Ensemble clustering in deterministic ensemble Kalman filters*

Javier Amezcua, Kayo Ide,
Craig Bishop and Eugenia Kalnay

* *Tellus A*, accepted May 2012.

DARC seminar, 4th July 2012, University of Reading

Outline

1. Ensemble Kalman filtering
    1. The ETKF family
2. Ensemble Clustering
3. A comprehensive study on Ensemble Clustering
4. Experiments
    1. Lorenz 1963 model
    2. More complicated models (with inflation, localization, etc.)
5. Summary

Outline

1. Ensemble Kalman filtering
    1. The ETKF family
2. Ensemble Clustering
3. A comprehensive study on Ensemble Clustering
4. Experiments
    1. Lorenz 1963 model
    2. More complicated models (with inflation, localization, etc.)
5. Summary

1. (Ensemble) Kalman filtering

Dynamical **model**

$$\mathbf{x}_t = f(\mathbf{x}_{t-1}) + \mathbf{w}_t$$

$$\mathbf{x} \in \Re^N$$

Observations

$$\mathbf{y}_t = h(\mathbf{x}_t) + \mathbf{v}_t$$

$$\mathbf{y} \in \Re^L$$

Model error

$$E(\mathbf{v}_t) = \mathbf{0}$$

$$Cov(\mathbf{v}_t) = \mathbf{Q}$$

Observation error

$$E(\mathbf{w}_t) = \mathbf{0}$$

$$Cov(\mathbf{w}_t) = \mathbf{R}$$

KF is **optimal** when:

- The **forecast** and **observation** operators are **linear**.
- The **errors** are **Gaussian**.

The **validity of these conditions** depends upon:

- **Length** of the **forecast window / frequency** of **observations**.
- **Magnitude** of **observational error**.
- **Nonlinearity** in the **model dynamics**.

1. Ensemble Kalman filtering

Updating the ensemble **mean** and **covariance** is **straightforward**, **updating** the perturbations **is not**.

$$\mathbf{X}^b \rightarrow \mathbf{X}^a$$

Deterministic EnSRF

- A **direct transformation** from background to analysis (not unique), can use **observations serially or all-at-once**.
- The KF **covariance equation** is **satisfied exactly**.
- Any **distortions** of the **ensemble** are prone to **persist**.

Stochastic EnKF

- Ensemble members are updated individually using **perturbed observations**.
- The **KF covariance equation** is **fulfilled** only **statistically**.
- The **ensemble** is constantly ‘**refreshed**’.

1. Ensemble Transform Kalman Filter family

Within the EnSRFs, the **ETKF** family relies on a **post-multiplication** to update perturbations **all-at-once**.

$$\mathbf{X}^a = \mathbf{X}^b\mathbf{W}^a, \mathbf{W}^a \in \Re^{M\times M} \qquad \mathbf{C}\boldsymbol{\Gamma}\mathbf{C}^{\mathrm{T}} = \left({\mathbf{Y}^b}^{\mathrm{T}}\mathbf{R}^{-1}\mathbf{Y}^b\right)/(\boldsymbol{M}-\mathbf{1})$$

- **One-sided** ETKF (Bishop et al., 2001) $\mathbf{W}^a = \mathbf{C}(\mathbf{I}+\boldsymbol{\Gamma})^{-\frac{1}{2}}$
- **Symmetric** ETKF (Wang et al., 2004; Hunt et al., 2007) $\mathbf{W}^a = \mathbf{C}(\mathbf{I}+\boldsymbol{\Gamma})^{-\frac{1}{2}}\mathbf{C}^{\mathrm{T}}$
- **No-symmetric** solutions (e.g. Sakov and Oke, 2008) $\mathbf{W}^a = \mathbf{C}(\mathbf{I}+\boldsymbol{\Gamma})^{-\frac{1}{2}}\mathbf{S}^{\mathrm{T}}$

The **one-sided ETKF is biased**, the **symmetric ETKF is unbiased**, for the **not symmetric ETKFs it depends upon the particular** (possibly random) **matrix S** (Livings et. al, 2008).

Outline

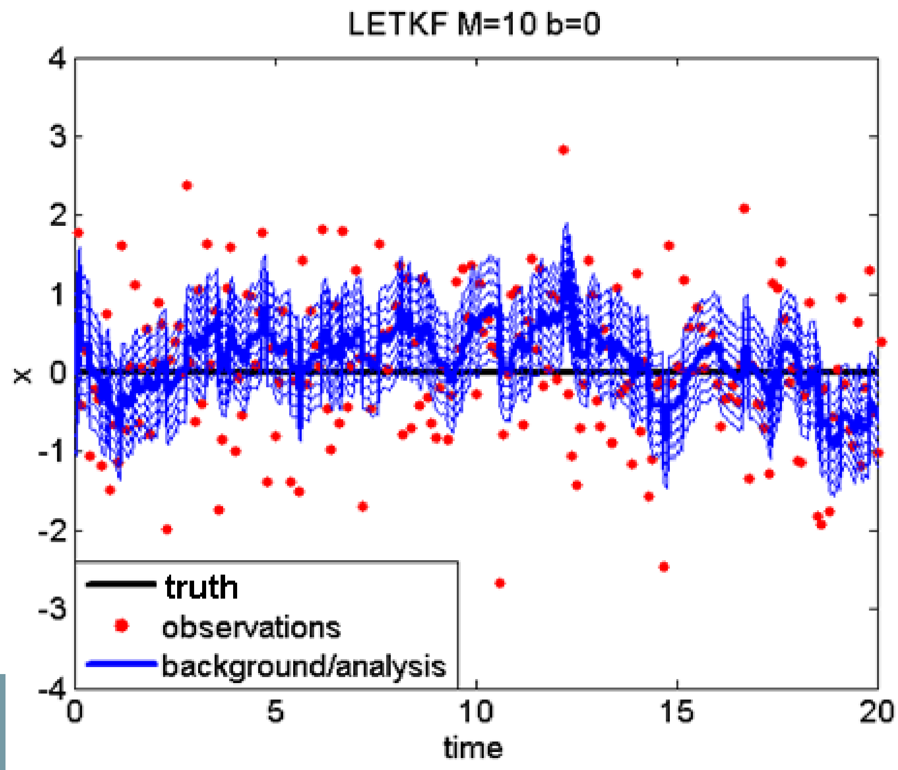
1. Ensemble Kalman filtering
    1. The ETKF family
2. Ensemble Clustering
3. A comprehensive study on Ensemble Clustering
4. Experiments
    1. Lorenz 1963 model
    2. More complicated models (with inflation, localization, etc.)
5. Summary

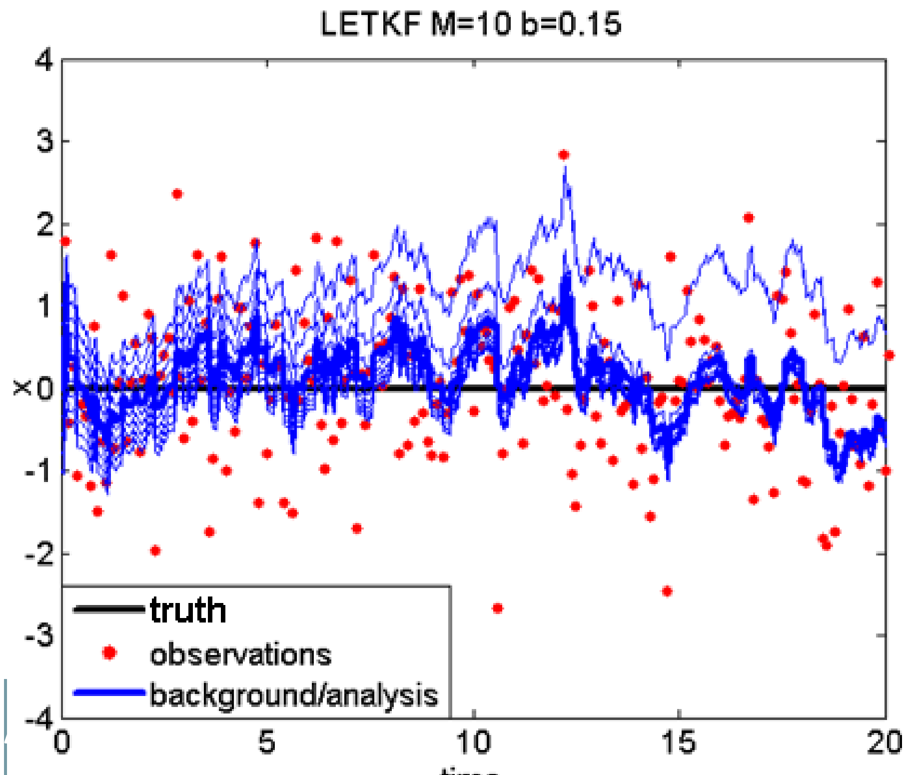
2. Ensemble clustering

Consider the **univariate quadratic model** (Anderson, 2010):

$$x_{t+1} = x_t + 0.05\left(x_t + \underline{b}\,|x_t|\,x_t\right)$$

It has an **unstable fixed point**; we use it as truth $x^t = 0$. The model is integrated with $\Delta t = 0.01$ and we observe ($\mathbf{H} = \mathbf{I}$) every 2 steps. We use S-ETKF for a **linear** $(b = 0)$ and a **nonlinear** $(b = 0.15)$ case, $M = 10$.



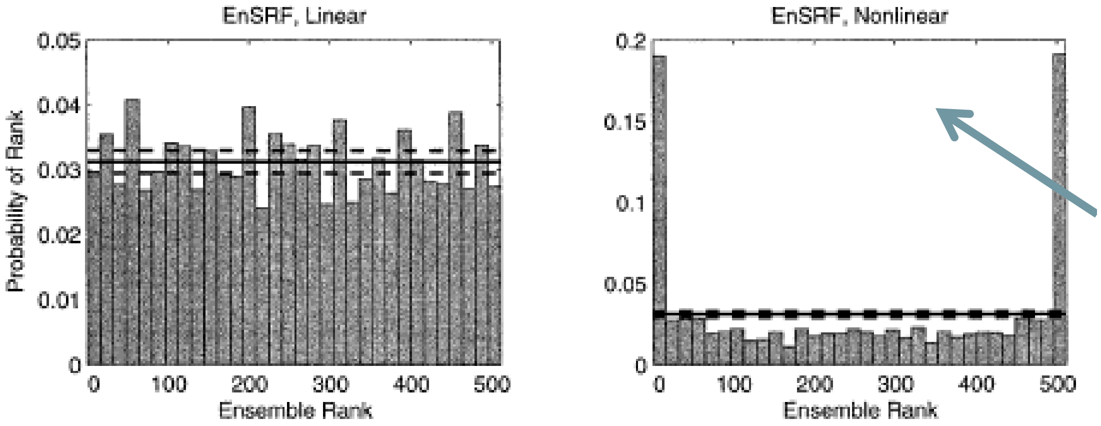


2. Ensemble clustering

As soon as nonlinearity strikes EC appears in EnSRFs. It **does not happen in the stochastic EnKF.** It results from the **disparity of nonlinear forecast and linear analysis** (Anderson, 2010).

