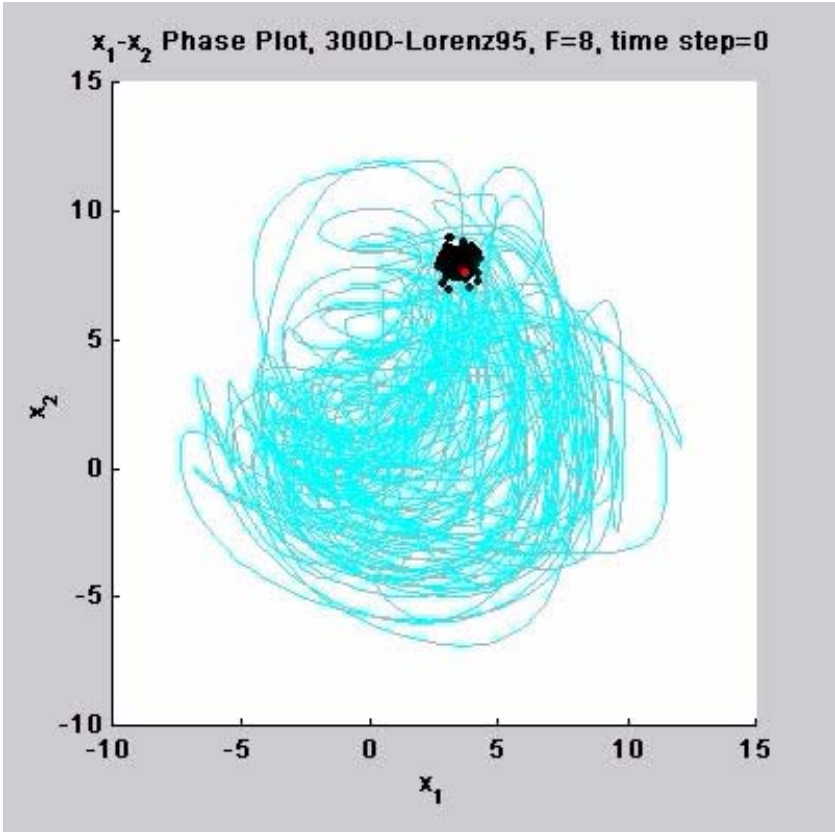
1. Update all replicates with high quality measurements – measurement spacing = 20, 80 time steps.
2. Use Kalman update based on prior "true" covariance at update time

Frequent measurement updates reduce variance and keep conditional distribution more Gaussian

Sampling procedure

Follow same approach as in unconditional case

Trajectory with update every 80 time steps



Note collapse of particle cloud around true value (red) at update times

Analysis of Covariance Sampling Error

Effectiveness of Kalman updating procedure depends on:

1. Adequacy of **normality assumption**
2. Accuracy of **sample covariances** estimated from ensemble.

Accuracy of covariance sample estimates can be expressed in terms of:

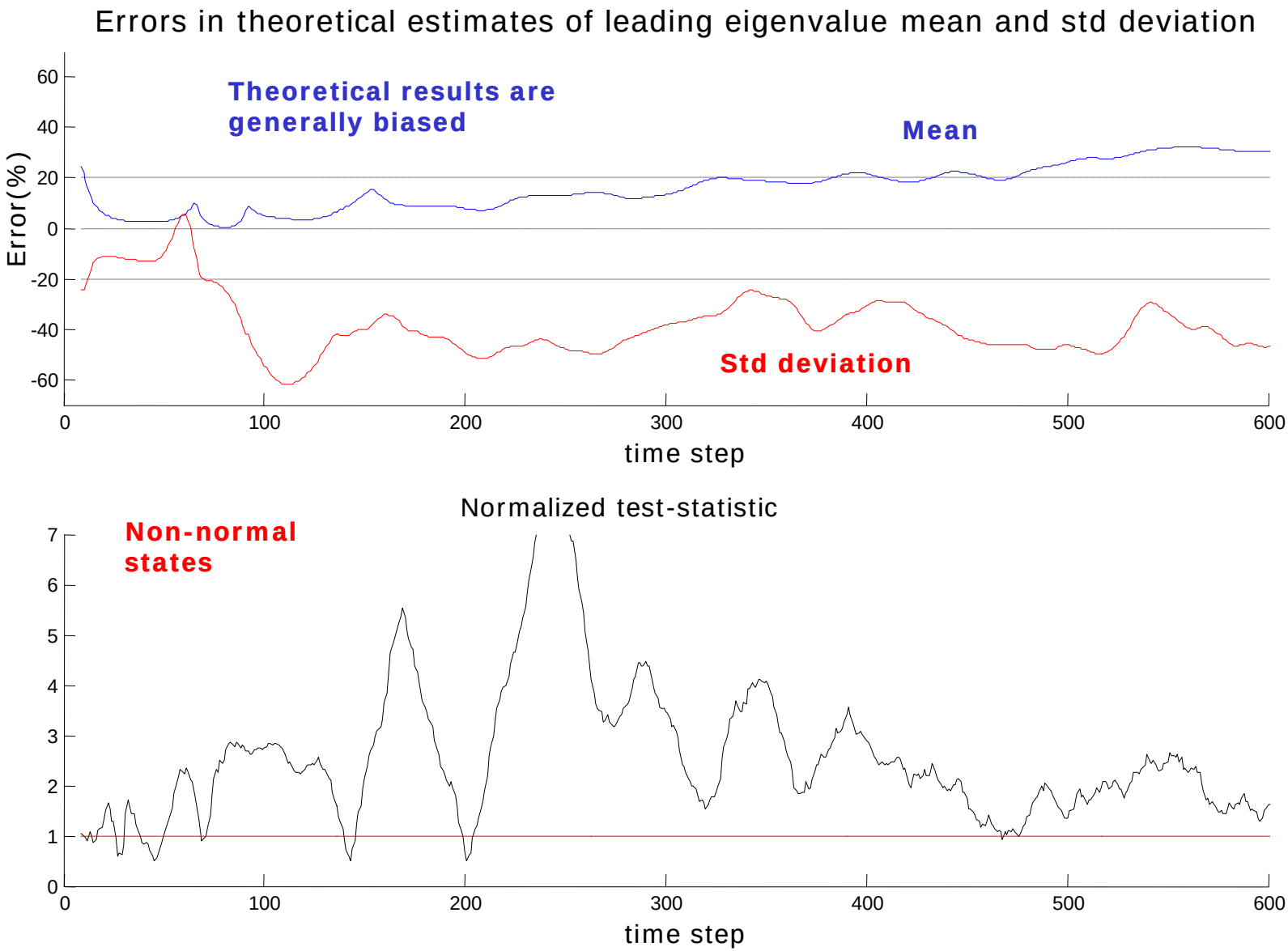
1. **Differences between eigenvalues** of "true" (n_{rep}=2000) and sample (n_{rep}=200) covariances
2. **Cosines of angles between eigenvectors** of "true" (nrep=2000) and sample (nrep=200) covariances

Closed form expressions statistics of these sampling error measures are given by **Paul (in press)** for the full rank "spiked" spectrum Gaussian case where

