Data assimilation using ensemble Kalman filters

Sebastian Reich in collaboration with Georg Gottwald (University of Sydney)
and Kay Bergemann (Uni Potsdam)

1. Problem statement

Consider a differential equation

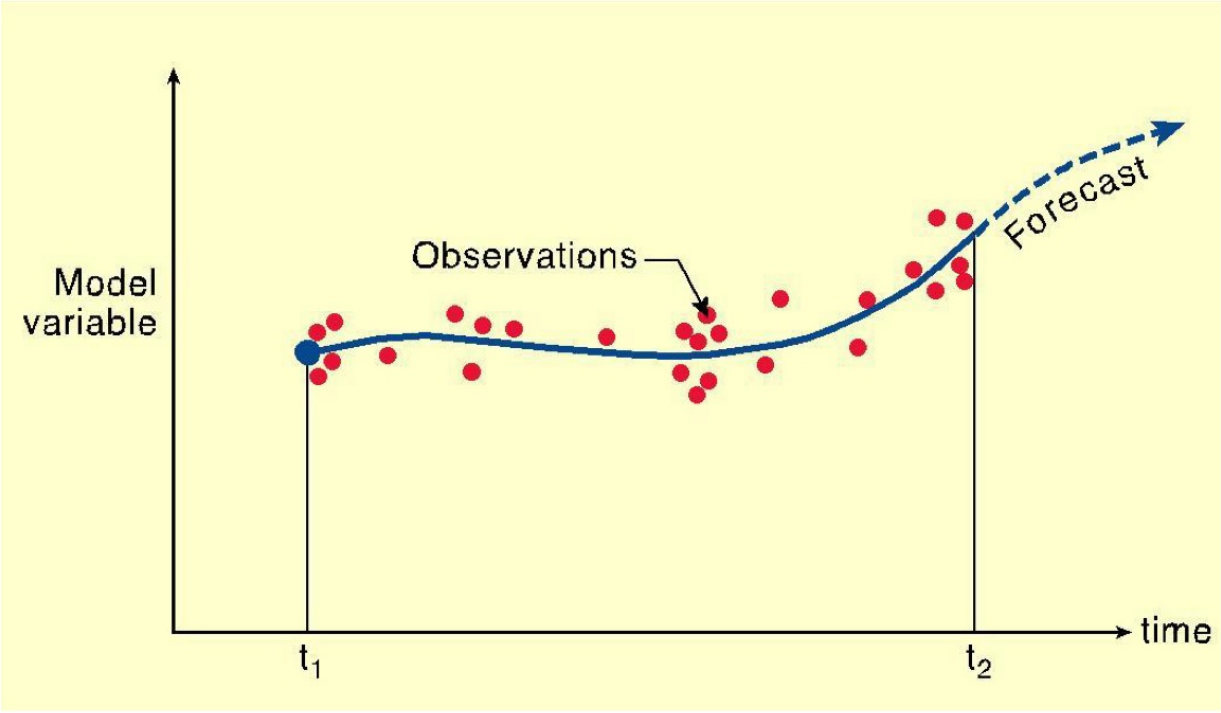
$$\dot{\mathbf{x}} = f(\mathbf{x}), \qquad \mathbf{x} \in \mathbb{R}^n,$$

for which we have some (but not precise) prior knowledge about its initial state $\mathbf{x}(0)$ and for which we can collect observations

$$\mathbf{y}(t_i) \approx \mathbf{H}\mathbf{x}(t_i), \qquad \mathbf{y} \in \mathbb{R}^k, \quad \mathbf{H} \in \mathbb{R}^{k \times n},$$

$k < n$, at discrete times $t_i > 0$, $i = 1, \ldots, I$, subject to some measurement errors.

We wish to find a solution $\mathbf{x}(t)$, $t \in [0, t_I]$, that makes optimal use of the available information in terms of initial data and observations.


Model variable
Observations
Forecast
time
t_1
t_2

A Gaussian distributed stochastic variable $\mathbf{z} \in \mathbb{R}^p$ with mean $\mathbf{a} \in \mathbb{R}^p$ and covariance matrix $\mathbf{B} \in \mathbb{R}^{p \times p}$ will be denoted by

$$\mathbf{z} \sim \mathsf{N}(\mathbf{a}, \mathbf{B}).$$

A typical setting of our data assimilation problem is then

$$\mathbf{x}(0) \sim \mathsf{N}(\mathbf{x}_B, \mathbf{B})$$

and

$$\mathbf{y}(t_i) - \mathbf{H}\mathbf{x}(t_i) \sim \mathsf{N}(\mathbf{0}, \mathbf{R}), \quad i = 1, \ldots, I.$$

2. Existing approaches

We write solution values at t_i with initial condition $\mathbf{x}_0$ at $t = 0$ as $\mathbf{x}(t_i, \mathbf{x}_0)$.