This has been studied in the Ikeda model (Lawson and Hansen, 2004), the Lorenz 1963 and 1996 models (Anderson, 2010) with **infrequent observations** and large **observational errors**.

It **does not affect the ensemble covariance**, but it does affect higher order moments.



The **truth is not statistically indistinguishable from the ensemble members** (from Lawson and Hansen, 2004).

2. Questions about EC

a) Is there a **simple way** to **diagnose** it?

b) Is it an **irreversible** phenomenon of EnSRFs?

c) How much does it **affect** the **accuracy** of EnSRFs? Does it **handicap** them?

d) Alternatives can be used to avoid it (e.g. **Non Symmetric ETKFs**, Anderson's Rank Histogram Filter). Are they **advantageous**?

Outline

1. Ensemble Kalman filtering
    1. The ETKF family
2. Ensemble Clustering
3. A comprehensive study on Ensemble Clustering
4. Experiments
    1. Lorenz 1963 model
    2. More complicated models (with inflation, localization, etc.)
5. Summary

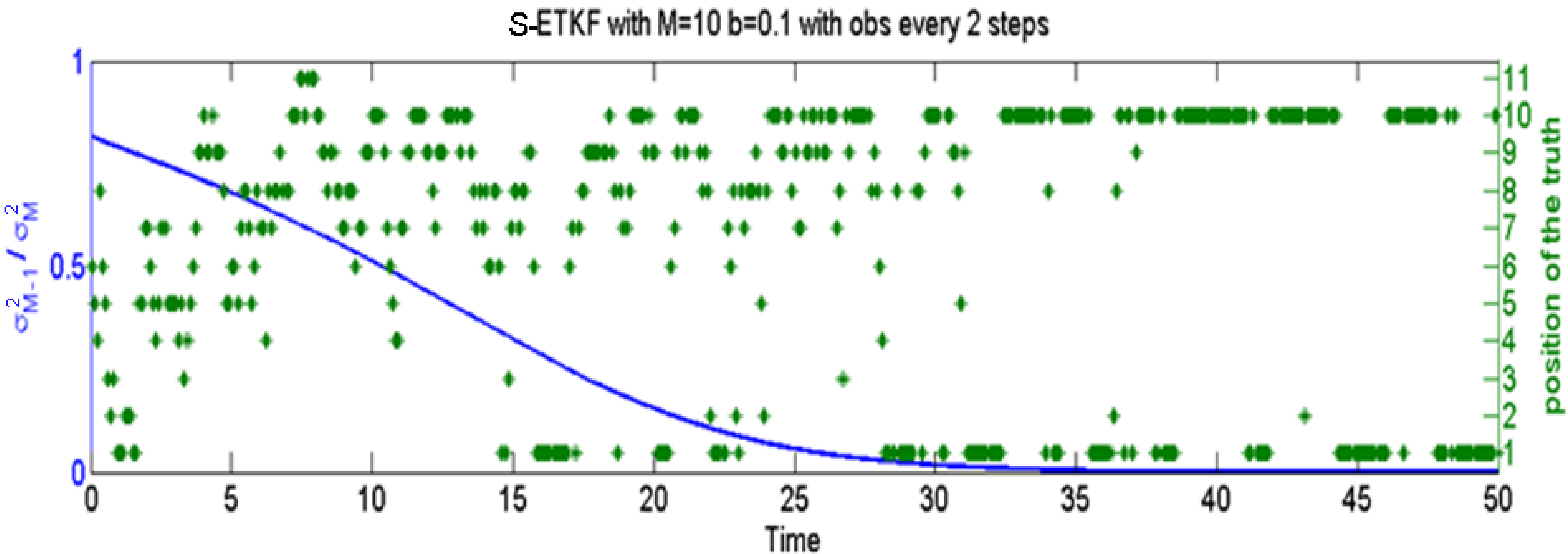
3. Measuring EC

We use the following **metric**, which we denominate '**clustering degree**'.

$$CD = \frac{\sigma^2_{M-1}}{\sigma^2_M} \qquad CD = \frac{Trace(P_{M-1})}{Trace(P_M)}$$

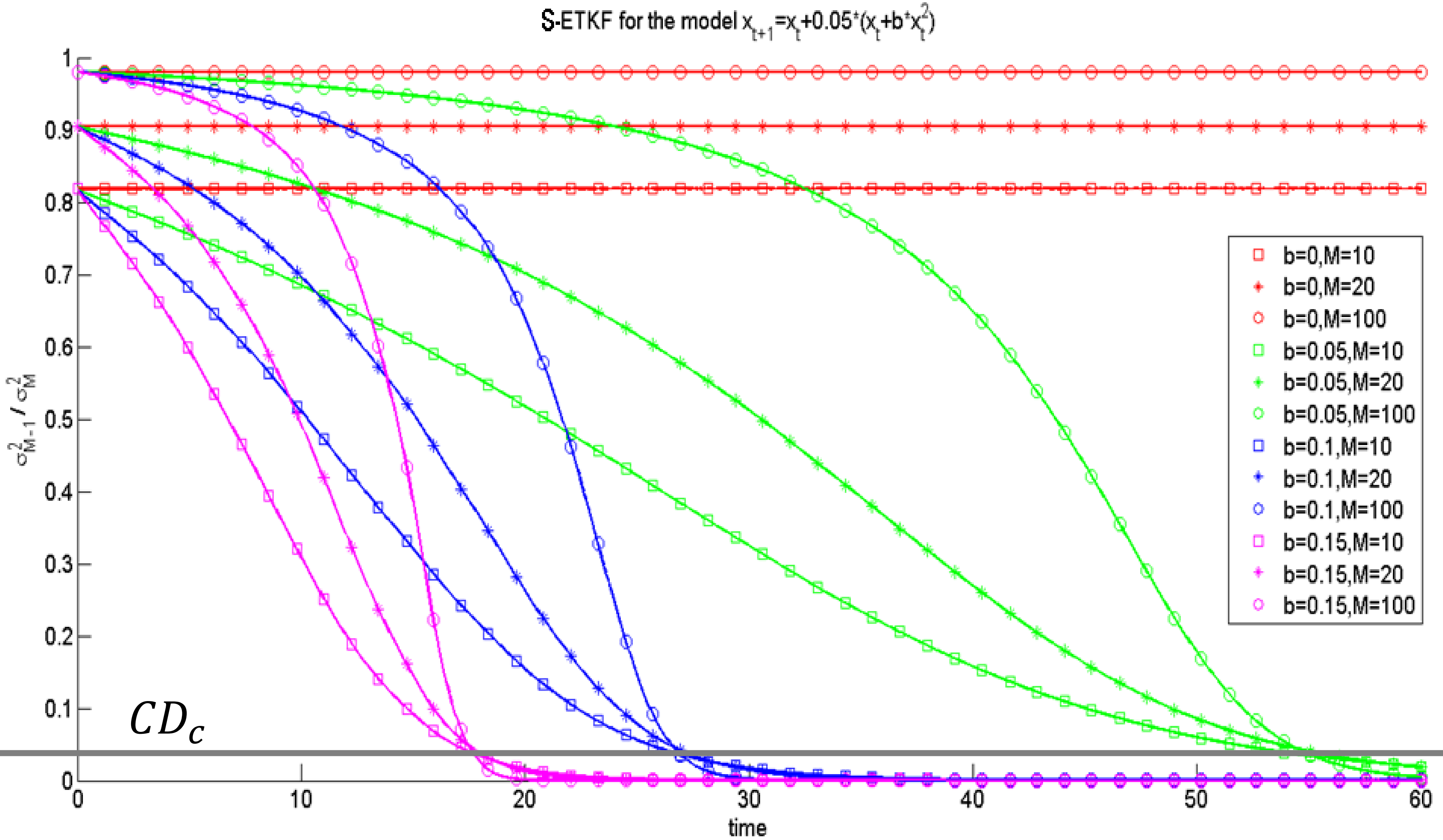
$$N = 1 \qquad N > 1$$

Considering again the model: $x_{t+1} = x_t + 0.05(x_t + b|x_t|x_t)$



3. Measuring EC

Varying the **ensemble size** and the **strength of the nonlinearity**:

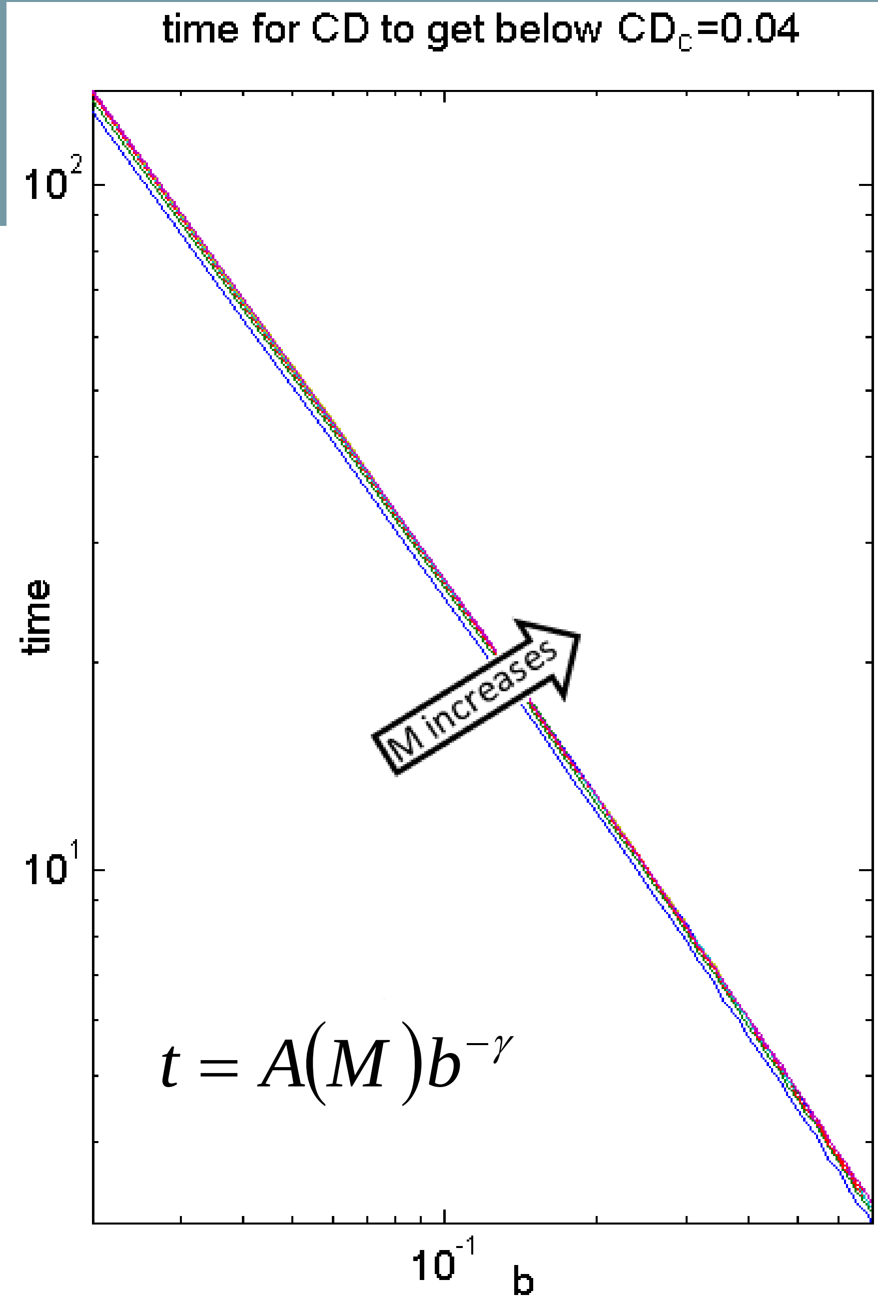


Outline

How **fast** does EC occur?

In this simple model it seems to follow a **power law** independent from *b* and weakly dependent on *M*.

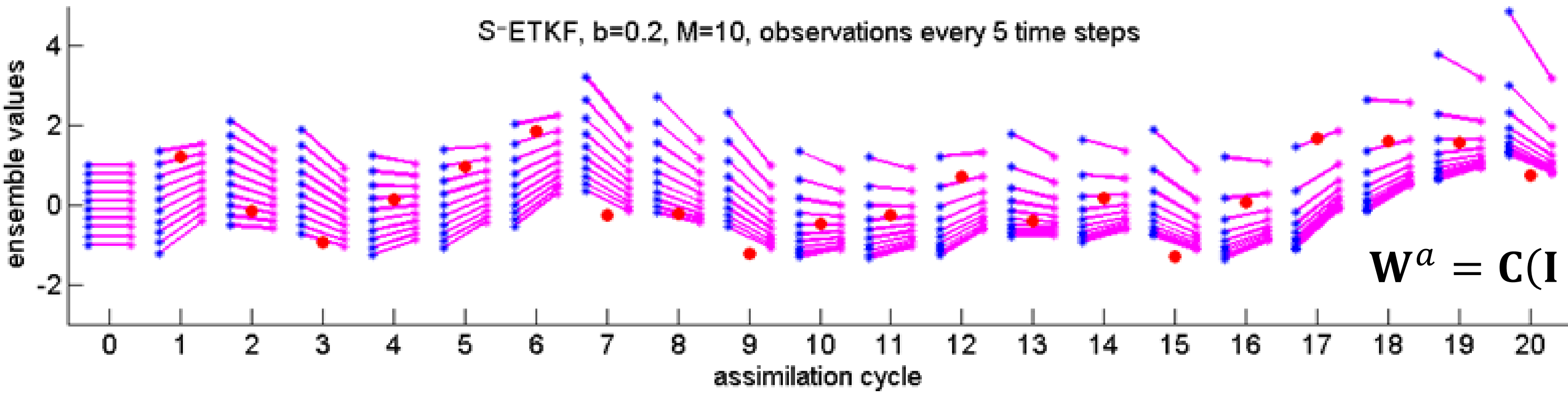
In **this case, clustering is inevitable.** Is this **always the case?**



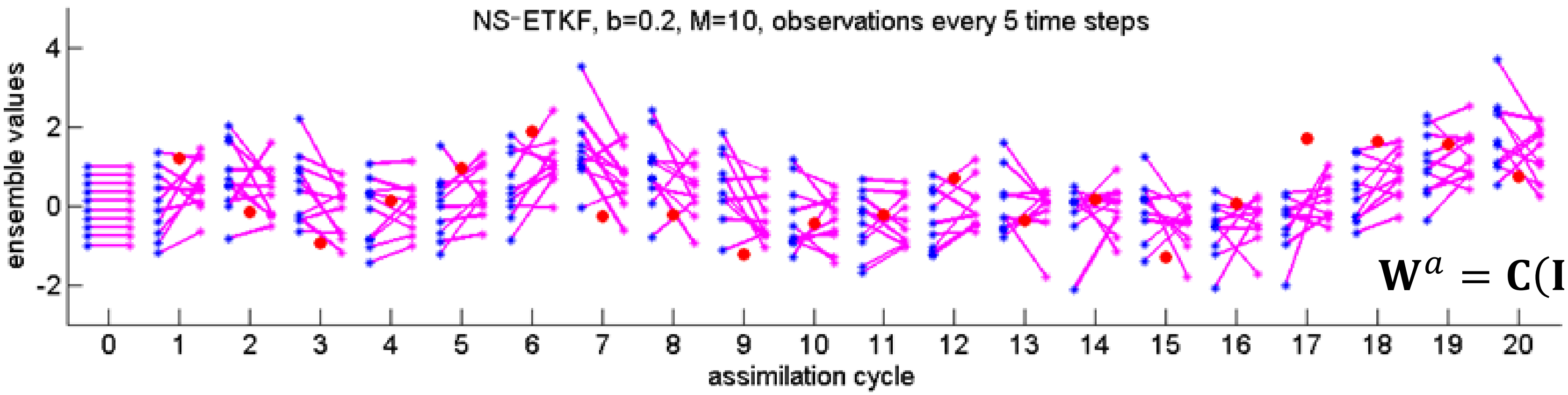
$$t = A(M)b^{-\gamma}$$

3. Avoiding EC: NS-ETKFs

(Unbiased) **randomly-rotated EnSRFs** **avoid clustering**.



$$\mathbf{W}^a = \mathbf{C}(\mathbf{I}+\boldsymbol{\Gamma})^{-\frac{1}{2}}\mathbf{C}^{\mathrm{T}}$$

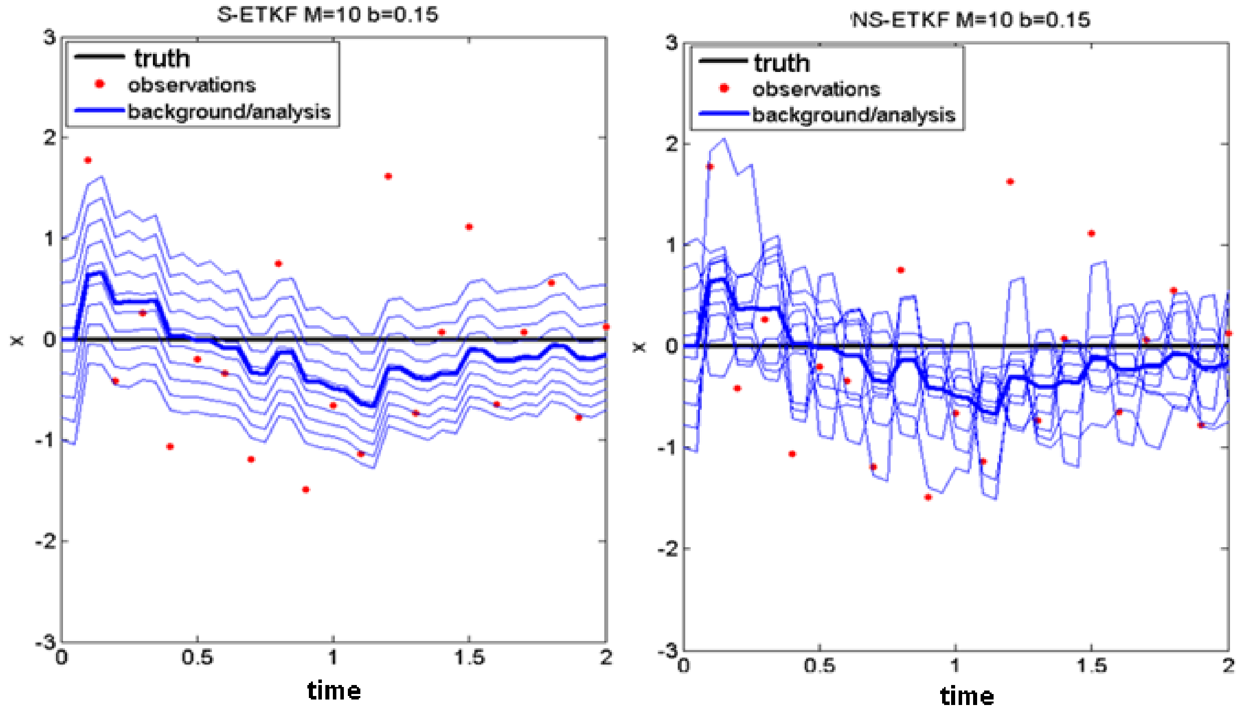


$$\mathbf{W}^a = \mathbf{C}(\mathbf{I}+\boldsymbol{\Gamma})^{-\frac{1}{2}}\mathbf{S}^{\mathrm{T}}$$

The **constant ‘scrambling’ of the ensemble** prevents the outlier from being persistent and eventually ‘escaping’.