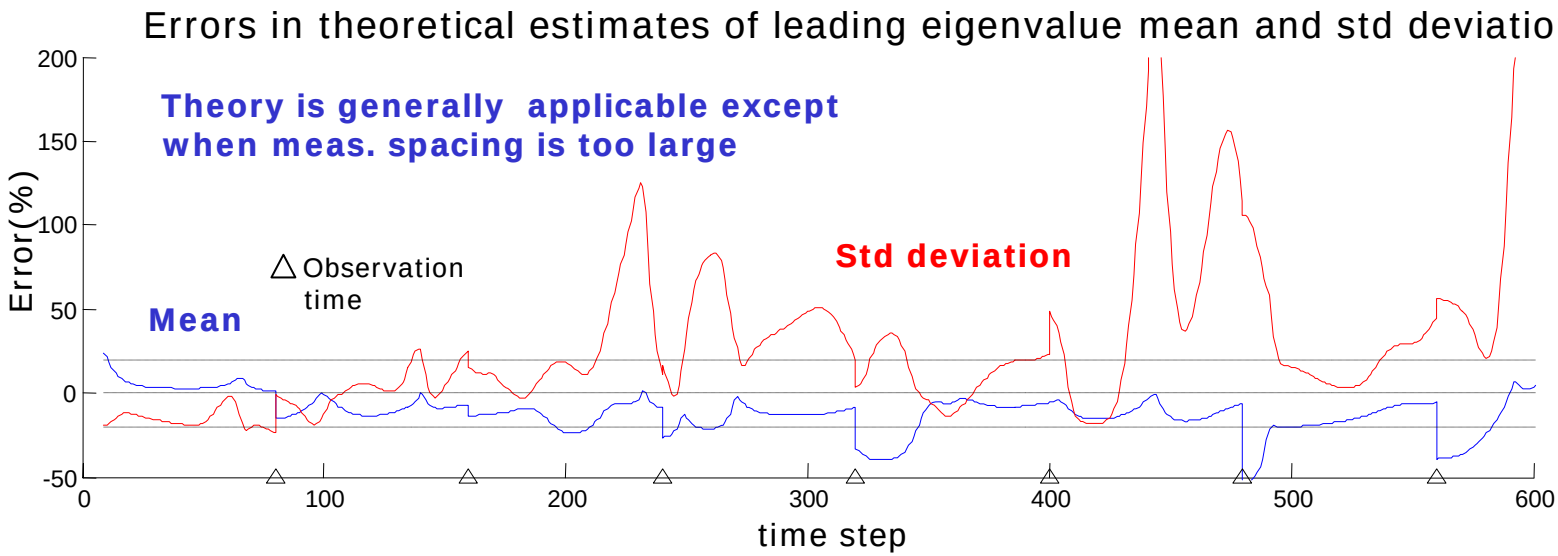
$$\gamma = \frac{n_{samp}}{n_x} \geq 1$$

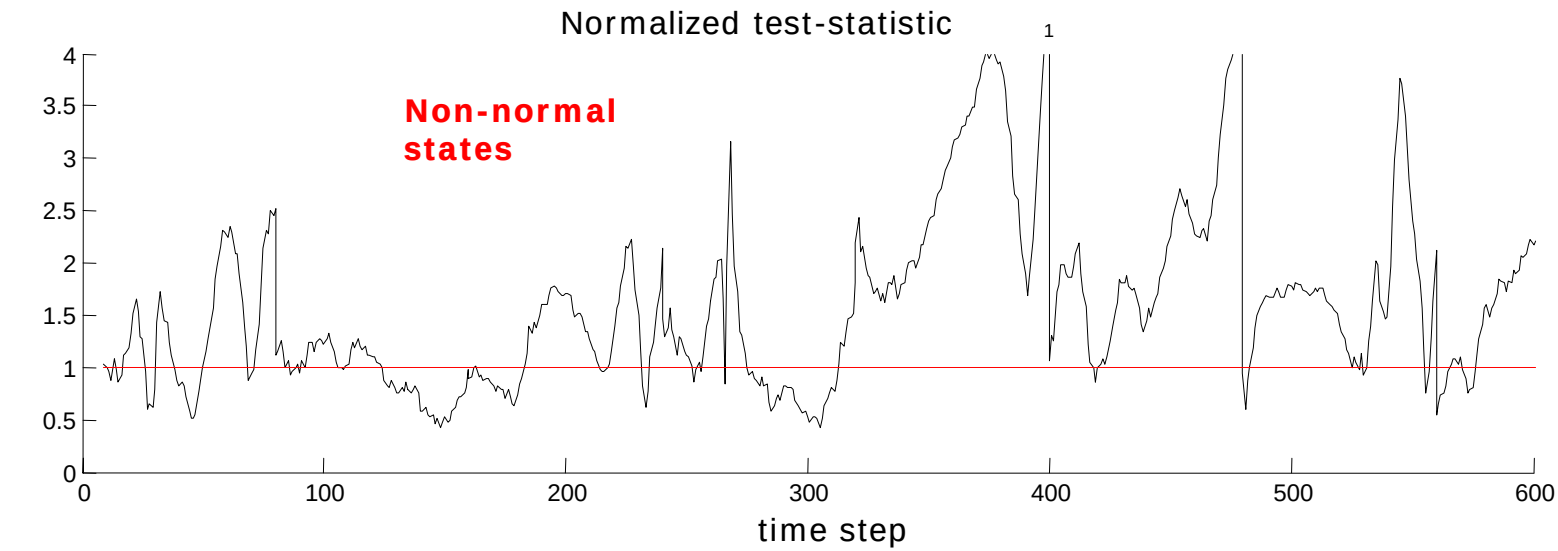
In this talk, focus on **eigenvalue errors**.