4DVar is a technique to find the optimal initial condition given all the available information. It leads to the minimization of the (nonlinear) cost functional

$$\begin{aligned} J(\mathbf{x}_0) &= \frac{1}{2}(\mathbf{x}_0 - \mathbf{x}_b)^T \mathbf{B}^{-1}(\mathbf{x}_0 - \mathbf{x}_B) \\ &\quad + \frac{1}{2}\sum_{i=1}^{I}(\mathbf{y}(t_i) - \mathbf{H}\mathbf{x}(t_i, \mathbf{x}_0))^T \mathbf{R}^{-1}(\mathbf{y}(t_i) - \mathbf{H}\mathbf{x}(t_i, \mathbf{x}_0)). \end{aligned}$$

Kalman filtering provides an approach in which one marches in time and alternates between a propagation and an analysis step:

Propagation step: Between observations propagate the mean and the covariance under the differential equation, i.e. prior to each observation, we have a most likely state $\mathbf{x}_f(t_i)$ and a covariance matrix $\mathbf{P}_f(t_i)$ and we make the simplifying (often highly questionable) assumption that

$$\mathbf{x}(t_i, \mathbf{x}_0) \sim \mathsf{N}(\mathbf{x}_f(t_i), \mathbf{P}_f(t_i)).$$

Kalman analysis step: Feed in the observations $\mathbf{y}(t_i)$ to obtain an improved most likely state $\mathbf{x}_a(t_i)$ and a covariance matrix $\mathbf{P}_a(t_i)$. We continue with a propagation step using the analyzed states

$$\mathbf{x}(t_i; \mathbf{x}_0) \sim \mathsf{N}(\mathbf{x}_a(t_i), \mathbf{P}_a(t_i)).$$

3. Kalman analysis step

Given the prior mean $\mathbf{x}_f$ and its covariance matrix $\mathbf{P}_f$, the inclusion of observations at t_i leads to the Kalman analysis step

$$\mathbf{x}_a = \mathbf{x}_f - \mathbf{K}\left(\mathbf{H}\mathbf{x}_f - \mathbf{y}\right)$$

with Kalman gain matrix

$$\mathbf{K} = \mathbf{P}_f\mathbf{H}^T\left(\mathbf{H}\mathbf{P}_f\mathbf{H}^T + \mathbf{R}\right)^{-1}$$

and

$$\mathbf{P}_a = (\mathbf{I} - \mathbf{K}\mathbf{H})\,\mathbf{P}_f.$$

I will come back to this step in my second talk. For now we take the formulas for granted.

4 Ensemble propagation and ensemble Kalman filter

We now consider a collection

$$\mathbf{X}(t) = [\mathbf{x}_1(t) \,|\, \mathbf{x}_2(t) \,|\, \cdots \,|\, \mathbf{x}_m(t)] \in \mathbb{R}^{n\times m}$$

of m independent solutions of the differential equation

$$\dot{\mathbf{x}} = f(\mathbf{x})$$

From this matrix we extract the time-dependent ensemble

$$\begin{aligned}
\text{mean:} \quad & \bar{\mathbf{x}} = m^{-1} \sum_i \mathbf{x}_i \in \mathbb{R}^n, \\
\text{deviations:} \quad & \mathbf{Y} = [\mathbf{x}_1 - \bar{\mathbf{x}} \,|\, \mathbf{x}_2 - \bar{\mathbf{x}} \,|\, \cdots \,|\, \mathbf{x}_m - \bar{\mathbf{x}}] \in \mathbb{R}^{n\times m}, \\
\text{covariance:} \quad & \mathbf{P} = \frac{1}{m-1} \mathbf{Y}\mathbf{Y}^T \in \mathbb{R}^{n\times n}.
\end{aligned}$$

The ensemble Kalman filter of Evensen (1996), combines ensemble propagation with the classical Kalman analysis step, i.e., at t_i there is a discontinuous change in $\mathbf{X}(t)$, i.e.

$$\mathbf{X}_f := \mathbf{X}(t_i - \varepsilon) \quad \rightarrow \quad \mathbf{X}(t_i + \varepsilon) = \mathbf{X}_a,$$

where $\mathbf{X}_a$ is the ensemble generated from the assimilated ensemble mean and deviation matrix:

$$\mathbf{X}_a := \mathbf{x}_a \mathbf{e}^T + \mathbf{Y}_a, \qquad \mathbf{e} = [1, 1, \ldots, 1]^T \in \mathbb{R}^m.$$

How to formulate the Kalman analysis step in terms of the ensemble deviations $\mathbf{Y}$ from the mean?

Kalman square root filters rely on a presentation/approximation of a covariance matrix $\mathbf{P} \in \mathbb{R}^{n\times n}$ as the product of a $n \times m$ matrix $\mathbf{Y}$ and its transpose, i.e.

$$\mathbf{P} = \frac{1}{m-1}\mathbf{Y}\mathbf{Y}^T.$$

The Kalman analysis becomes

$$\mathbf{Y}_a = \mathbf{Y}_f\mathbf{T}_r$$

or equivalently

$$\mathbf{Y}_a = \mathbf{T}_l\mathbf{Y}_f.$$

The matrices $\mathbf{T}_r \in \mathbb{R}^{m\times m}$ and $\mathbf{T}_l \in \mathbb{R}^{n\times n}$ are given in terms of square roots of symmetric matrices involving $\mathbf{P}$, $\mathbf{R}$, and $\mathbf{H}$.

We are interested in particular in the case $m \ll n$ and then $\mathbf{T}_r$ is to be preferred. Note that $\mathbf{T}_r$ is not uniquely defined.

We now got a complete filter implementation for nonlinear ODEs.

We have ignored model and numerical time-stepping errors.

There are problems for small ensemble sizes m and very large values of the phase space dimension n due to a poor representation of the ensemble covariance matrix $\mathbf{P} \sim \mathbf{Y}\mathbf{Y}^T$.

These problems are addressed (partially) by ensemble inflation and localization.

