The Relation between Ensemble Size and Expected Error in Ensemble Filter Data Assimilation

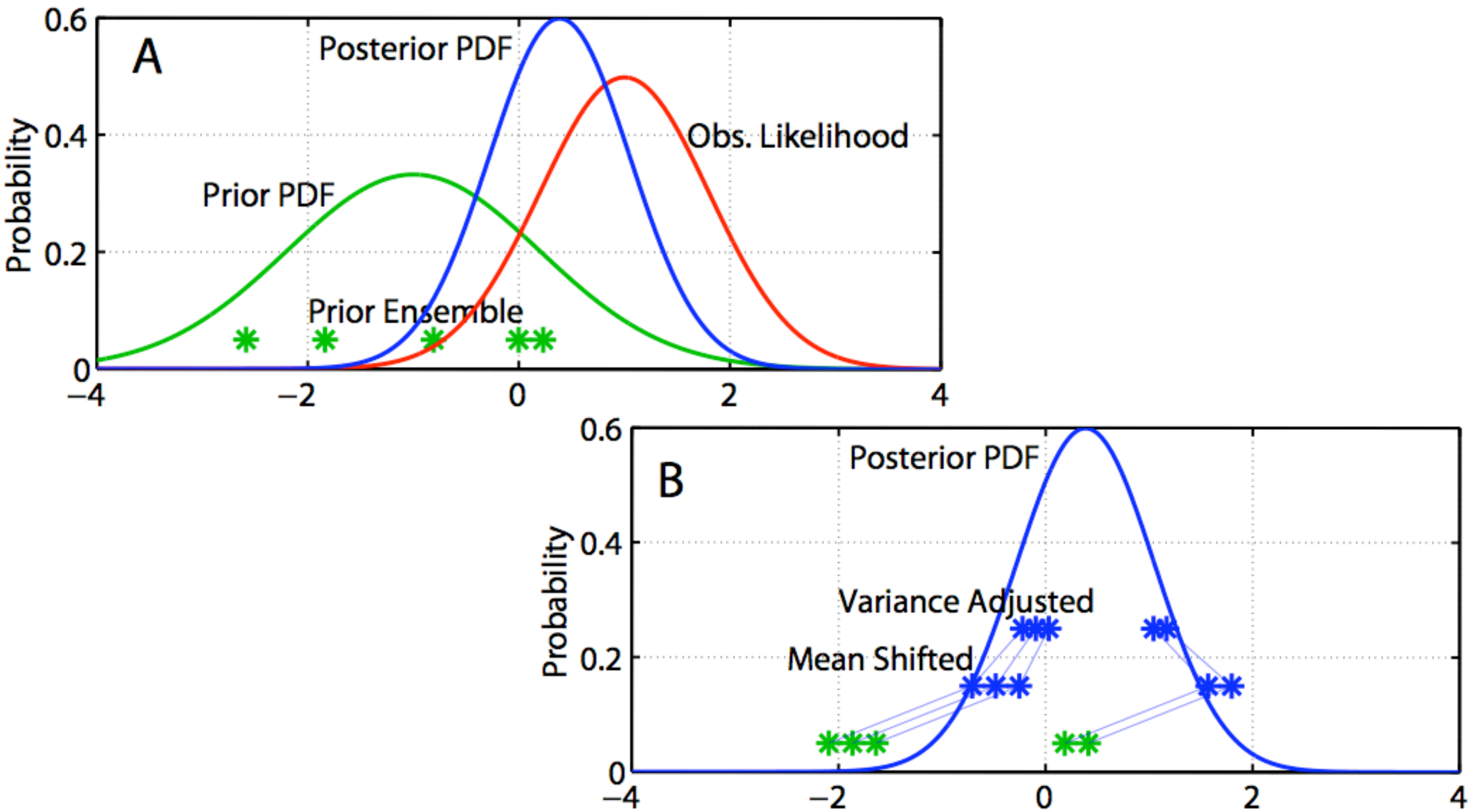
Jeffrey Anderson
Data Assimilation Research Section
NCAR/IMAGe

AMS Annual Meeting
22 January, 2008

- Atmospheric Models are Enormous.
- Number of Observations is Enormous.
- Can we get nearly optimal results with tiny ensembles?
- What is the main challenge?

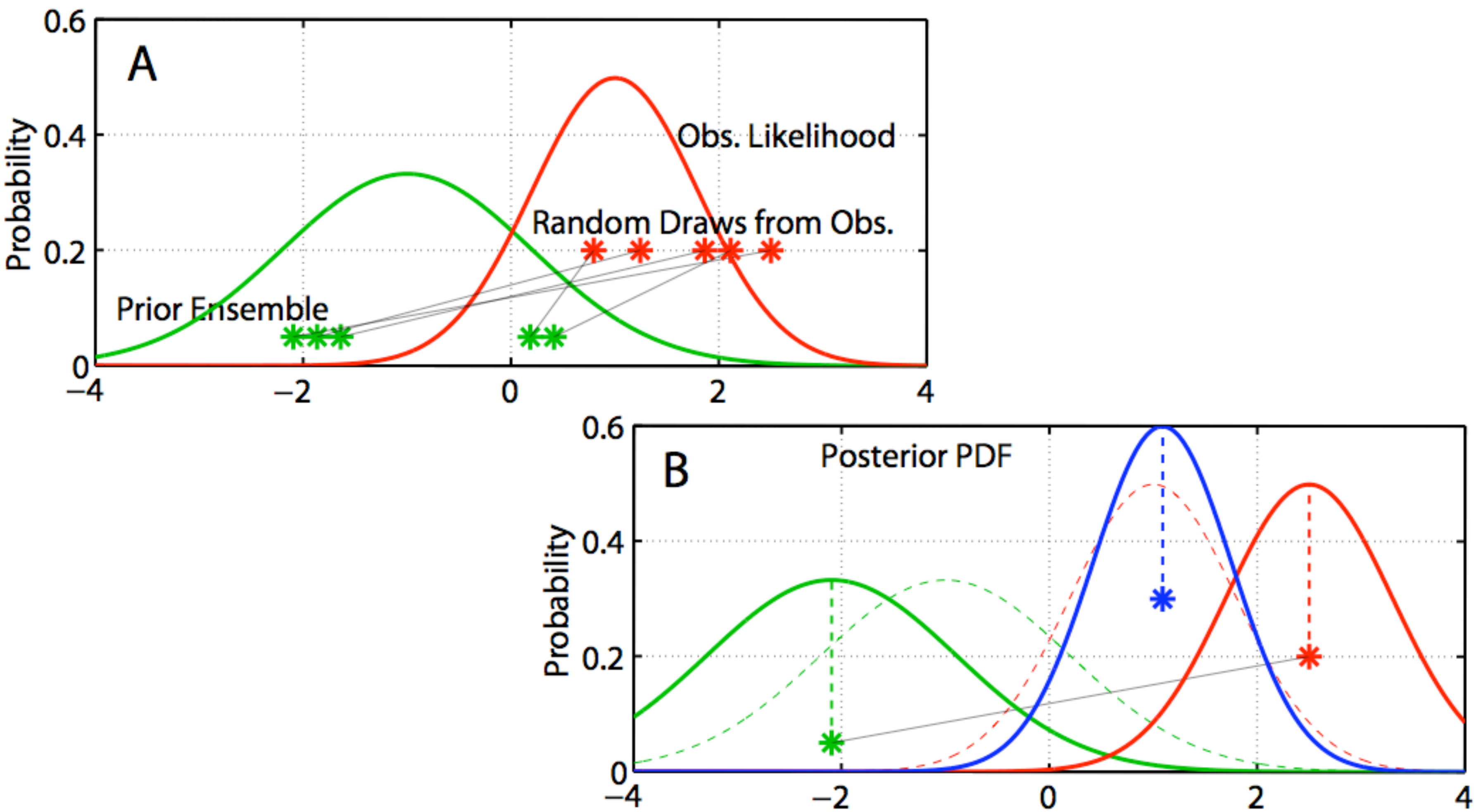
A Deterministic Ensemble Kalman Filter (EAKF)