Unconditional Sample Properties

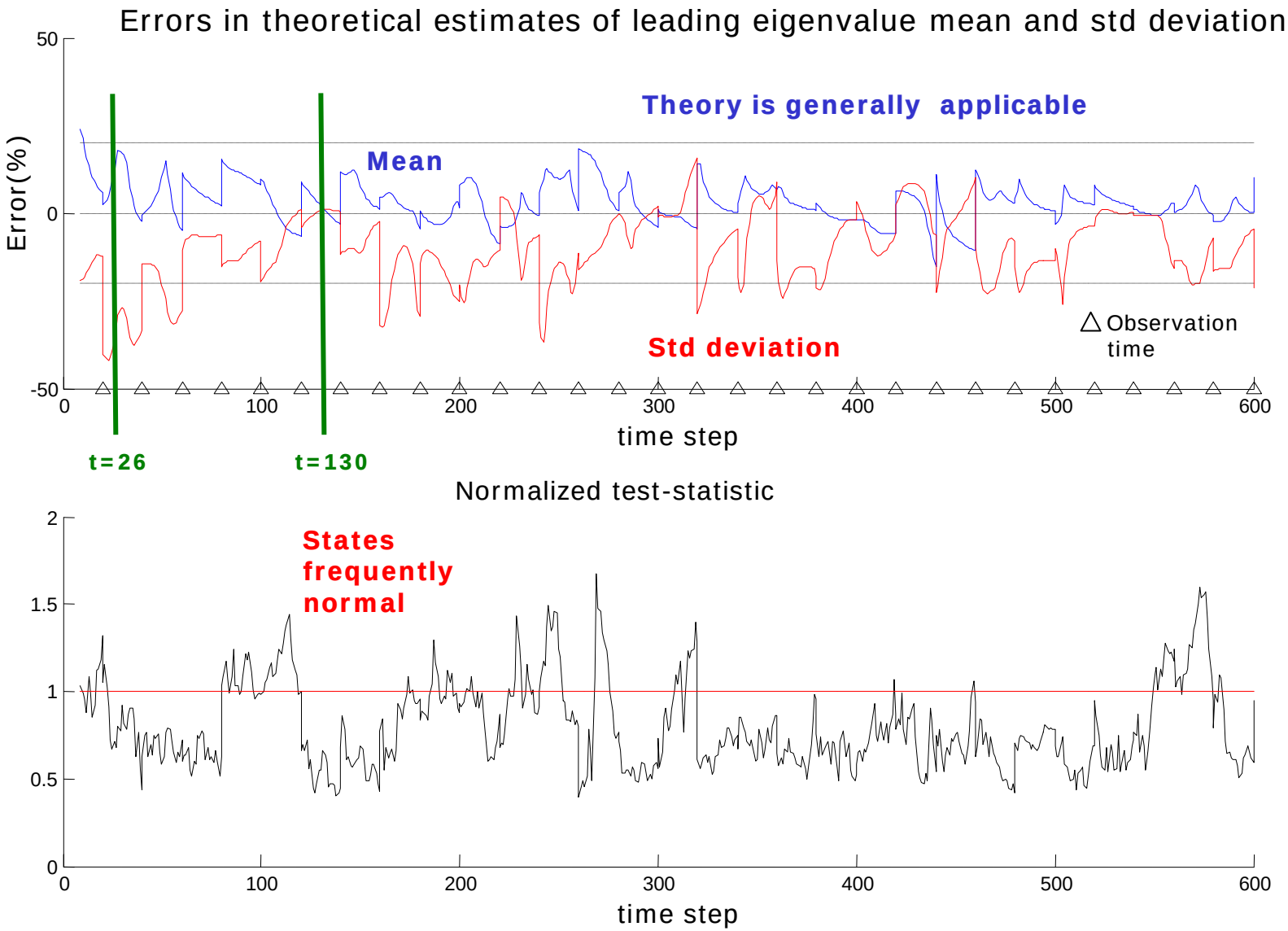


Conditional Sample Properties
Measurement Update every 80 Time Steps