5. Continuous factorization algorithms

In Bergemann et al, 2009, we have introduced two new ideas for the implementation of ensemble Kalman filters:

The Kalman analysis step for the ensemble deviations $\mathbf{Y}$ is formulated in terms of an ODE

$$\frac{\mathrm{d}\mathbf{Y}}{\mathrm{d}s} = -\frac{1}{2m-2}\mathbf{Y}\mathbf{Y}^T\mathbf{H}^T\mathbf{R}^{-1}\mathbf{H}\mathbf{Y}$$

with initial condition $\mathbf{Y}(0) = \mathbf{Y}_f$ and $\mathbf{Y}_a = \mathbf{Y}(1)$. The equation is typically solved using two time-steps with forward Euler.

(NB. This ODE is, of course, well known; but has not been used in the context of ensemble Kalman filters before.)

The second idea is to write the matrix $\mathbf{Y}(t)$ of ensemble deviations as a (generalized) SVD decomposition

$$\mathbf{Y}(t) = \mathbf{U}(t)\mathbf{\Sigma}(t)\mathbf{V}(t)$$

with orthogonal $\mathbf{U} \in \mathbb{R}^{n\times m}$, diagonal $\mathbf{\Sigma} \in \mathbb{R}^{m\times m}$ and orthogonal $\mathbf{V} \in \mathbb{R}^{m\times m}$.

To compute such a decomposition from scratch at each time-step would be rather time consuming. Instead one can consider continuous/incremental updates in terms of underlying ODEs for the factors $\mathbf{U}(t)$, $\mathbf{\Sigma}(t)$, $\mathbf{V}(t)$.

Similar ideas have been developed previously for the computation of Lyapunov exponents.

6. Localization

The empirical covariance matrix $\mathbf{P}$ contains spurious long-distance correlations due to under-sampling ($m \ll n$). This problem has led to the idea of localization, i.e., $\mathbf{P}_f\mathbf{H}^T$ in the Kalman analysis step is replaced by

$$\mathbf{P}_f\mathbf{H}^T \to \mathbf{C} \circ \left(\mathbf{P}_f\mathbf{H}^T\right),$$

where $\mathbf{C}$ is a "local" (in some distance) covariance matrix and $\circ$ denotes the Schur product of two matrices $\mathbf{C}, \mathbf{A} \in \mathbb{R}^{n\times k}$, i.e. $(\mathbf{C} \circ \mathbf{A})_{i,j} = (\mathbf{C})_{i,j}\,(\mathbf{A})_{i,j}$.

While it is straightforward to apply localization to the update of the ensemble mean, localized updates of the deviation matrix $\mathbf{Y}_f \to \mathbf{Y}_a$ are the subject of ongoing research.

Based on our continuous update formulation, we can implement localization easily.

For example, we take the (linearized) ODE update equation

$$\frac{\mathrm{d}\mathbf{Y}}{\mathrm{d}t} = -\frac{1}{2}\mathbf{P}_f\mathbf{H}^T\mathbf{R}^{-1}\mathbf{H}\mathbf{Y}$$

with $\mathbf{P}_f$ constant and temper/localize it to

$$\frac{\mathrm{d}\mathbf{Y}}{\mathrm{d}t} = -\frac{1}{2}\left[\mathbf{C}\circ\left(\mathbf{P}_f\mathbf{H}^T\right)\right]\mathbf{R}^{-1}\mathbf{H}\mathbf{Y}.$$

This is a linear, constant coefficient ODE in $\mathbf{Y}(s)$ with $\mathbf{Y}(0) = \mathbf{Y}_f$ and the "Kalman" update is provided by $\mathbf{Y}_a = \mathbf{Y}(1)$.

See Bergemann and Reich, 2009, for details.

7. Numerical results for a barotropic fluid model

We use a 1.5 layer reduced-gravity quasi-geostrophic model with double-gyre wind forcing and biharmonic friction:

$$q_t = -\psi_x - \varepsilon J(\psi, q) - A\Delta^3\psi + 2\pi \sin(2\pi y),$$

where $q = \Delta\psi - F\psi$, $J(\psi, q) = \psi_x q_y - \psi_y q_x$. See Sakov & Oke, Tellus A, 2008, for details.

The number of degrees of freedom (phase space) after spatial discretization is $n = 127 \times 127 = 16129$, the number of observables at each t_i is $k = 300$, and the size of the ensemble is $m = 25$. The variance of the observation error is 4, i.e., $\mathbf{R} = 4\mathbf{I}$. We also use

$$\mathbf{C}_{ij,i'j'} = \exp\left(-(i-i')^2/r_0^2 - (j-j')^2/r_0^2\right)$$

for grid points x_{ij} and $x_{i'j'}$; r_0 the localization radius.

Observations are obtained from a reference numerical trajectory with added noise of variance $\mathbf{R}$, i.e., we treat our numerical model as "perfect". Simulations are run over 4000 time-steps with data assimilated every fourth time-step.