➢Observation Space Algorithm Schematic



A Monte Carlo Ensemble Kalman Filter (EnKF)

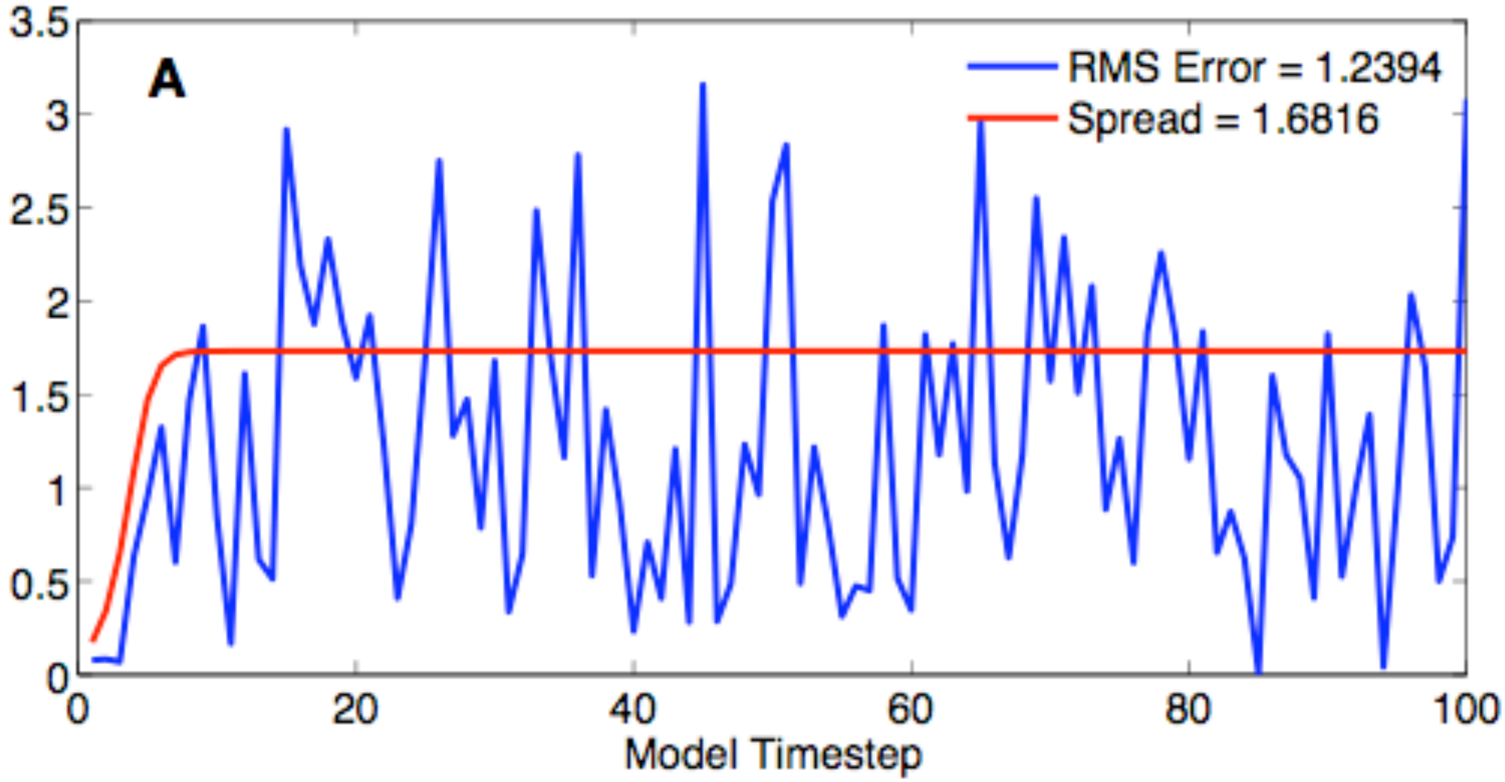
➢Observation Space Algorithm Schematic



1-Dimensional Linear Model: $x_{t+1} = \alpha x_t$

Observe x after each advance, obs. error is Normal(0, 1).