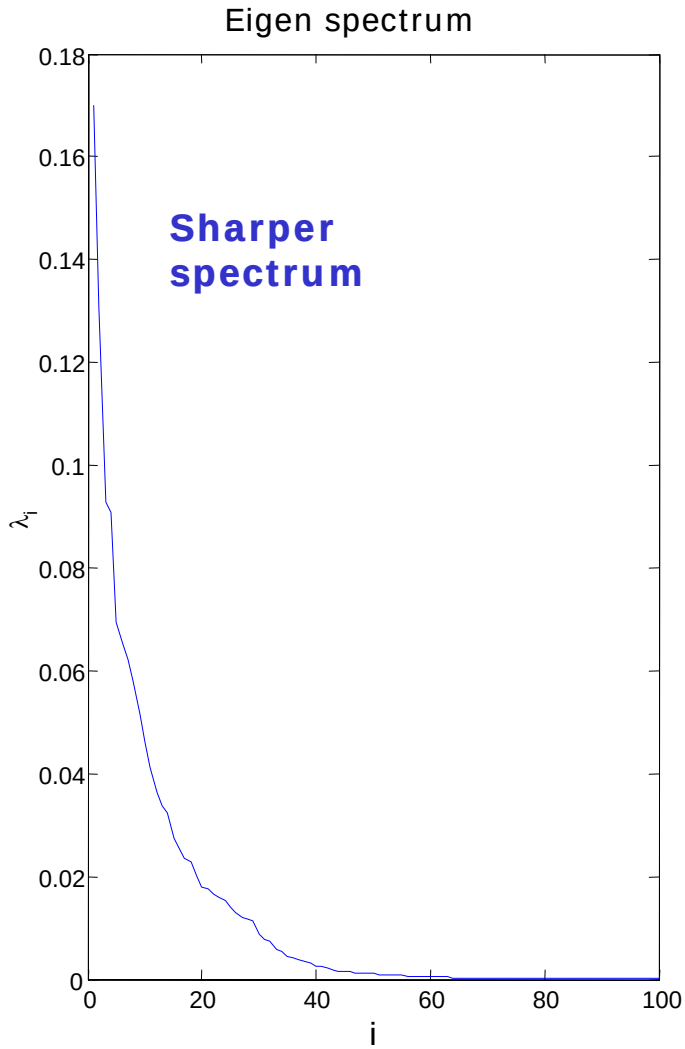


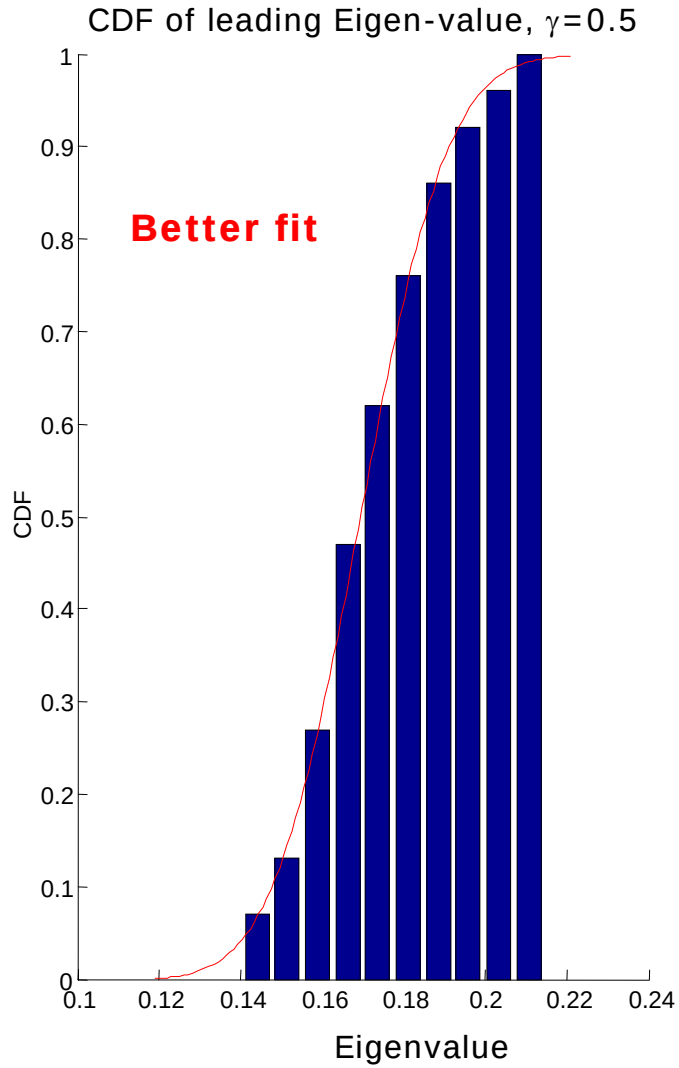
Conditional Sample Properties
Measurement Update every 20 Time Steps