3. Avoiding EC: NS-ETKFs

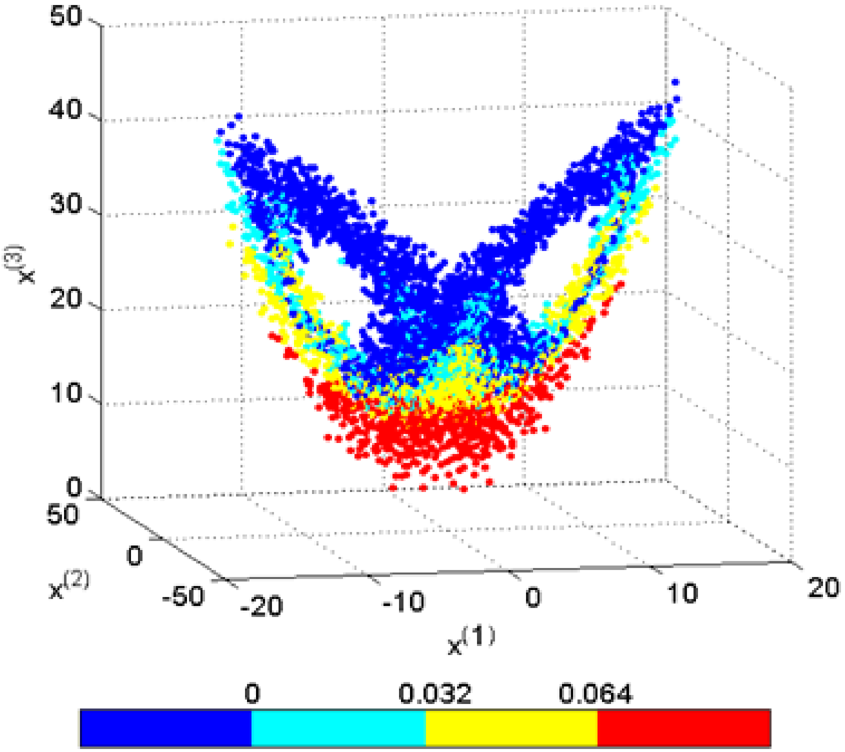
However, this **constrained resampling** of the ensemble **erases** the memory from **individual trajectories**, the effect of the '**errors of the day**' is modified. It is like "**rebooting**" at each analysis instant.



Following individual trajectories was one of the advantages of EnSRFs (Anderson, 2001). **Is it worth losing this ability**?

3. 'Local' nonlinearities

The **growth/decay perturbations is not constant** (neither linearly nor nonlinearly) **in an attractor**.



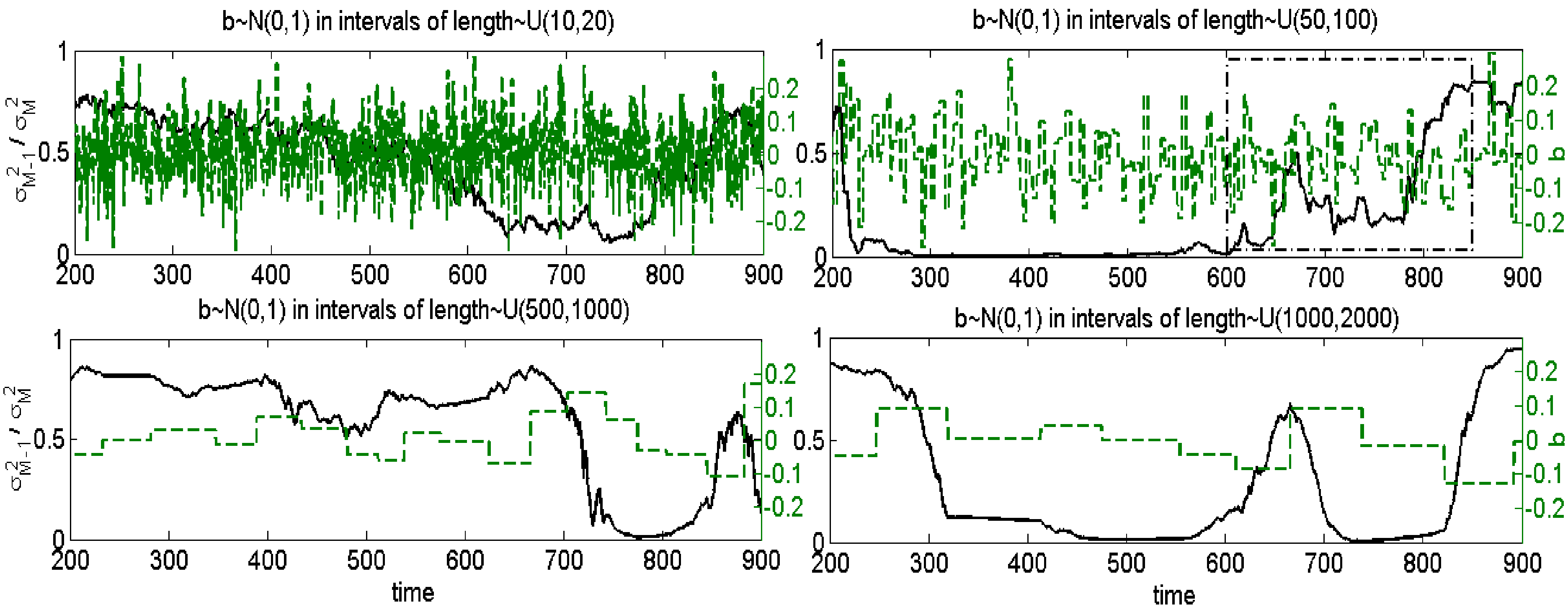
Reproduced from Evans *et al.*, 2004.

Bred vector growth in the Lorenz 1963 model showing the **growth rate for perturbations**. For different regions, there can be **decay** or growth (**slow**, **moderate**, **fast**).

3. ‘Local’ nonlinearities

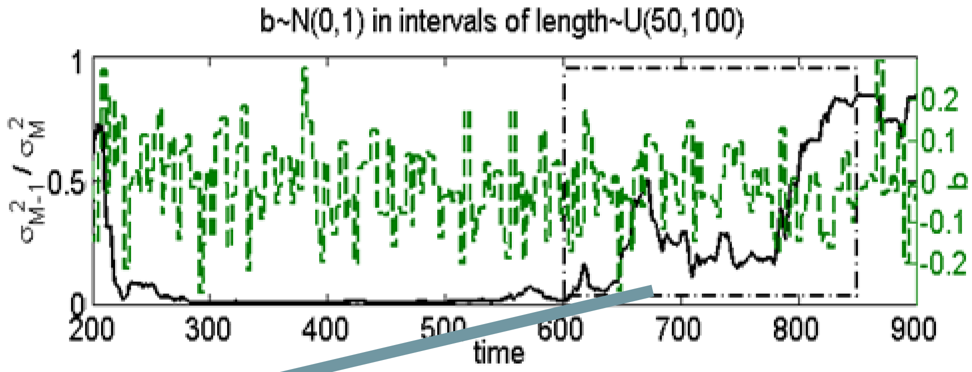
$$x_{t+1} = x_t + 0.05\left(x_t + b|x_t|x_t\right)$$

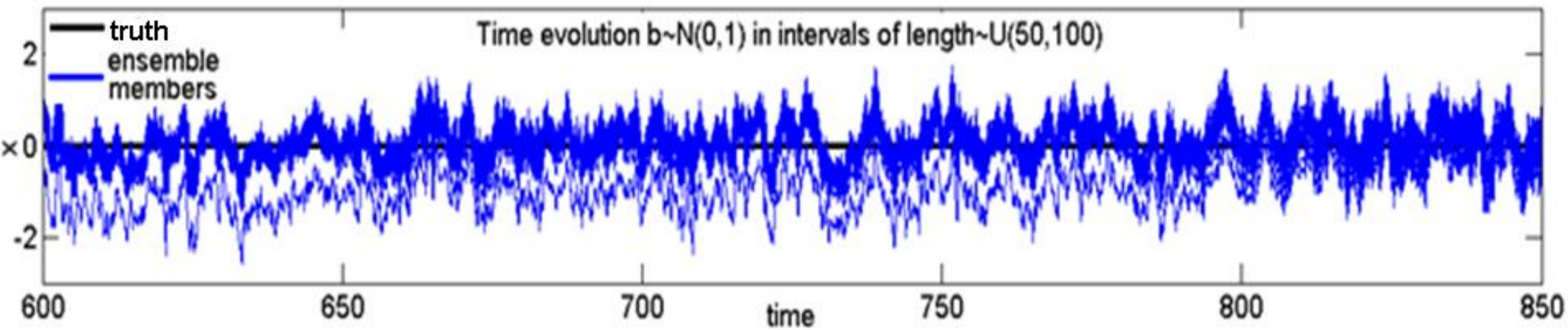
Let’s draw **a new** $b \sim N(0,1)$ **every** $L \sim Unif\left(L_0, 2L_0\right)$ **steps**, perform DA in this model, and measure the clustering degree.



3. 'Local' nonlinearities

$$x_{t+1} = x_t + 0.05\left(x_t + b|x_t|x_t\right)$$





Clustering can be **reverted** by the **alternation of nonlinear growth and decay**. It is an **intermittent phenomenon**.

Outline

1. Ensemble Kalman filtering
   1. The ETKF family
2. Ensemble Clustering
3. A comprehensive study on Ensemble Clustering
4. Experiments
   1. Lorenz 1963 model
   2. More complicated models (with inflation, localization, etc.)
5. Summary

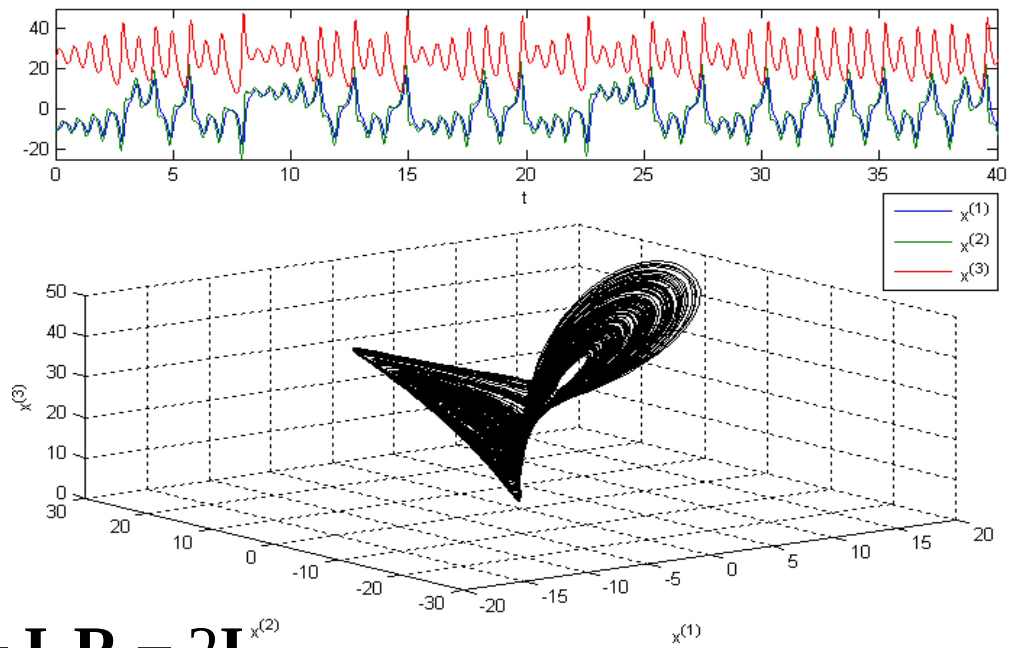
4. EC in the Lorenz 1963 (L63) model

The system:

$$\dot{x}^{(1)} = \sigma\left(x^{(2)} - x^{(1)}\right) \qquad \sigma = 10$$

$$\dot{x}^{(2)} = x^{(1)}\left(r - x^{(3)}\right) - x^{(2)} \qquad r = 8/3$$

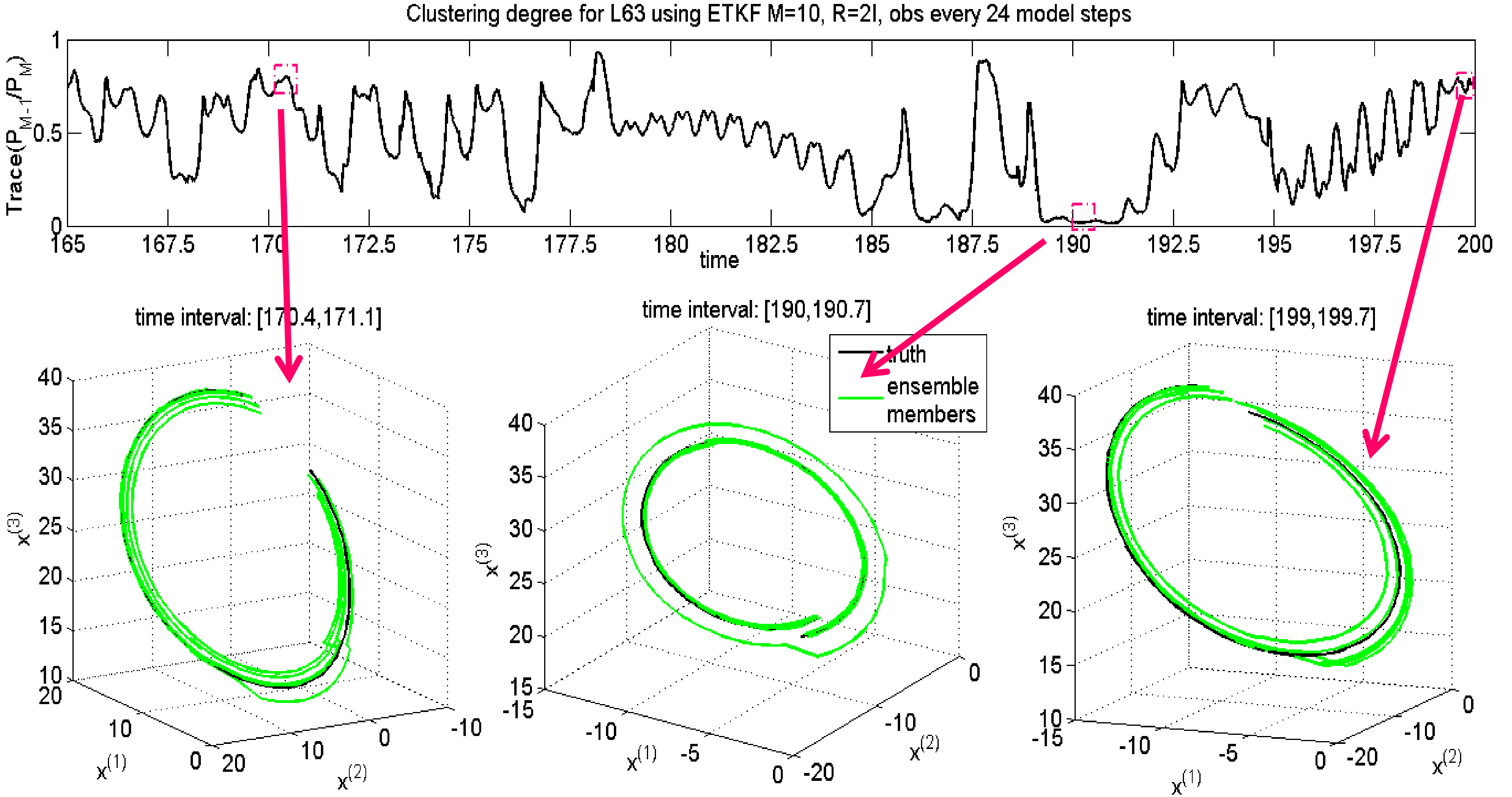
$$\dot{x}^{(3)} = x^{(1)}x^{(2)} - bx^{(3)} \qquad b = 28$$



Settings (Miller et al., 1997; Kalnay et al., 2007): $\Delta t = 0.01, \mathbf{H} = \mathbf{I}, \mathbf{R} = 2\mathbf{I}$

- "**Frequent**" **observations**: every **8 steps**, **linear** regime.
- "**Infrequent**" **observations**: every **24 steps**, **nonlinear** regime.

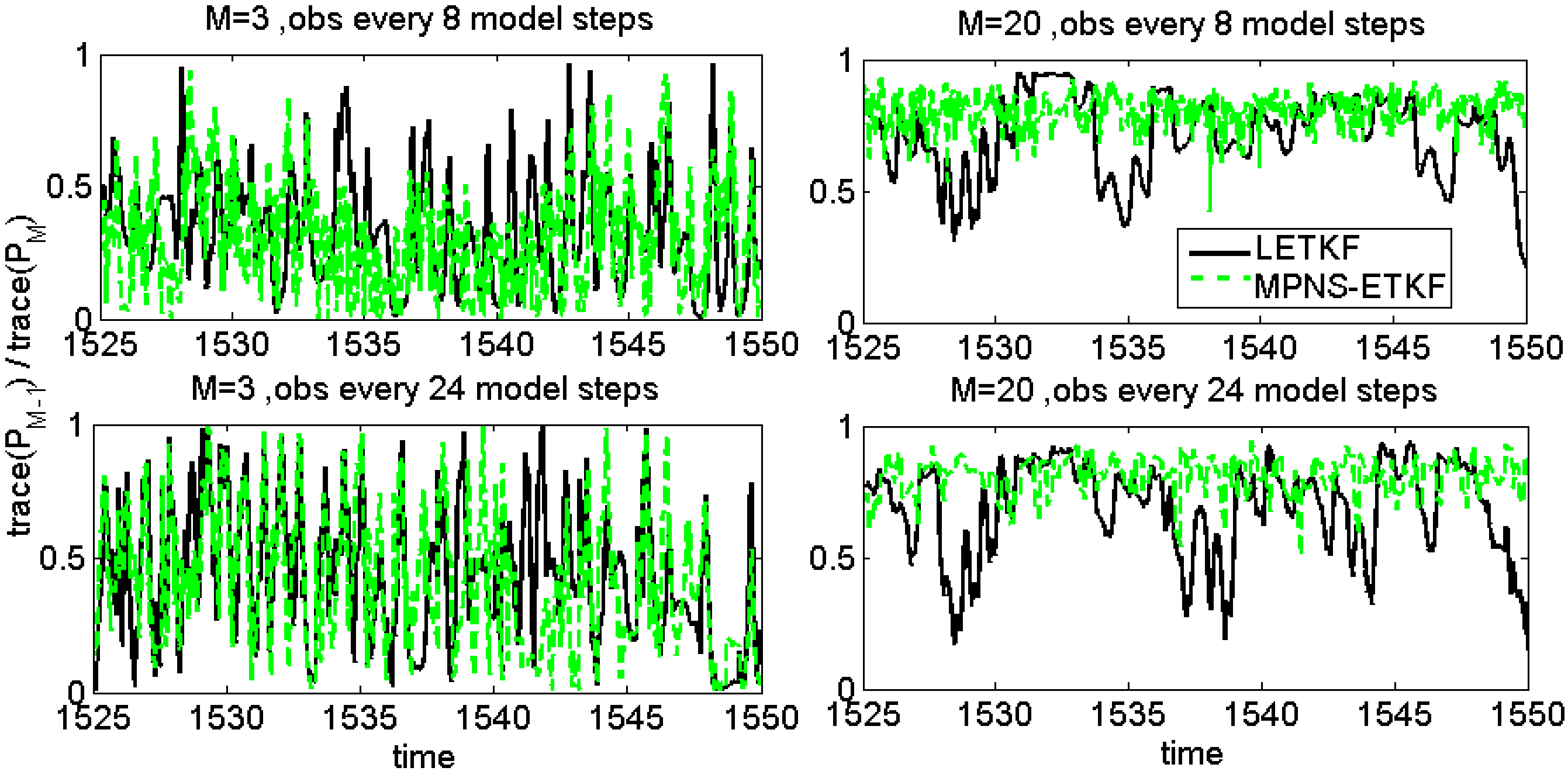
4. EC in L63



Clustering is intermittent.

4. EC in L63

As usual, **the Non-Symmetric ETKF does not present clustering**.



4. EC in L63

Clustering is **intermittent**, and **less persistent than in the univariate quadratic model**. Why?

In the **univariate model**, only the **magnitude of *b* could vary**. Plus, this model didn't present mixing.