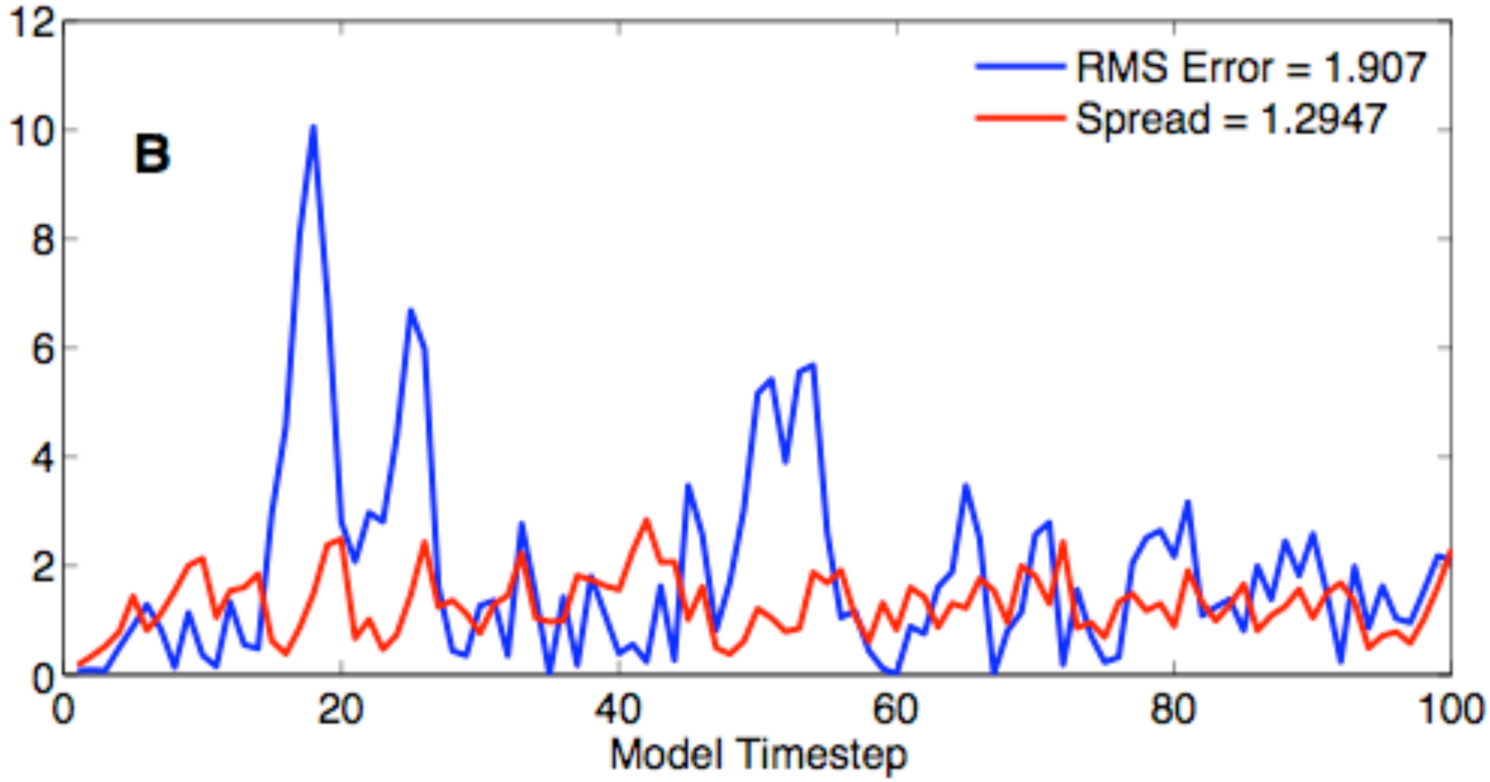
EAKF converges to exact spread, sample of mean (Same as KF).



1-Dimensional Linear Model: $x_{t+1} = \alpha x_t$

Observe x after each advance, obs. error is Normal(0, 1).

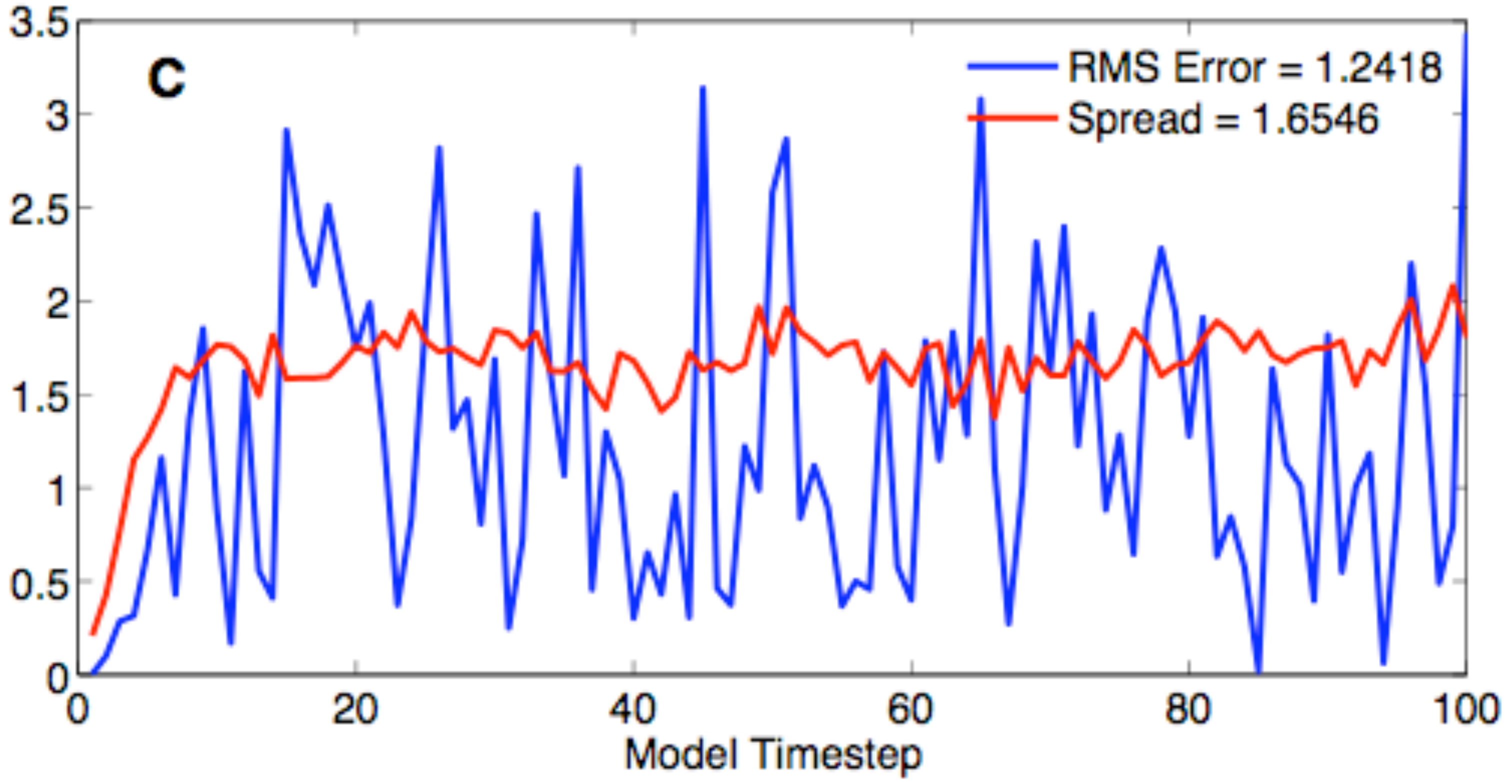
EnKF is Monte Carlo: 4-member ensemble is noisy.



1-Dimensional Linear Model: $x_{t+1} = \alpha x_t$

Observe x after each advance, obs. error is Normal(0, 1).

EnKF is Monte Carlo: 100-member ensemble is less noisy.