Conditional Eigenvalue Spectrum and CDF at t=130

Meas. spacing = 20

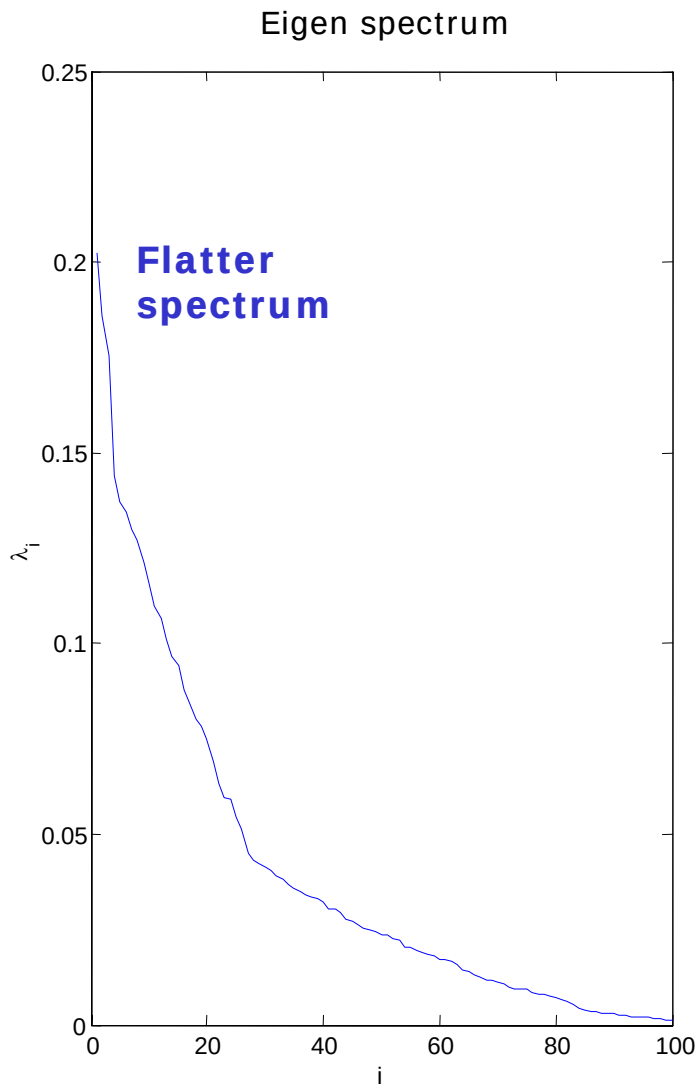


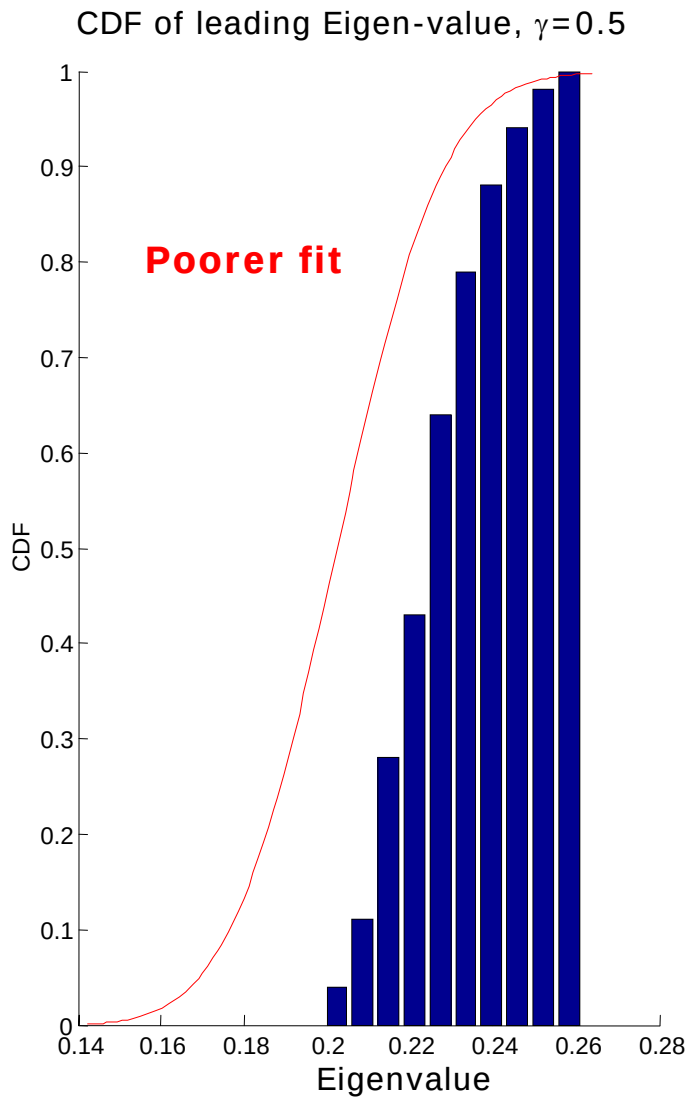


Conditional Eigenvalue Spectrum and CDF at t=26

Meas. spacing = 20

a4





a4 The plots for openloop and the 80 time step between observations, almost always look like this.
aahanin, 7/10/2006

Conclusions

- Initial uncertainties in states of chaotic nonlinear systems (e.g. meteorological and oceanographic) gradually **expand to fill entire attractor** if there are **no measurement updates**

- Transition to non-Gaussian behavior can be slowed or stopped if replicates are **periodically updated** with measurements.

  ↳ **Measurement updating extends the validity of the assumptions** used in the ensemble Kalman filter approx. and in random matrix theory

  Real spectra are rarely flat beyond leading eigenvalue – this raises **normalization issues** for theoretical approximations

- For large nonlinear problems we need a **small set of very informative replicates**. Random matrix theory may help identify efficient replicate generation procedures.