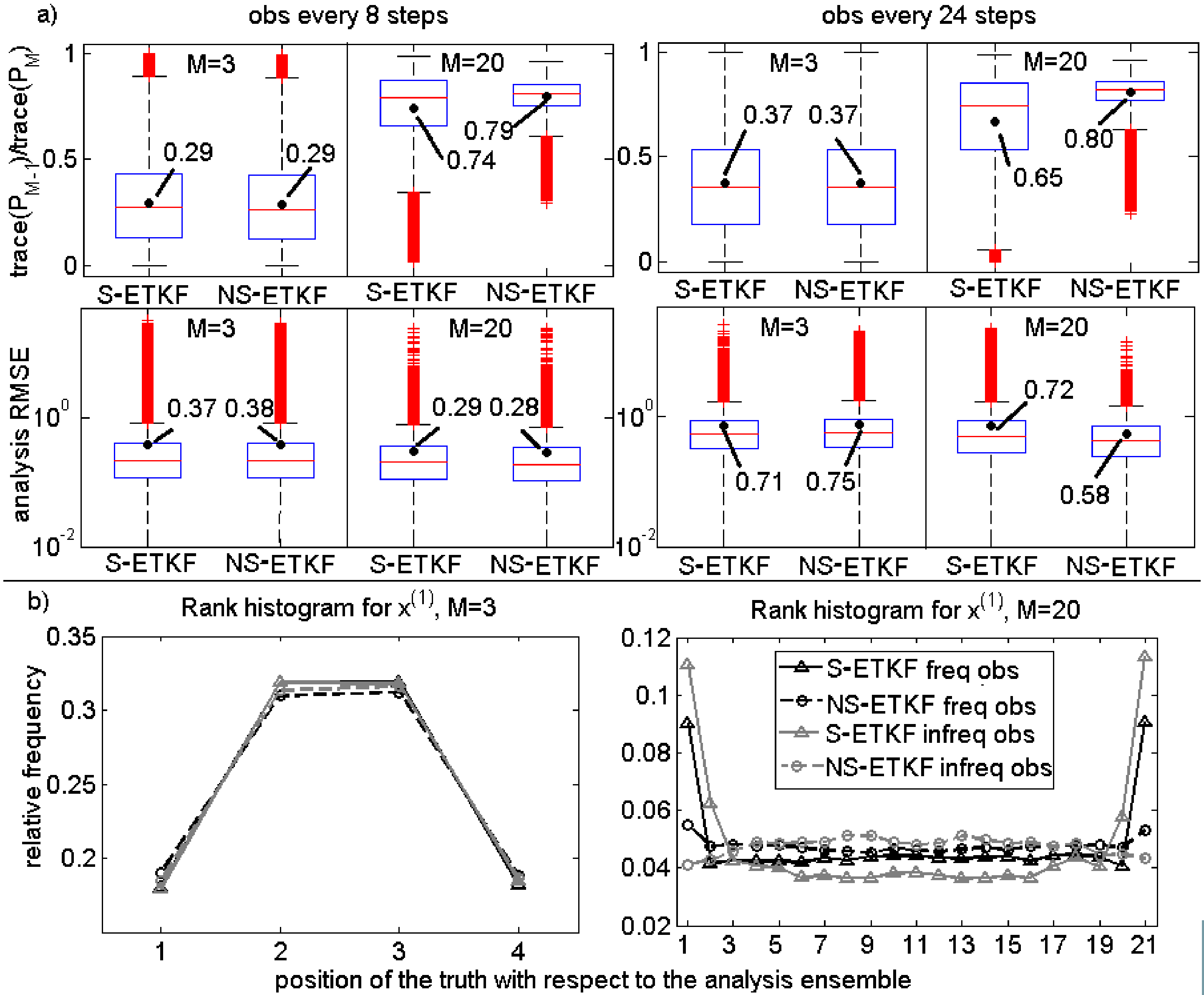
In the **Lorenz 1963 model**, both the **direction and magnitude of the nonlinear growth can vary**. Besides, this model presents mixing.

4. EC in L63

A **statistical summary** with 2 ensemble sizes (*M=3,10*).

Clustering affects the performance of S-ETKF (in terms of RMSE) **for large size ensembles**.

Random rotations **improve** the **rank histograms**, but **not when inflation is used** (*M=3*).



4. Larger models: localization

Larger models require **localization**. A natural choice for the **ETKF family** is **grid-point R localization**.

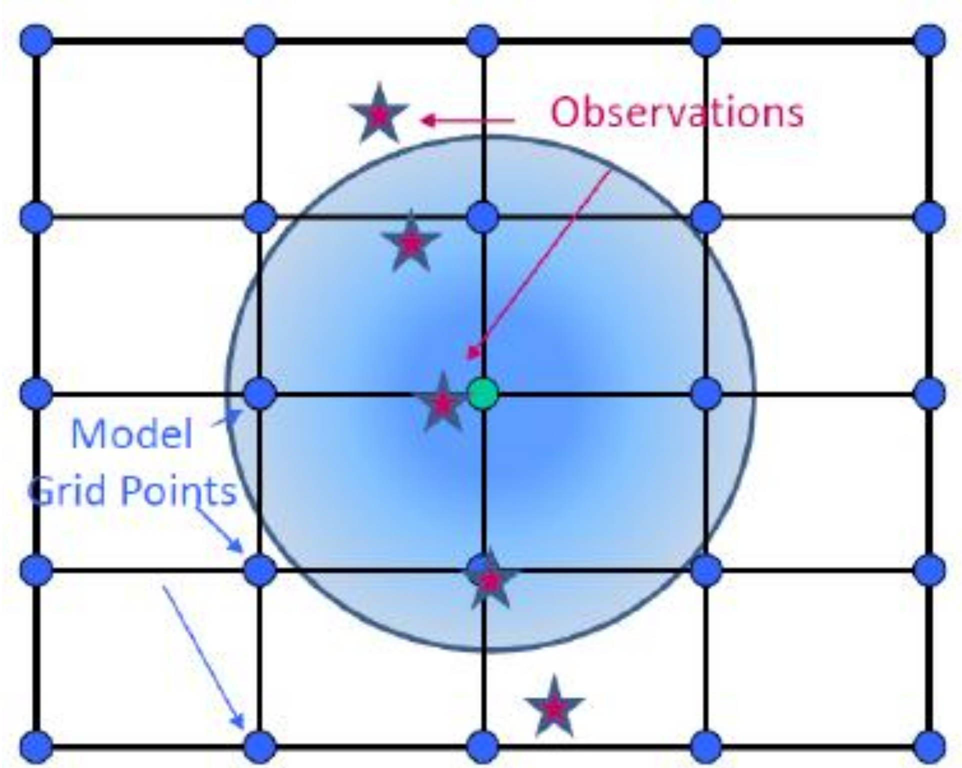


*Figure courtesy of Steven Greybush.

An **independent analysis** is carried out for **each grid-point** considering the neighboring observations.

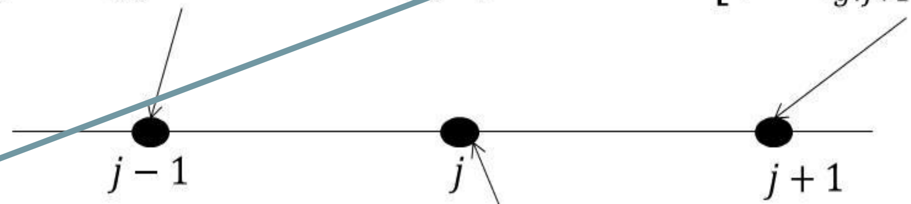
The **analysis ensemble** is constructed by **sets of rows**.

R-localization allows for **spatially varying adaptive inflation** (Miyoshi, 2011).

4. Larger models: localization

$$\mathbf{W}_{g:j-1}=\begin{bmatrix}(w_{1,1})_{g:j-1} & \cdots & (w_{1,M})_{g:j-1}\\ \vdots & \ddots & \vdots\\ (w_{M,1})_{g:j-1} & \cdots & (w_{M,M})_{g:j-1}\end{bmatrix}$$

$$\mathbf{W}_{g:j+1}=\begin{bmatrix}(w_{1,1})_{g:j+1} & \cdots & (w_{1,M})_{g:j+1}\\ \vdots & \ddots & \vdots\\ (w_{M,1})_{g:j+1} & \cdots & (w_{M,M})_{g:j+1}\end{bmatrix}$$



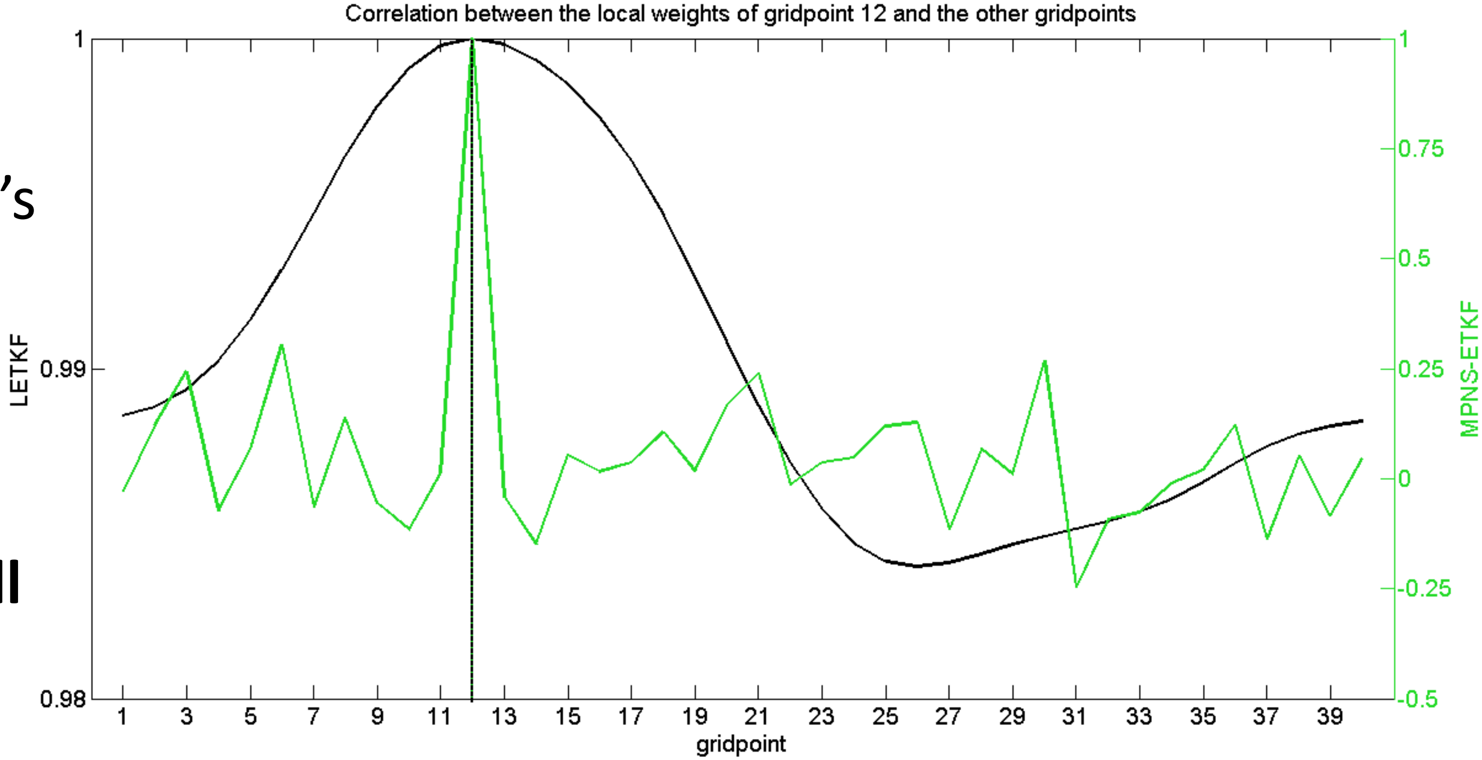
$$\mathbf{W}_{g:j-1}=\begin{bmatrix}(w_{1,1})_{g:j} & \cdots & (w_{1,M})_{g:j}\\ \vdots & \ddots & \vdots\\ (w_{M,1})_{g:j} & \cdots & (w_{M,M})_{g:j}\end{bmatrix}$$

The **transition between these matrices must be as smooth as possible** from one point *j* to the neighbors *j-1, j+1*

This was one of the reasons why the **symmetric square root was used in LETKF** (Hunt et al., 2007).