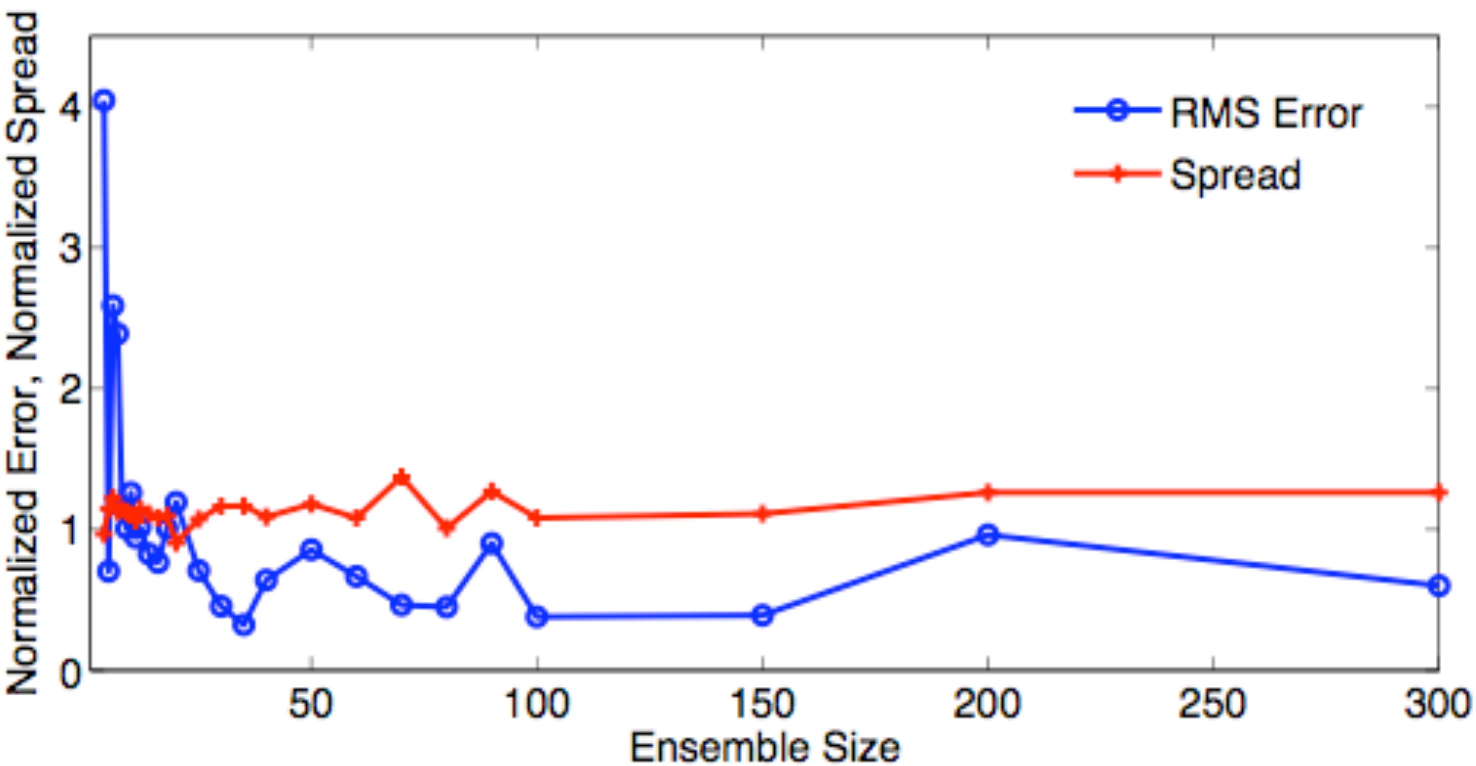


1-Dimensional Linear Model: $x_{t+1} = \alpha x_t$

Observe x after each advance, obs. error is Normal(0, 1).

EnKF error and spread – correct multiplied by ens. size, N.

RMS error surplus, spread shortfall, inversely proportional to N.

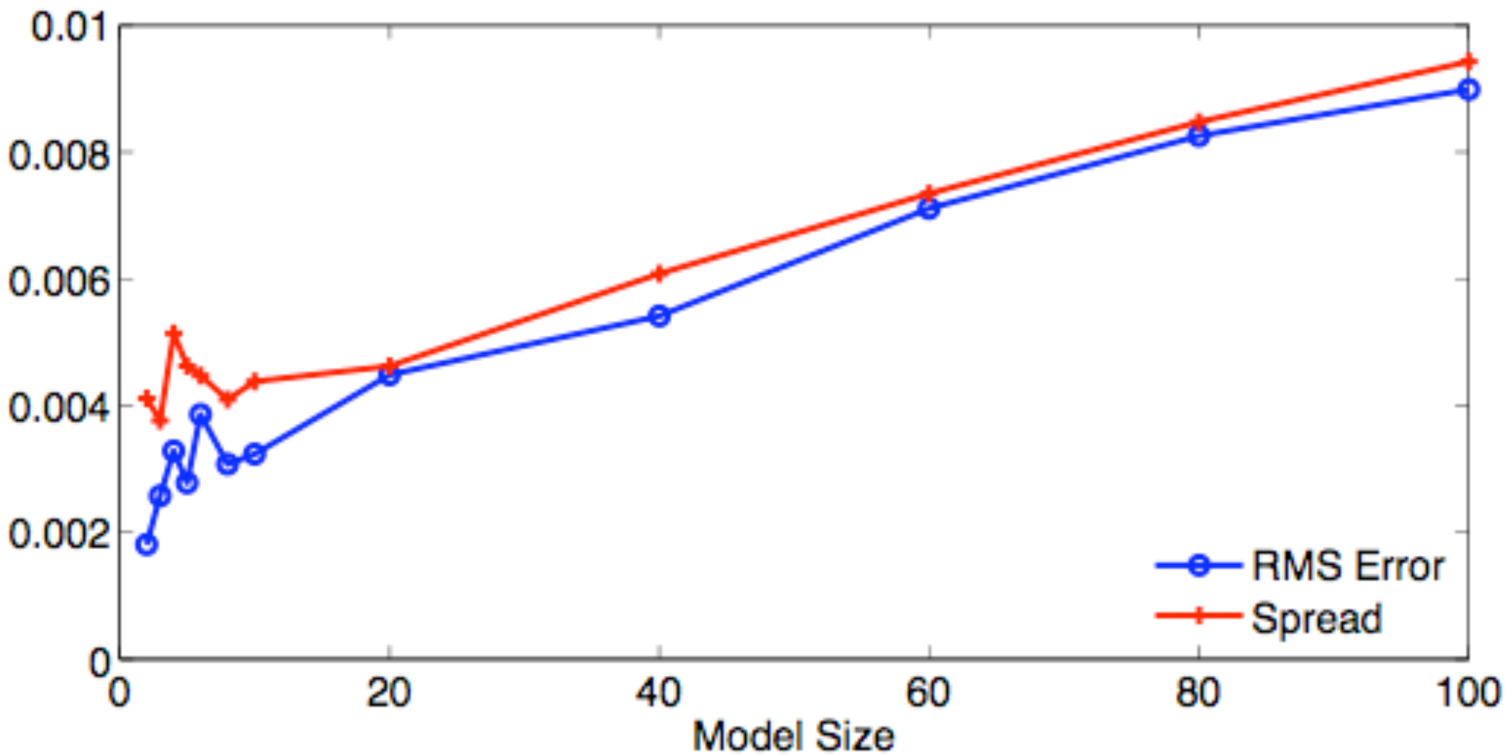


EnKF 101-Member Ensemble

Error as function of linear model size from 1 to 100.