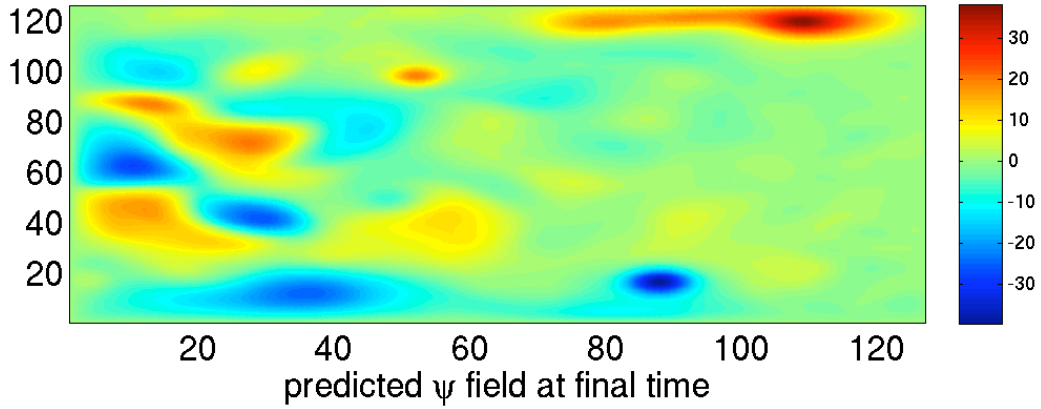
We compute the standard deviation (STD) of the difference between the "true" (unperturbed) trajectory and the ensemble mean at each observation point t_i. Roughly speaking the filter yields "skill" if the STD is less than 2 on average.

We study the behavior of the localized filter for different values of the localization radius r_0 ($r_0 \to \infty$ corresponds to no localization) and ensemble inflation

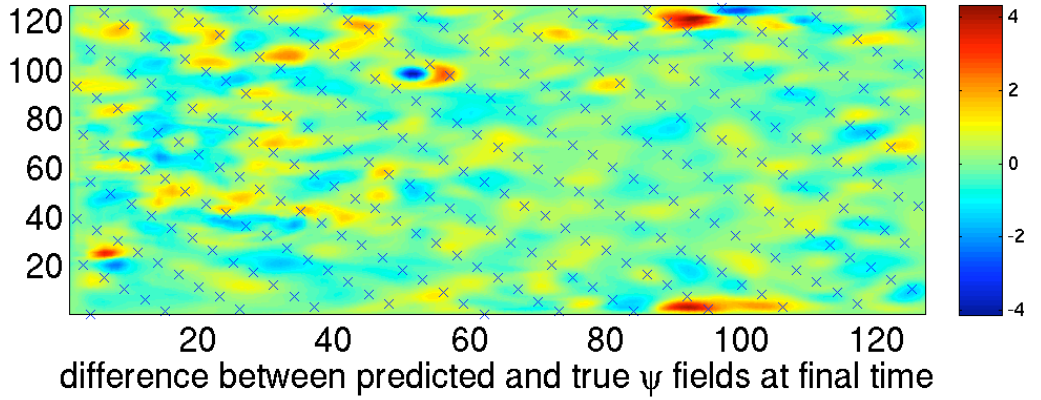
$$\mathbf{Y}(t_i) \to \delta\, \mathbf{Y}(t_i), \qquad \delta > 1$$

after each assimilation step.

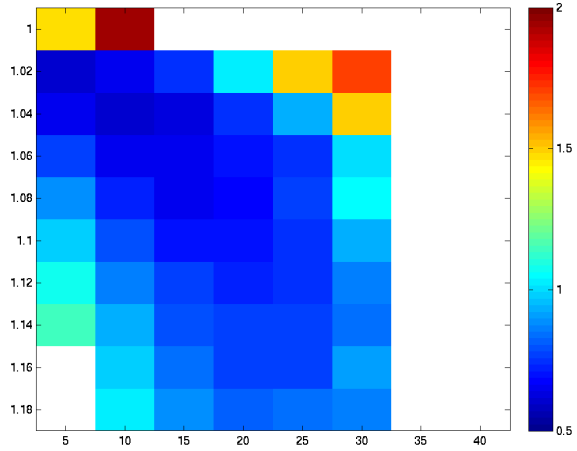
Results at final time:



predicted ψ field at final time



difference between predicted and true ψ fields at final time



$\delta \backslash r_0$	**5**	**10**	**15**	**20**	**25**	**30**	**35**	**40**
1.00	1.46	1.95	Inf	Inf	Inf	Inf	Inf	Inf
1.02	0.60	0.64	0.75	1.03	1.49	1.70	Inf	Inf
1.04	0.65	0.61	0.62	0.74	0.92	1.49	Inf	Inf
1.06	0.76	0.66	0.64	0.69	0.76	1.02	Inf	Inf
1.08	0.88	0.73	0.66	0.68	0.77	1.05	Inf	Inf
1.10	0.97	0.80	0.71	0.69	0.74	0.92	Inf	Inf
1.12	1.06	0.87	0.76	0.72	0.75	0.86	Inf	Inf
1.14	1.14	0.93	0.80	0.76	0.77	0.84	Inf	Inf
1.16	Inf	0.99	0.84	0.78	0.76	0.90	Inf	Inf
1.18	Inf	1.04	0.88	0.81	0.83	0.86	Inf	Inf

Mean RMS error for the ensemble mean update over the last 3000 time steps as a functions of the localization radius r_0 (over grid point indices) and the inflation factor δ. For clarity, the value Inf is assigned if the RMS error exceeds the value 2.0 (no filter skill).