4. Larger models: localization

Using the **40-variable Lorenz 1996 model**, let's compute the **correlation among the weights of 1 gridpoint and all others.**

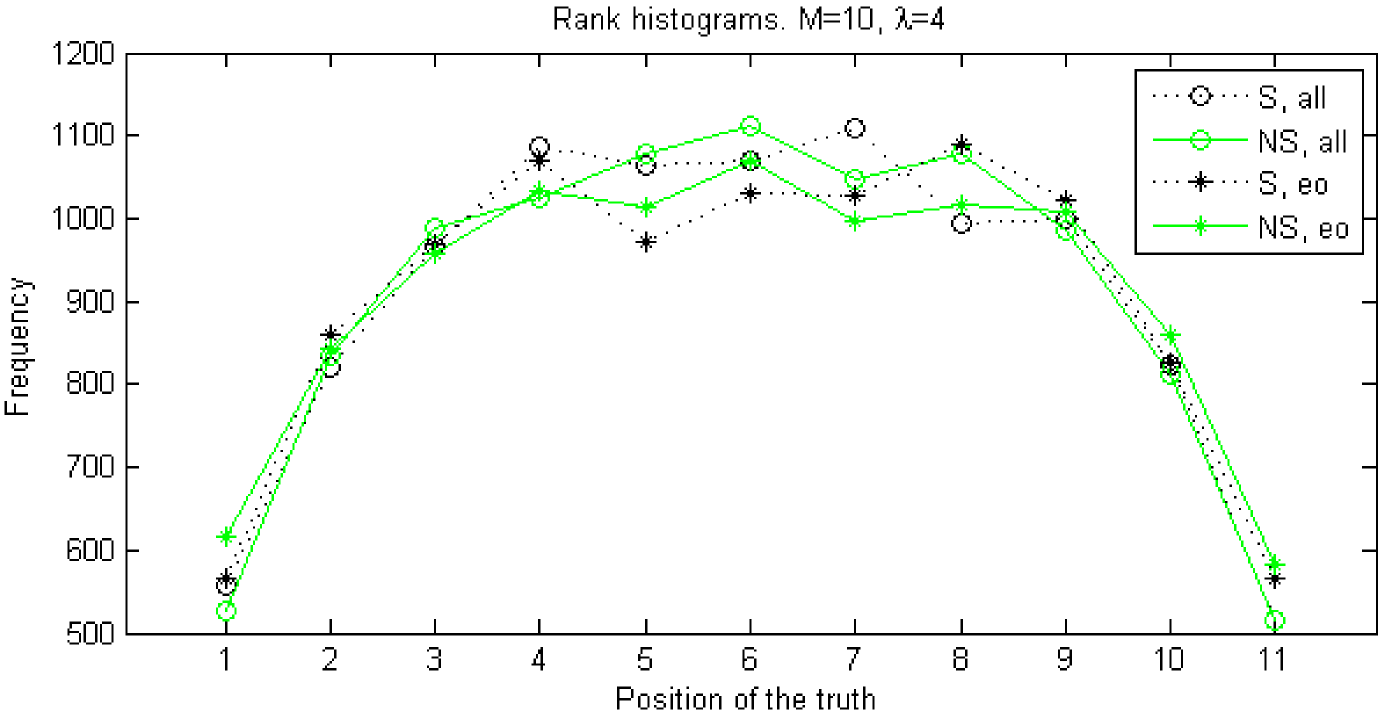


The randomly rotated ETKF **cannot be applied directly**, since a smooth transition among weight matrices (**W**) of neighboring gridpoints is not guaranteed.

4. Larger models: localization/inflation

Using **L96**, we perform experiments **observing** (a) **all variables** and (b) every **other variable** using **R-localization and adaptive inflation**. The observations are taken every 2 model steps ($\mathbf{R} = \mathbf{I}$).

No **perceivable difference in analysis RMSE** is noted. What happens with the **rank verification of truth**?

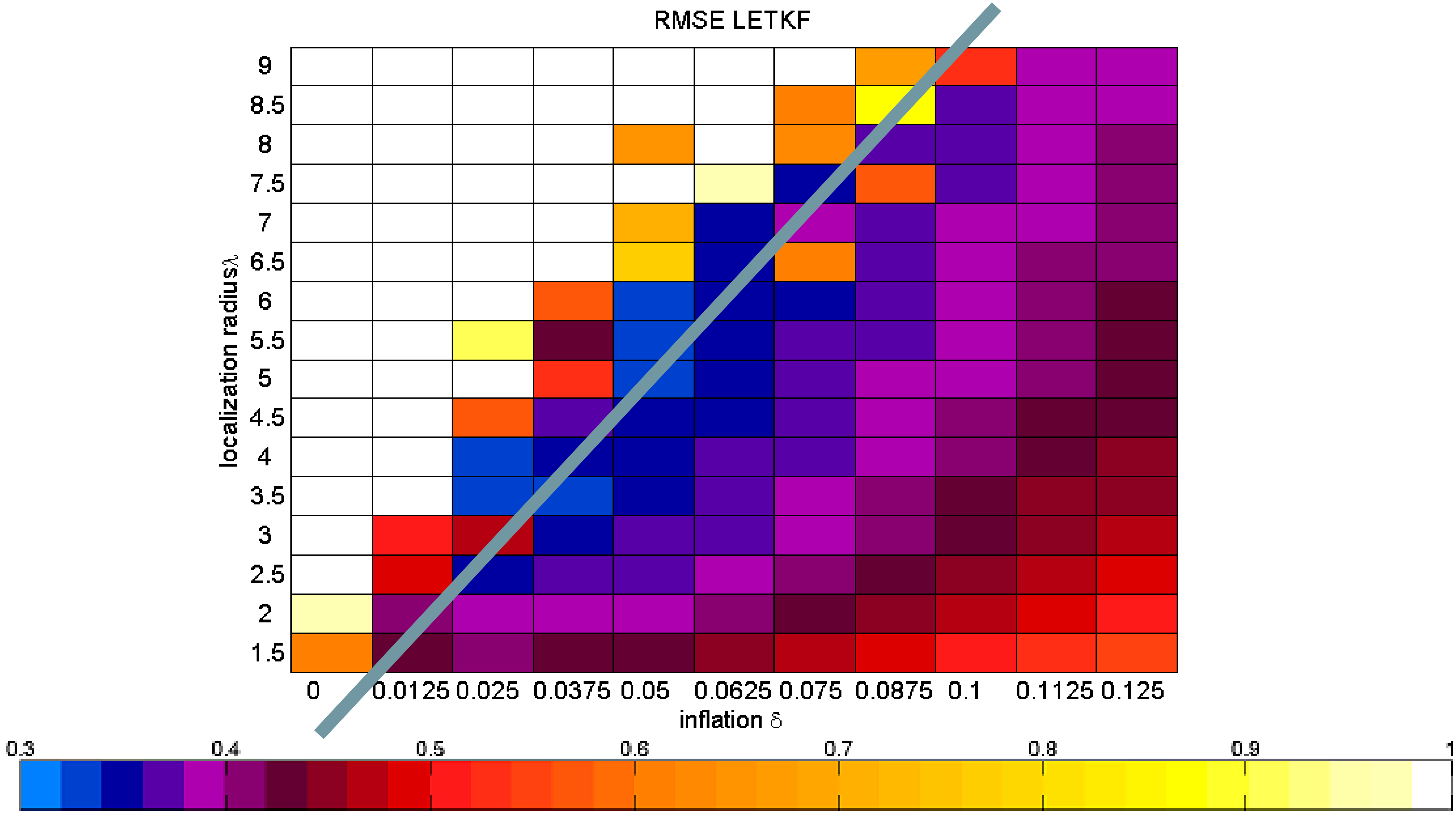


Inflation can lead to **over-dispersive ensembles** in all cases for these settings.

4. Larger models: localization/inflation

The **inflation is adaptive**. Why are the **ensembles overdispersive**?

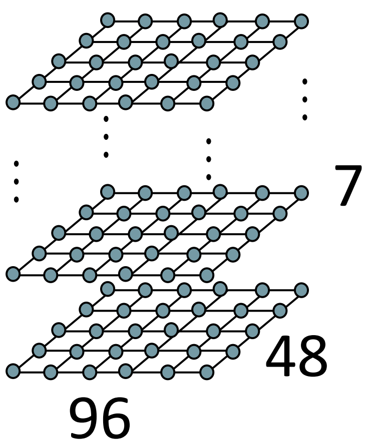
It is tough to find the **optimal inflation**, it **is close to filter divergence**.