Total error proportional to model size.

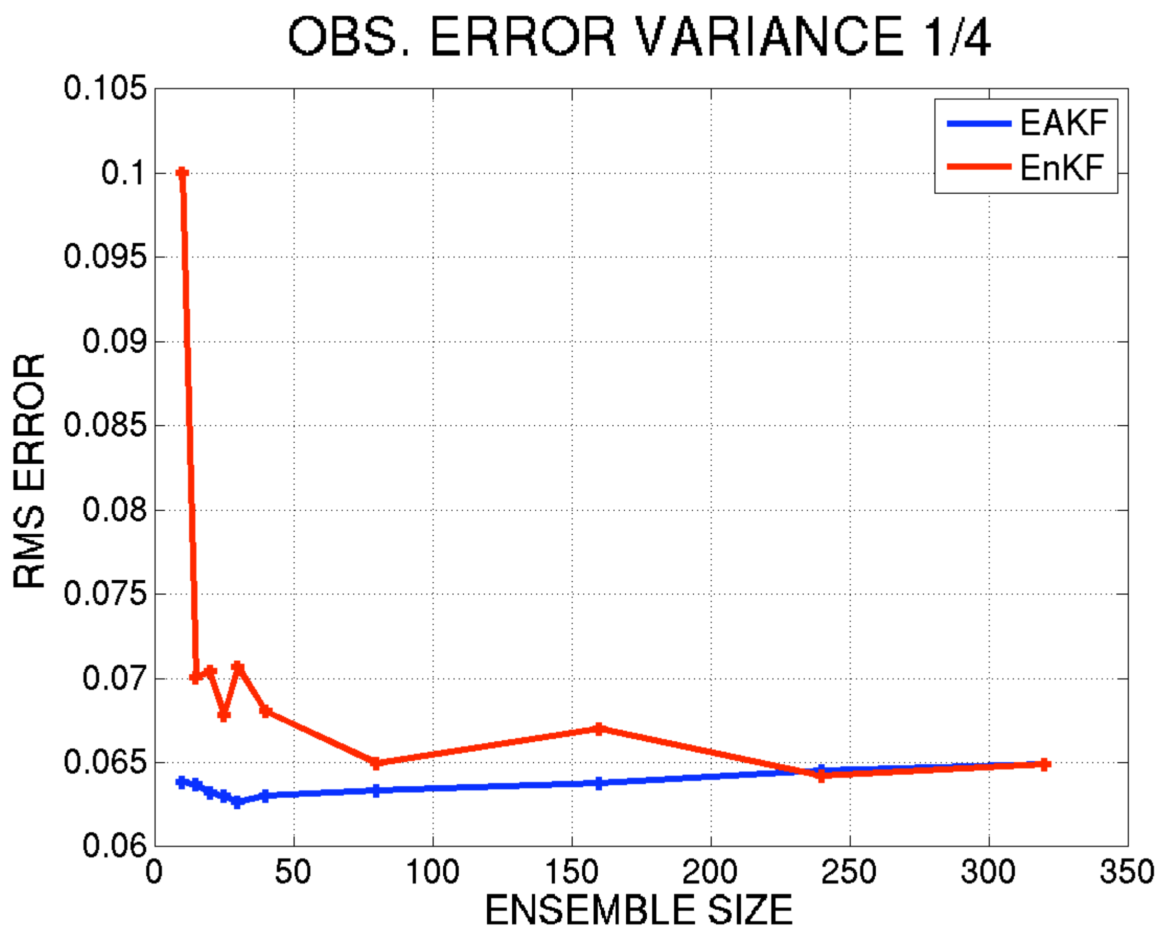
Component errors not affected by model size!



Nonlinear Dynamics and Sampling Error

Lorenz-63. Observations of x+y, y+z, z+x.

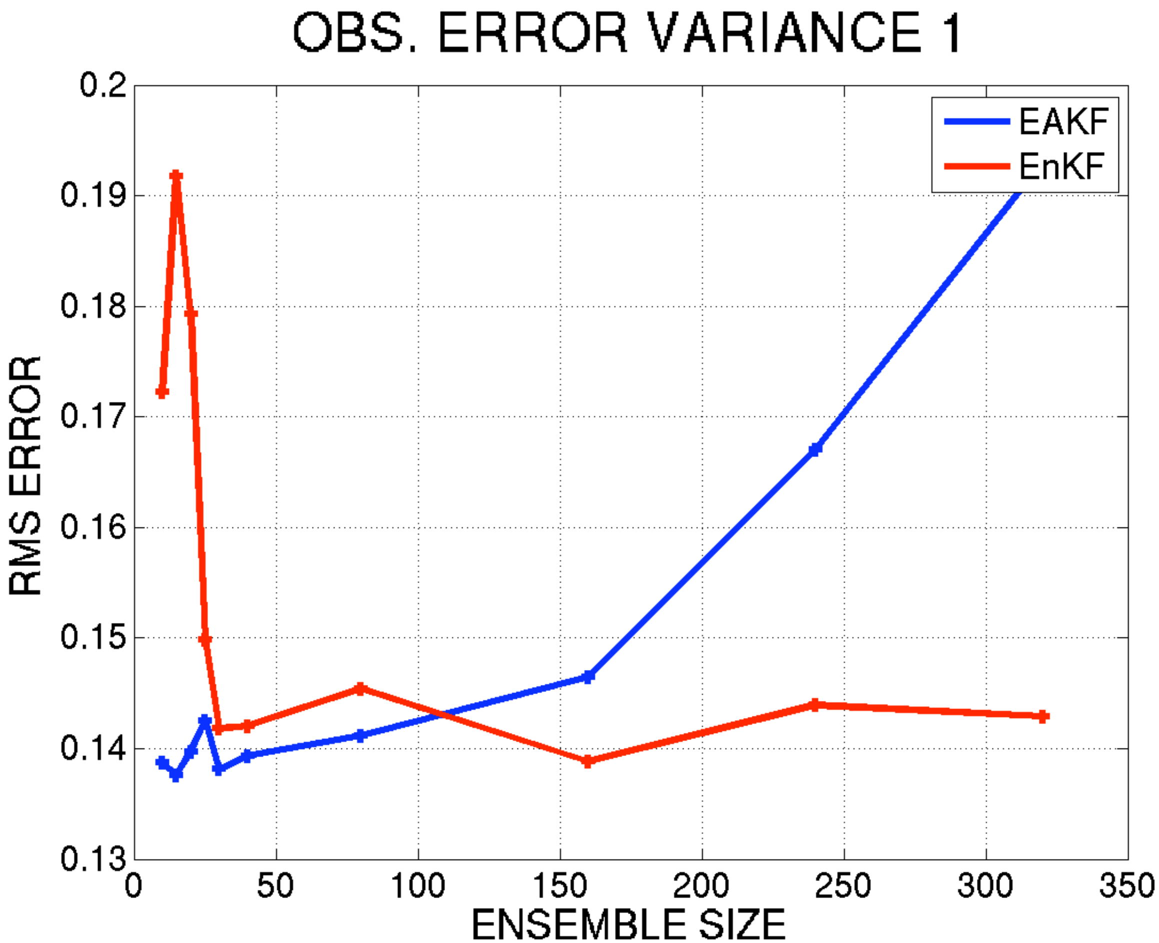
Nearly linear.