8. Numerical results from the Lorenz 96 model

The standard implementation of the Lorenz-96 model has state vector $\mathbf{x} = (x_1, \ldots, x_n)^T$ with $n = 40$ and its time evolution is given by the equation

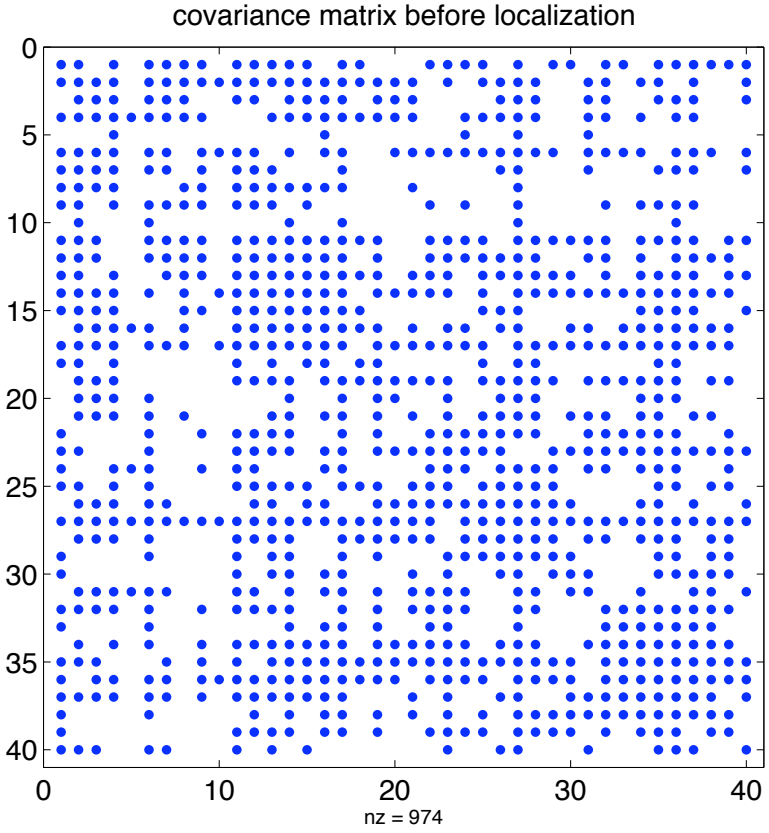
$$\dot{x}_j = (x_{j+1} - x_{j-2})x_{j-1} - x_j + 8$$

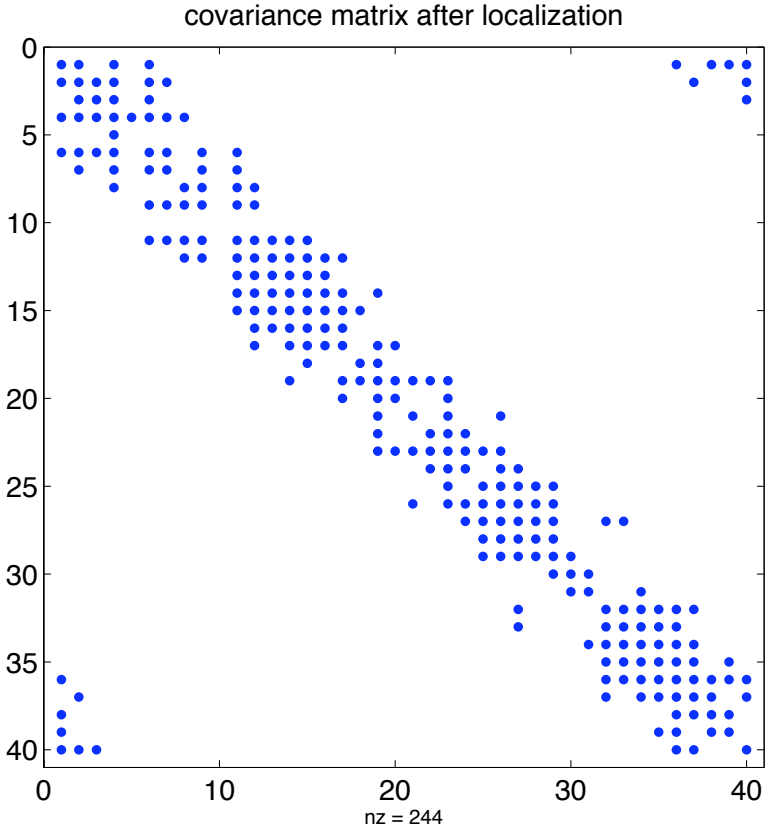
with periodic boundary conditions.

We observer every second grid point, i.e., $k = 20$, with measurement variance $\mathbf{R} = \mathbf{I}$.

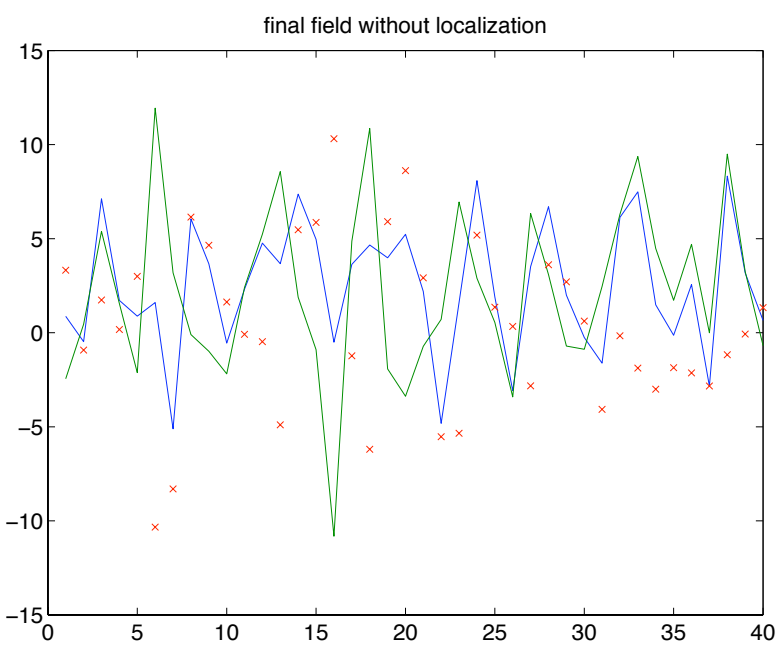
The model is chaotic with 13 positive Lyapunov exponents. One would expect that the necessary ensemble size m should be larger than the number of positive Lyapunov exponents. But this is not the case if localization is used! We show results for $m = 10$.