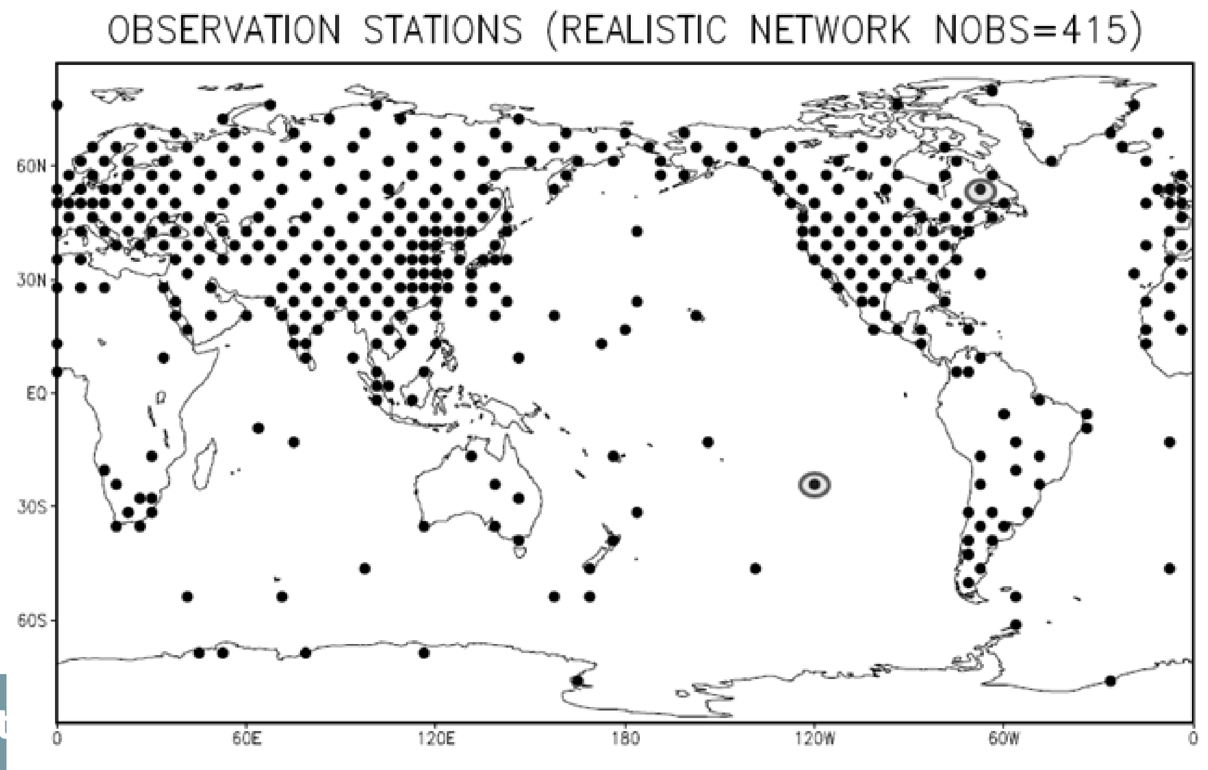


3. Larger models: a simplified AGCM

SPEEDY (Simplified Parameterizations, primitivE-Equations Dynamics, Molteni 2003)

- Time step is 40 minutes.
- Model variables: *u, v, T, q, ps*
- Spectral model with T30L7 resolution using σ-coordinates.

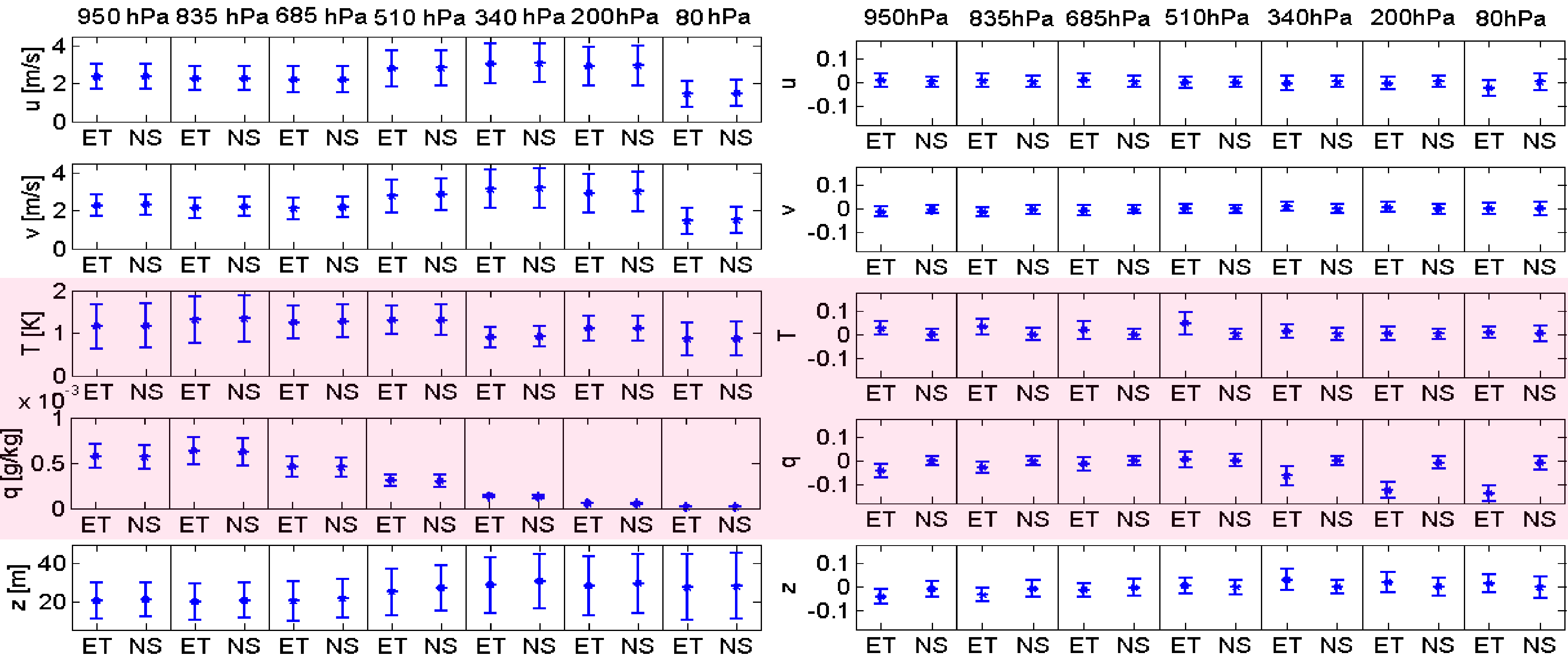




3. Larger models: a simplified AGCM

Computing **analysis RMSE and sample skewness** for all variables.

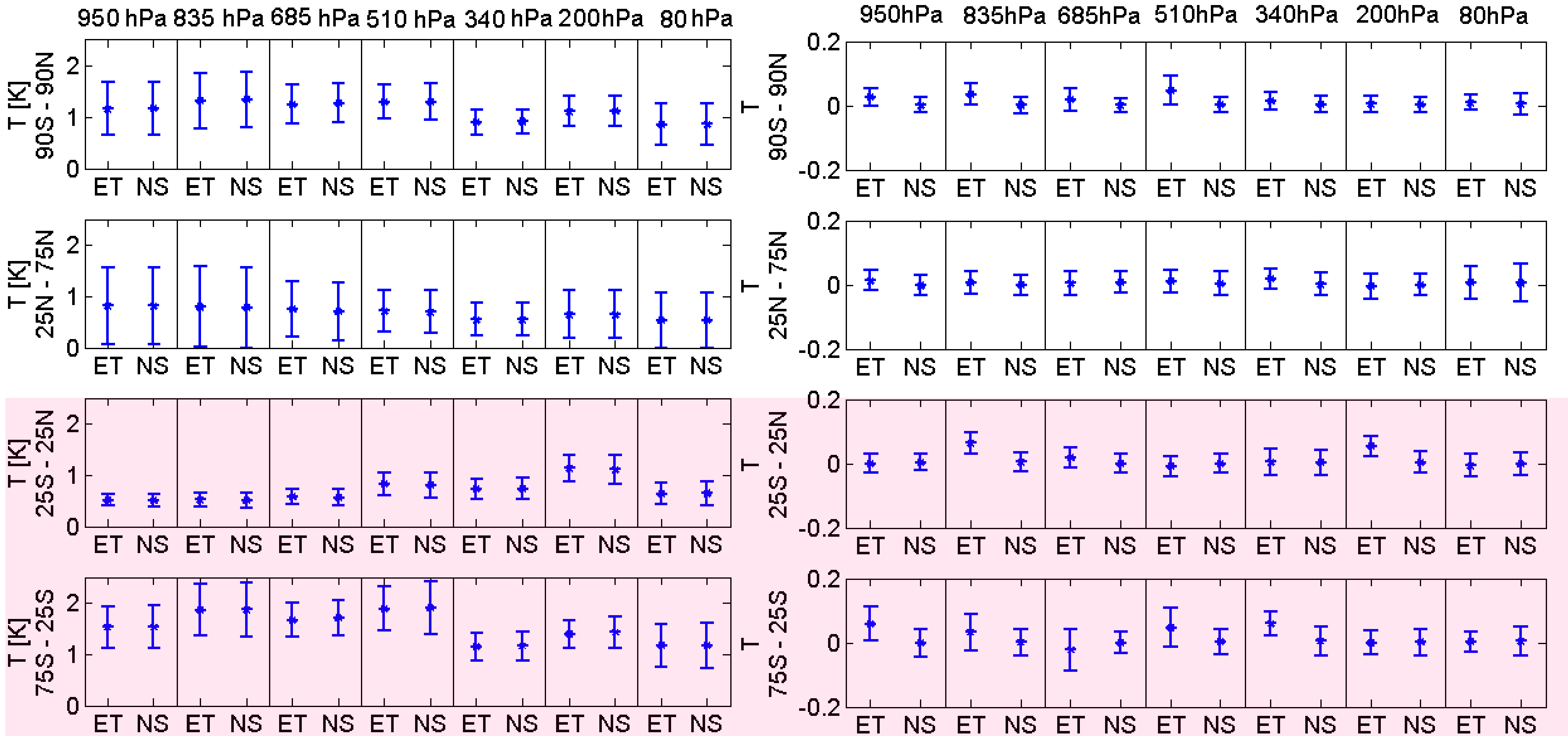
2 months of experiments, *M=20*, R-localization and adaptive inflation.



For some variables (***T* and *q***) we get **asymmetric ensembles**.

3. Larger models: a simplified AGCM

SPEEDY: **What happens with *T*?**



Asymmetric ensembles mainly in the **tropics** and in the **SH**. This **does not seem to affect RMSE**.

Outline

1. Ensemble Kalman filtering
    1. The ETKF family
2. Ensemble Clustering
3. A comprehensive study on Ensemble Clustering
4. Experiments
    1. Lorenz 1963 model
    2. More complicated models (with inflation, localization, etc.)
5. Summary

5. Summary

- **Clustering** is **not** an **irreversible** phenomenon of (deterministic) EnSRFs. It is **intermittent**.

- The **local** (in time and space) **nonlinear** **expansion/contraction** of the **ensemble** **triggers/reverses** **clustering**.

- As the **model** **grows**, the **persistence** **of clustering is** **shorter**.

- We only found **EC affecting** the performance (in terms of **RMSE**) of data assimilation when the **ensemble size is much larger than the state dimension** (***M>>N***).

5. Summary

- **Non-symmetric** solutions of the **ETKF** **do not present clustering**. Nonetheless, they **lose track** **of individual trajectories**.

- For **R-localization**, it is indispensable to have a **symmetric solution** (for **smoothness**). One can apply **rotations** as an **extra step**.

- When **localization and inflation** are used, their **effects** tend to **dominate** and **clustering** is more **difficult to find**.

- In a **simplified AGCM** we find evidence of **asymmetric ensembles** for **some variables**, but this **does not affect the RMSE**. **No episodes of clustering were observed**.