Nonlinear Dynamics and Sampling Error

Lorenz-63. Observations of x+y, y+z, z+x.

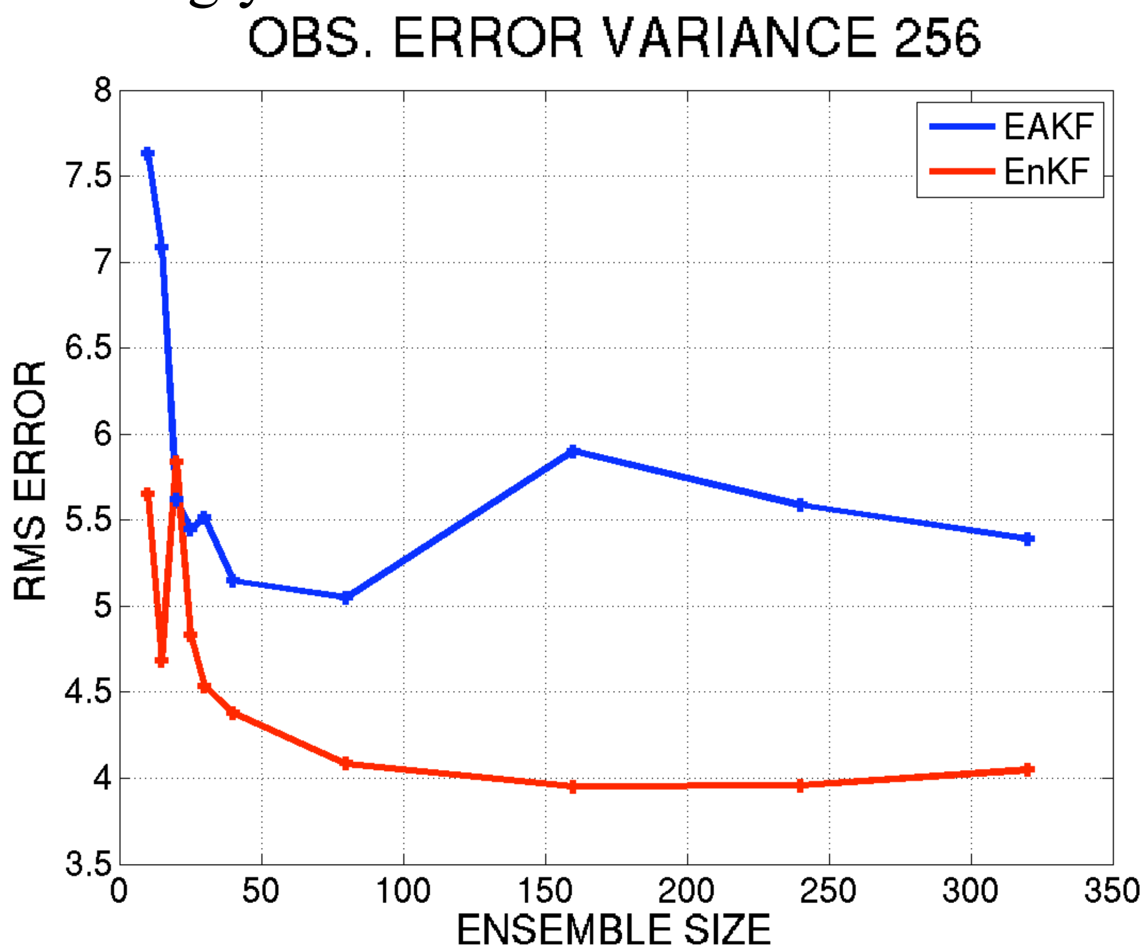
Mildly non-linear.



Nonlinear Dynamics and Sampling Error

Lorenz-63. Observations of x+y, y+z, z+x.

Strongly nonlinear.



Degeneracy, small ensembles, and localization.

100-Dimensional Model, EnKF and EAKF fail for N<101.

But, can localize.

Modify correlation between observations and state variables.

Statistical approach (hierarchical filter):

- There is correlation signal and noise,
- Run a group of ensemble filters, differ in initial members,
- Get a sample of correlations,
- Filter them to retain signal.

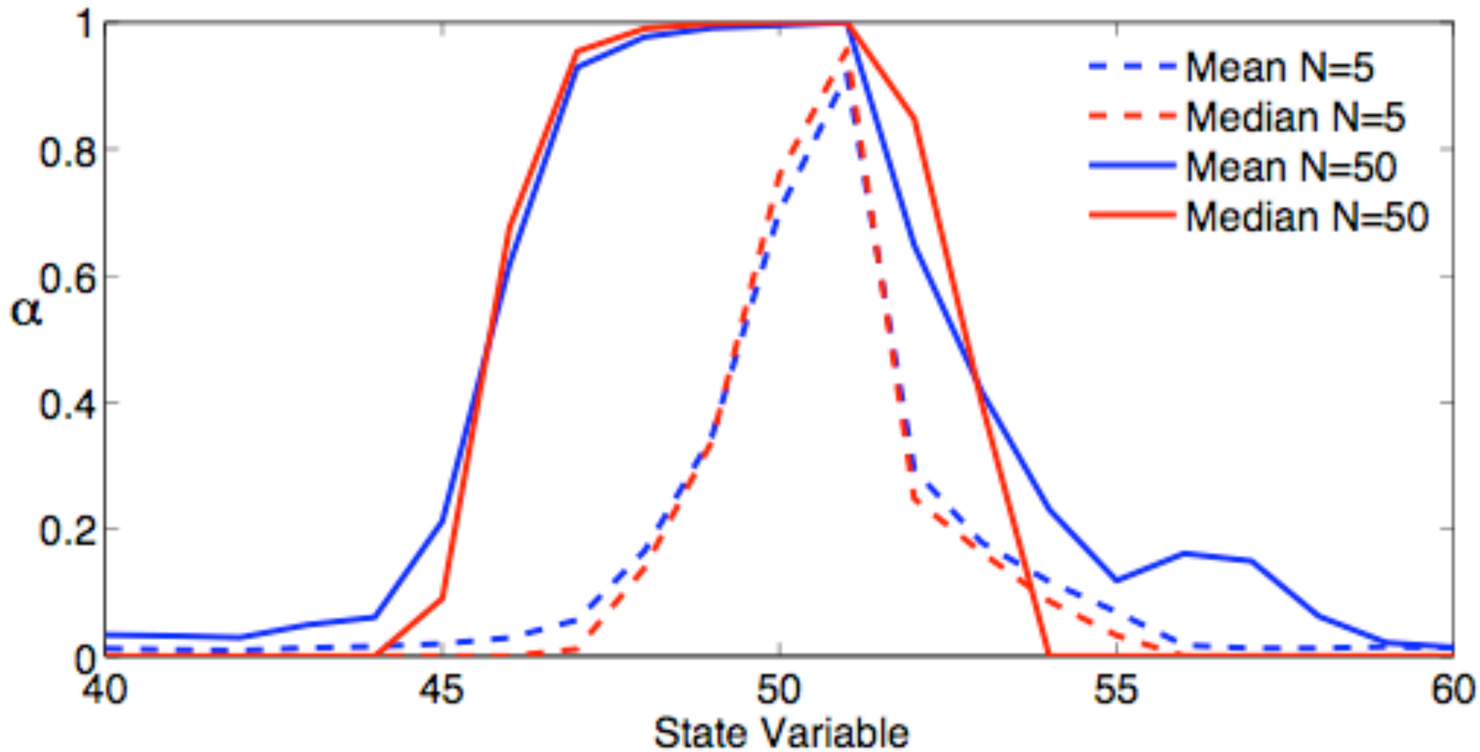
Localization for 100-Dimensional Linear Model