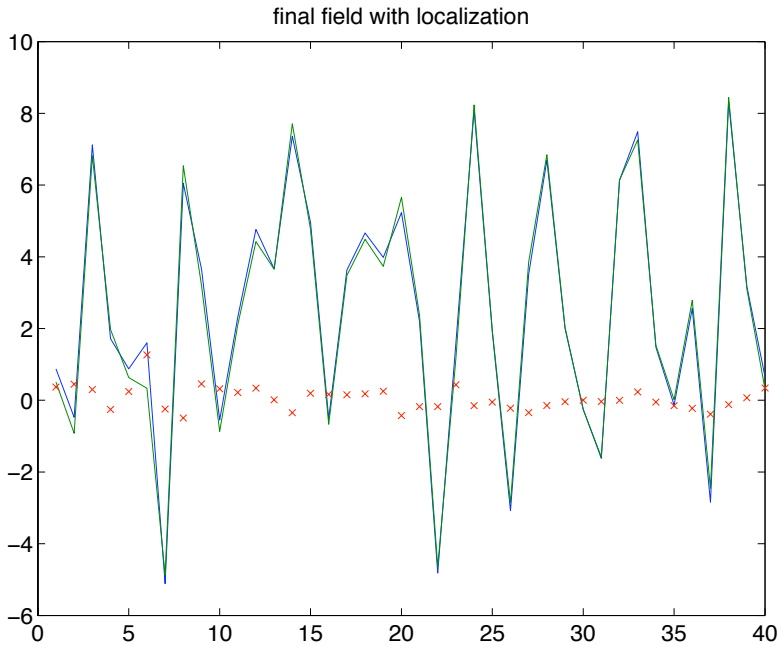
Impact of localization on 40 x 40 covariance matrix. We indicate all entries that are above a certain threshold.





Final fields with and without localization (ensemble size $m = 10$):





(truth green, prediction blue, difference red)

9. Numerical results from a linear balance model

We consider two dynamic variables (PV and geopotential) in an idealized setting of linear advection

$$q_t = q_x, \qquad \Psi_t = \Psi_x.$$

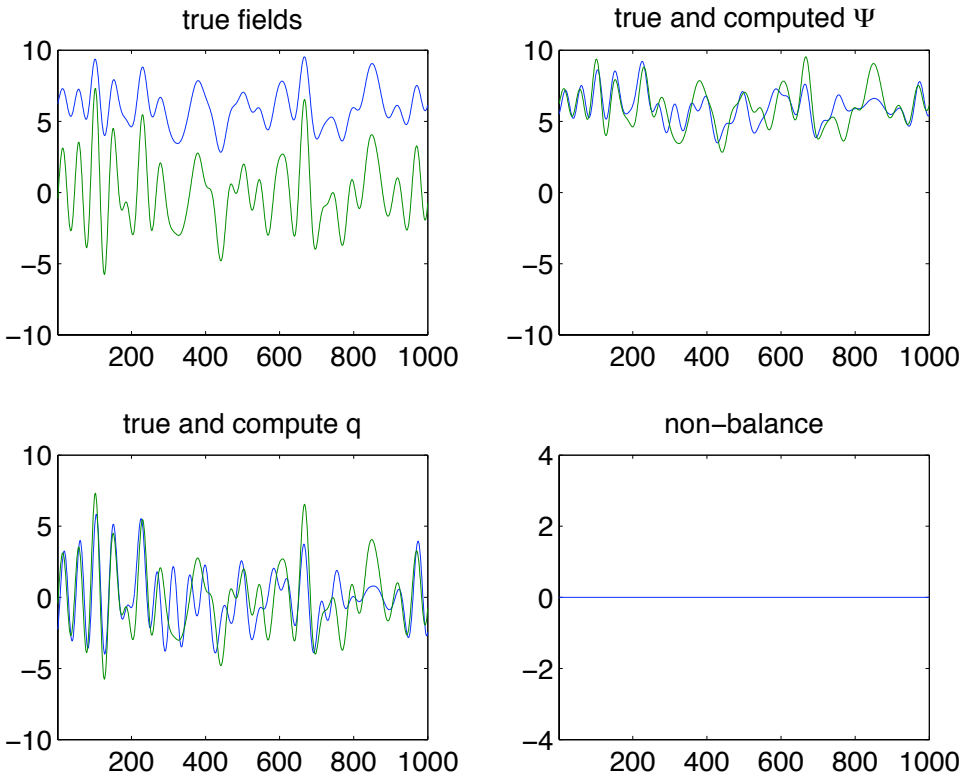
We assume that these two variables are related by the balance relation (PV inversion)

$$\Psi - \alpha^2 \Psi_{xx} = q. \tag{1}$$

We only (partially) observe Ψ and initialize the ensemble members $\mathbf{x}_i = (\Psi_i, q_i)$ such that the balance relation (1) holds.

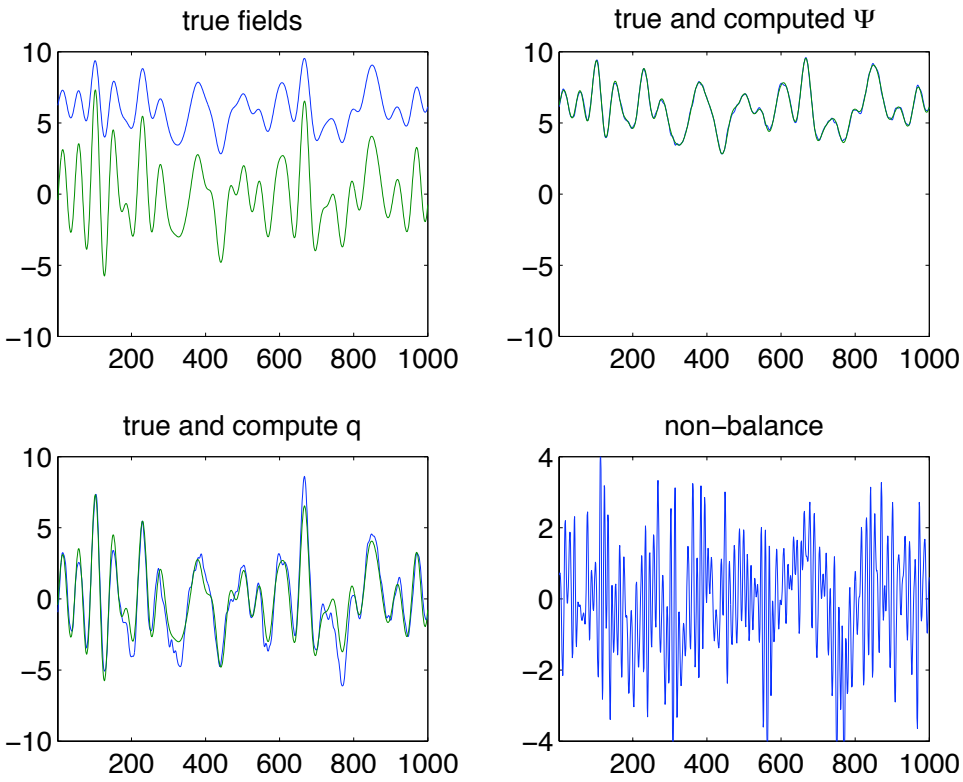
We study the impact of localization on the conservation of (1) for localization radii $r_0 \sim \alpha$.

No localization:



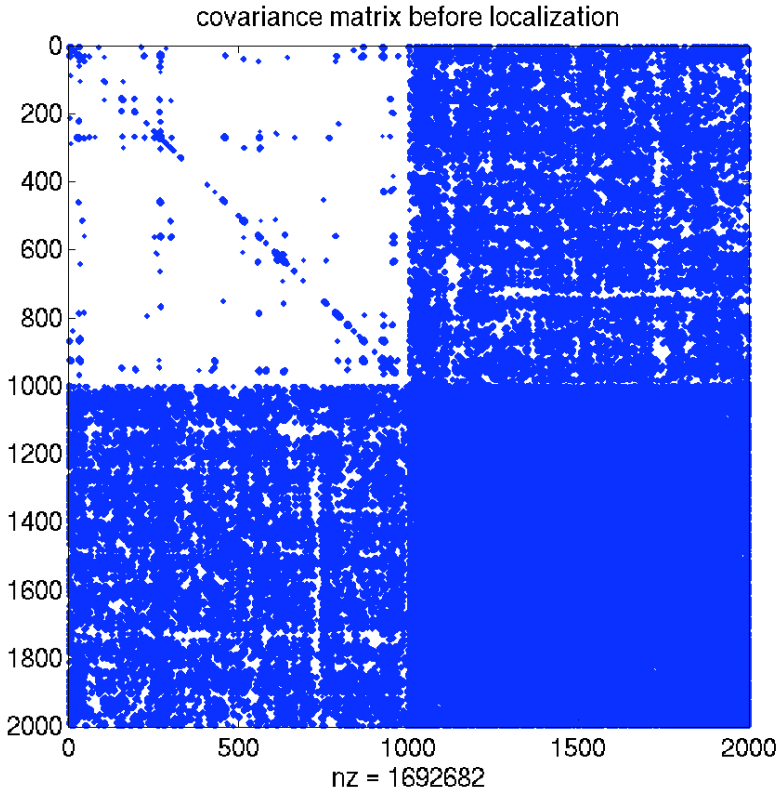
unable to track the reference solution, maintains balance