Observation $y_i = 0.7x_i + 0.3\ x_i$

Run groups of N-member ensembles.

Keep time mean/median of localization.

Results for observation 50.

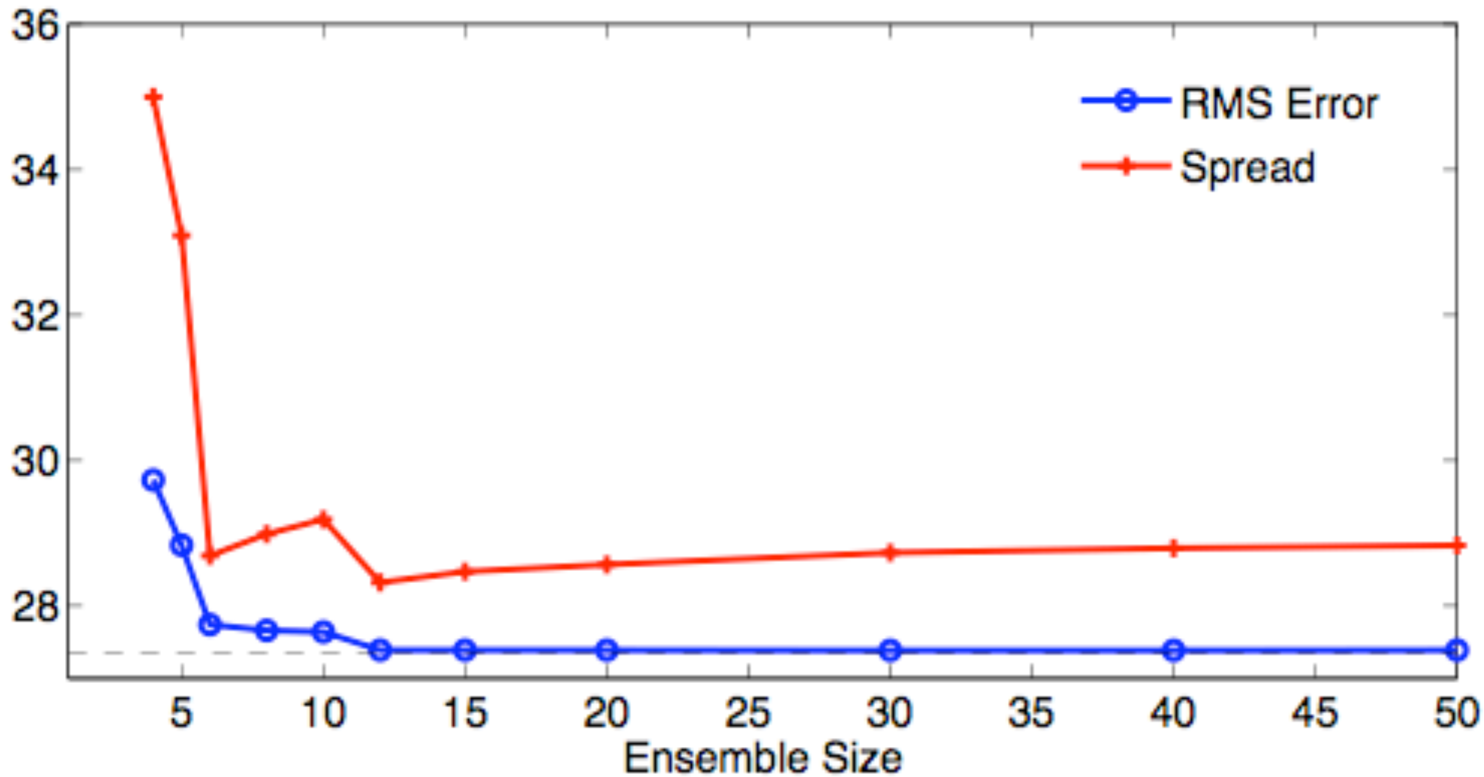


Localization for 100-Dimensional Linear Model

Observation $y_i = 0.7x_i + 0.3\ x_i$

Use time median from groups for single N-member EAKF.

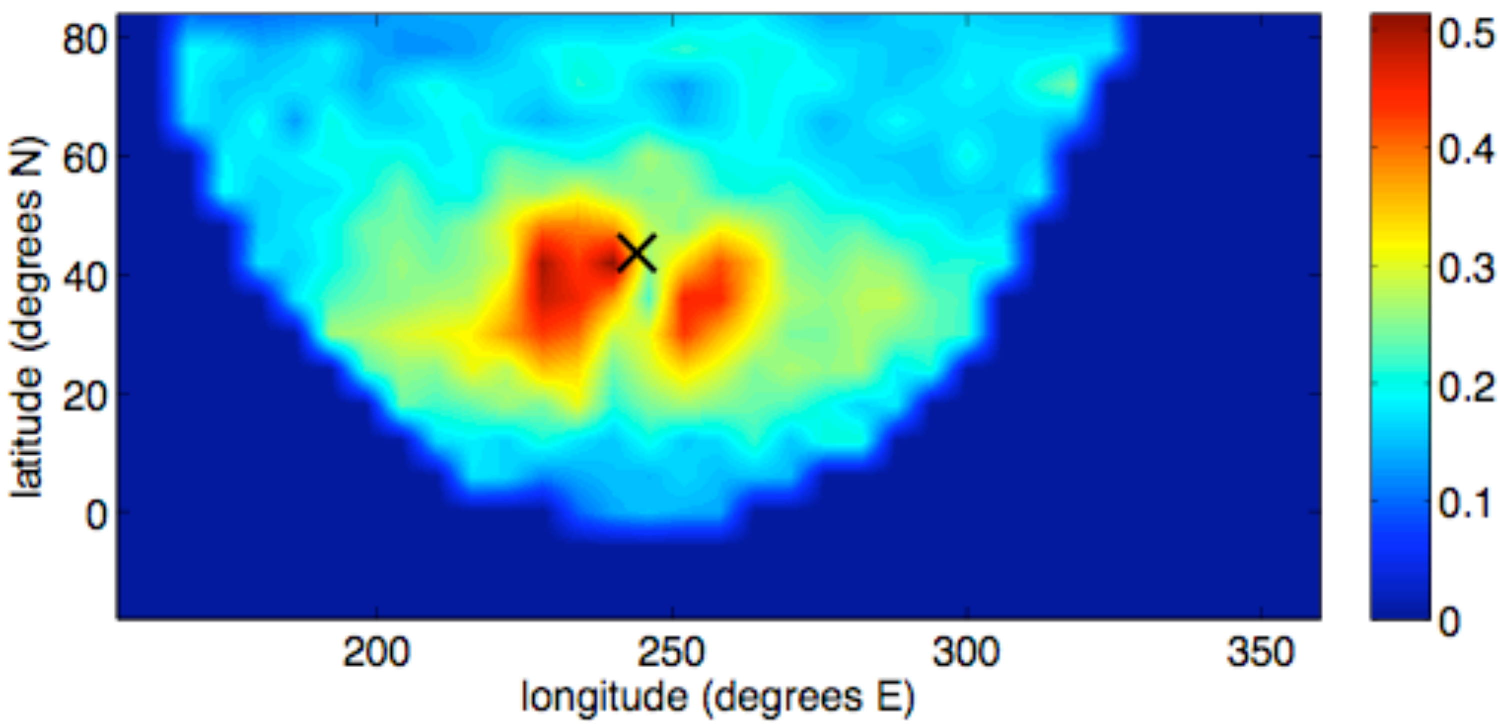
Can make small ensembles work very well.



Localization get complex in atmospheric models.

Localization for T obs. in mid-troposphere of dry AGCM core.

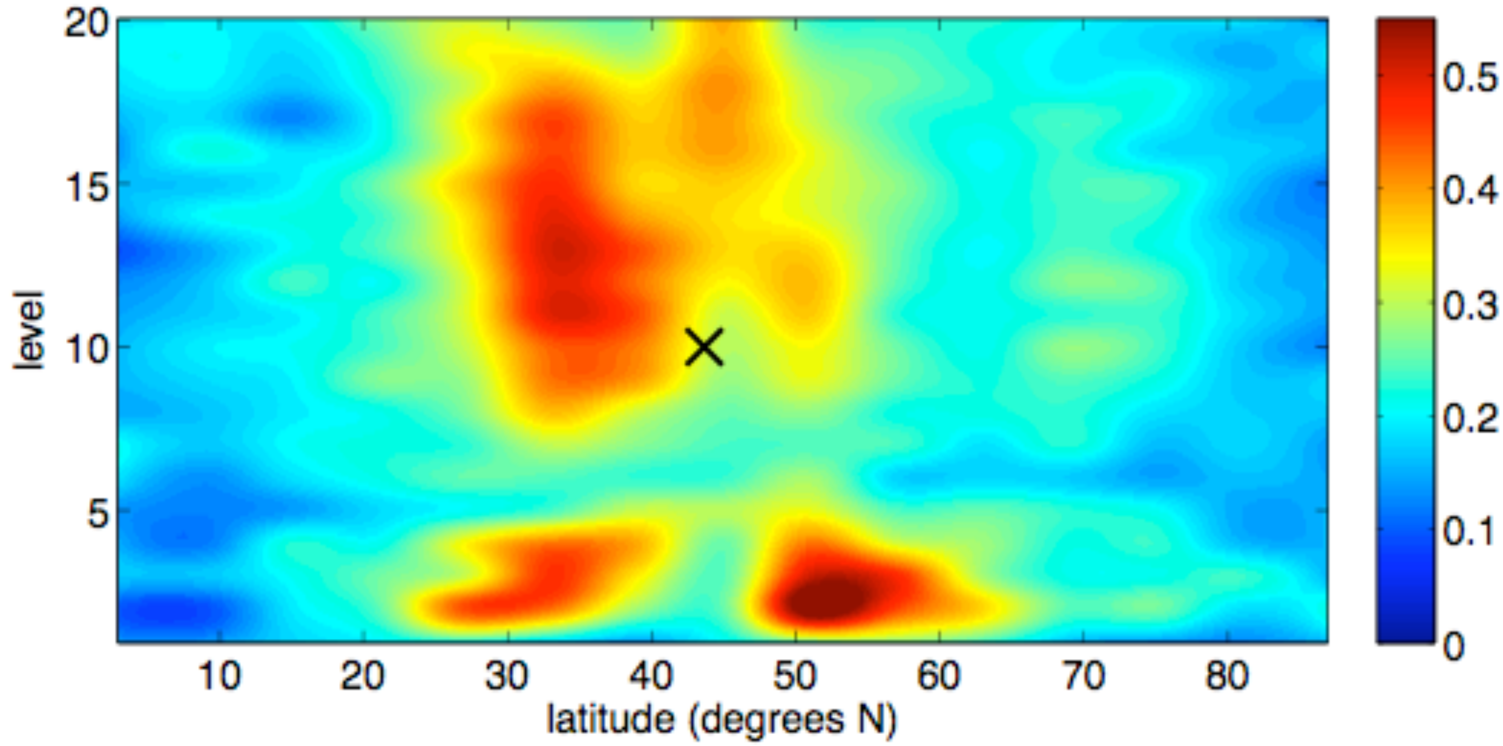
State variables are meridional wind components.



Localization get complex in atmospheric models.

Localization for U obs. in mid-troposphere of dry AGCM core.

State variables are temperature.



Model error reduces need for large ensembles

- If error is in mean, model will never sample it
- Have to correct errors by additional means.
- If error is in covariance, more confidence is a bad thing.

Localization remains biggest challenge/opportunity

- Remote correlations are only thing requiring large ensembles.
- No good theory, even in small, linear systems.
- Become non-linear when filter is applied.
- Gaussian univariate localization is sub-optimal.
- Lots of structure in statistically derived localizations.
- These work better, even in simple problems.

Questions:

- Can we estimate the minimum non-diverging ensemble size?
- Is there an efficient way to find good localization?
- Can small ensembles do nearly perfectly in large models?

GREAT PROBLEMS FOR GRAD. STUDENTS

Note: Non-linear filters would change things a lot, but…
too expensive?