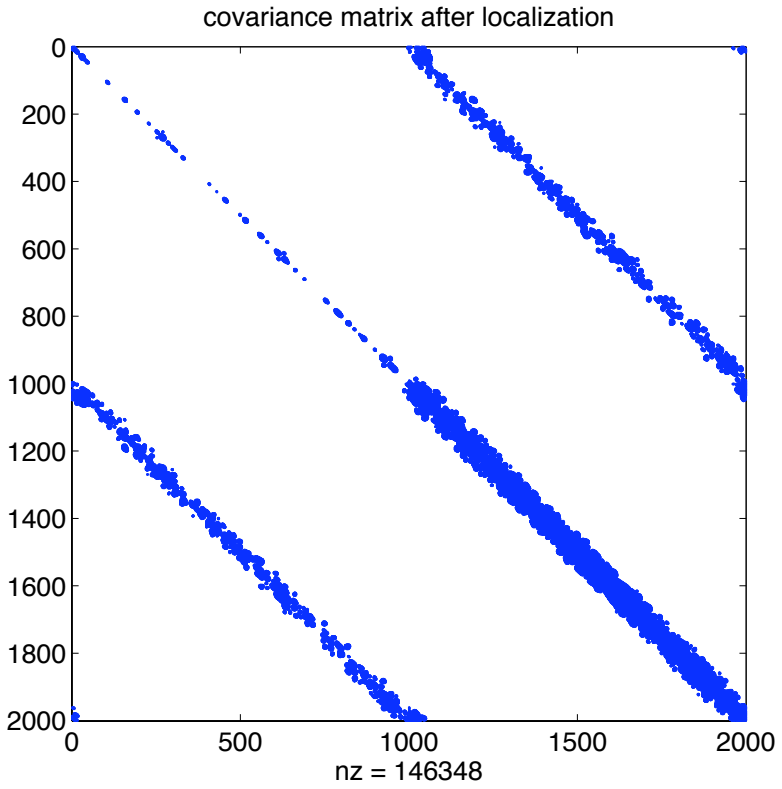
localization with $\alpha = r_0$:



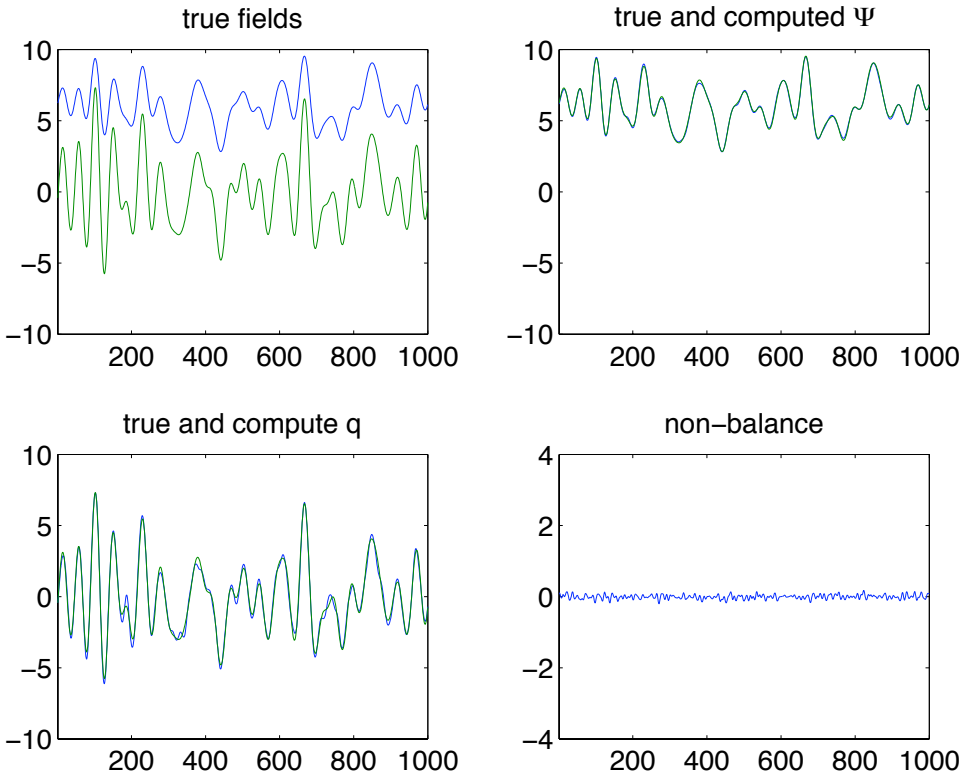
good tracking in Ψ (which is partially observed), balance is disrupted, less optimal tracking of the q field.

Impact of localization on 2000 x 2000 covariance matrix. We indicate all entries that are above a certain threshold.





Here is the magic solution combinging balance with localization (work in progress, keep watching my homepage)



good tracking in Ψ (which is partially observed), balance is maintained, good tracking of the q field (which is not observed!)

10. Conclusion

- Ensemble Kalman filters are a very powerful tool for data assimilation now being used, for example, by the DWD.
- Problems arise for small ensemble sizes due to underestimation of variances and spurious covariances.
- Ensemble inflation and localization are common approaches to overcome these limitations.
- These techniques can be implemented efficiently and robustly within the continuous in time covariance updates.
- Localization is problematic for multi-scale problems and is subject to ongoing research.
- The numerical model is, of course, not "perfect". What is the impact of systematic discretization errors?
- The assumption of Gaussian distributed solutions in the Kalman analysis step is also questionable.

Reference:

[1] Evensen, 2006, Data assimilation, the ensemble Kalman filter. Springer-Verlag,

[2] Bergemann, Gottwald, Reich, 2009, Ensemble propagation and continuous matrix factorization algorithms, Quarterly Journal Royal Meteorological Society, in press.

[3] Bergemann, Reich, 2009, Localization techniques for ensemble transform Kalman filters, submitted.