Data Assimilation Methods

Georgia Sfakianaki

Department of Mathematics and Applied Mathematics,
University of Crete
`gsfakian@tem.uoc.gr`

Introduction

Let $X = \mathbb{R}^n$ be the *model space* and $Y = \mathbb{R}^p$ the *observation space*. Our aim is to produce an analysis estimate $\mathbf{x}^a \in X$ of the true state $\mathbf{x}^t \in X$ of a system, given a background estimate $\mathbf{x}^b \in X$ and observations $\mathbf{y}^o \in Y$ about the true state. We assume that $\mathbf{x}^b$ is evolved from time t_{i-1} to time t_i through the forecast model

$$\mathbf{x}^b(t_i) = M_{i-1}[\mathbf{x}^a(t_{i-1})],$$

where $M_{i-1} : X \to X$ is the (nonlinear) *model operator*, whereas the observations $\mathbf{y}^o$ are generated by the *observation model*

$$\mathbf{y}^o = H(\mathbf{x}^t) + \varepsilon_o,$$

where $H : X \to Y$ is the (nonlinear) *observation operator* and ε_o the observation error.
We consider that the background, the analysis and the observations contain zero-mean normally distributed errors:

$$\varepsilon_b = \mathbf{x}^b - \mathbf{x}^t, \quad \varepsilon_a = \mathbf{x}^a - \mathbf{x}^t, \quad \varepsilon_o = \mathbf{y}^o - H(\mathbf{x}^t).$$

Their respective covariance matrices are defined as

$$\mathbf{B} = E\{\varepsilon_b \varepsilon_b^T\}, \quad \mathbf{P}^a = E\{\varepsilon_a \varepsilon_a^T\}, \quad \mathbf{R} = E\{\varepsilon_o \varepsilon_o^T\}.$$

Moreover, we assume that the errors in the background and the observations are uncorrelated, i.e., $E\{\varepsilon_o \varepsilon_b^T\} = 0$.

3D-Var

In 3D-Var we seek the optimal analysis $\mathbf{x}^a$ that minimizes the cost function of the state $\mathbf{x}$:

$$J(\mathbf{x}) = \frac{1}{2}(\mathbf{x} - \mathbf{x}^b)^T \mathbf{B}^{-1}(\mathbf{x} - \mathbf{x}^b) + \frac{1}{2}[\mathbf{y}^o - H(\mathbf{x})]^T \mathbf{R}^{-1} [\mathbf{y}^o - H(\mathbf{x})],$$

which measures the distance between $\mathbf{x}$ and the background $\mathbf{x}^b$, weighted by the inverse of the background error covariance $\mathbf{B}$, plus the distance between $\mathbf{x}$ and the observations $\mathbf{y}^o$, weighted by the inverse of the observation error covariance $\mathbf{R}$.
The minimum of the functional is attained when its gradient is equal to zero. The gradient of $J(\mathbf{x})$ with respect to $(\mathbf{x} - \mathbf{x}^b)$ is

$$\nabla J(\mathbf{x}) = \mathbf{B}^{-1}(\mathbf{x} - \mathbf{x}^b) + \mathbf{H}^T\mathbf{R}^{-1}\mathbf{H}(\mathbf{x} - \mathbf{x}^b) - \mathbf{H}^T\mathbf{R}^{-1}[\mathbf{y}^o - H(\mathbf{x}^b)],$$

where $\mathbf{H}$ is the matrix of the first-order partial derivatives of the nonlinear operator H, i.e., its elements are $h_{ij} = \partial H_i / \partial x_j$. Taking $\nabla J(\mathbf{x}^a) = 0$ and solving for $\mathbf{x}^a$, yields the 3D-Var analysis

$$\mathbf{x}^a = \mathbf{x}^b + \mathbf{W}[\mathbf{y}^o - H(\mathbf{x}^b)],$$
$$\mathbf{W} = [\mathbf{B}^{-1} + \mathbf{H}^T\mathbf{R}^{-1}\mathbf{H}]^{-1}\mathbf{H}^T\mathbf{R}^{-1},$$

where $\mathbf{W}$ is the optimal weight matrix, in the sense that it minimizes the analysis error variance.
Formally, this is the solution of the 3D-Var minimization problem. In practice, the minimization of the cost function is carried out using iterative minimization algorithms, such as the Conjugate Gradient or quasi-Newton methods.

In Figures 1 and 2, there are presented results of the 3D-Var assimilation implementation for the Lorenz-96 model.

The Lorenz-96 Model

The Lorenz-96 is a system of N ordinary differential equations:

$$\frac{dX_i}{dt} = (X_{i+1} - X_{i-2})X_{i-1} - X_i + F, \quad i = 1, \dots, N,$$

with cyclic boundary conditions $X_{-1} = X_{N-1}$, $X_0 = X_N$, $X_{N+1} = X_1$ and constant forcing term F.
We have considered $N = 40$ and $F = 8$ for which the system exhibits chaotic behavior. Furthermore, as initial condition we have assumed the steady state solution $X_i = F$, $\forall\ i = 1, \dots, N$, with a small perturbation introduced in $X_{20} = F + 0.008$. A fourth-order Runge-Kutta scheme with time-step $\Delta t = 0.05$ has been used for the numerical integration.

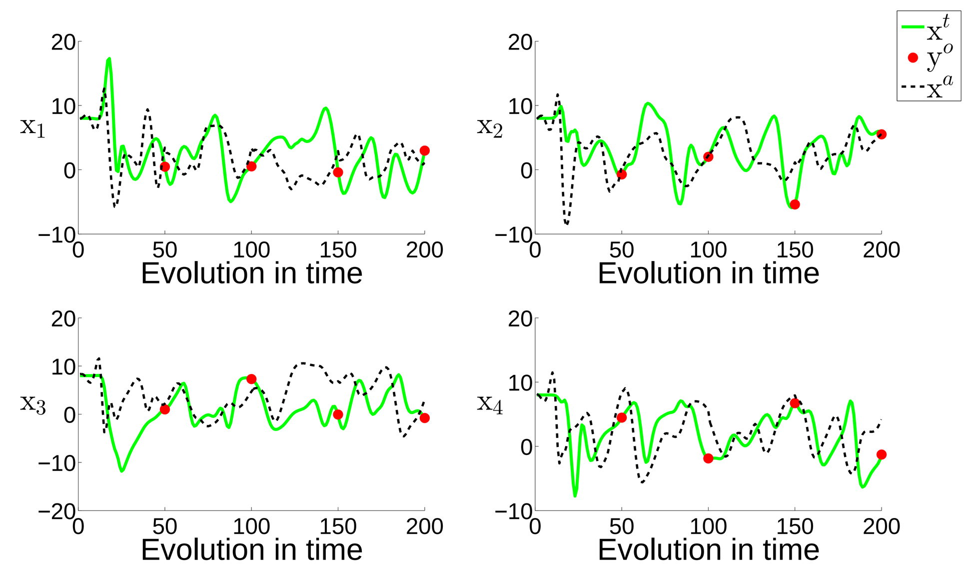

Figure 1: *3D-Var assimilation for the Lorenz-96 model. The background error covariance is a diagonal matrix, localized around the last 20 sites. Observations are available at each site, with an error $\varepsilon_o \sim N(0, \sigma_o^2)$, $\sigma_o = 0.20$ and assimilation is performed every 50 integration steps. We display the first four components of the analysis (black dashed line) and the true state (green solid line), as well as the available observations (red dots).*

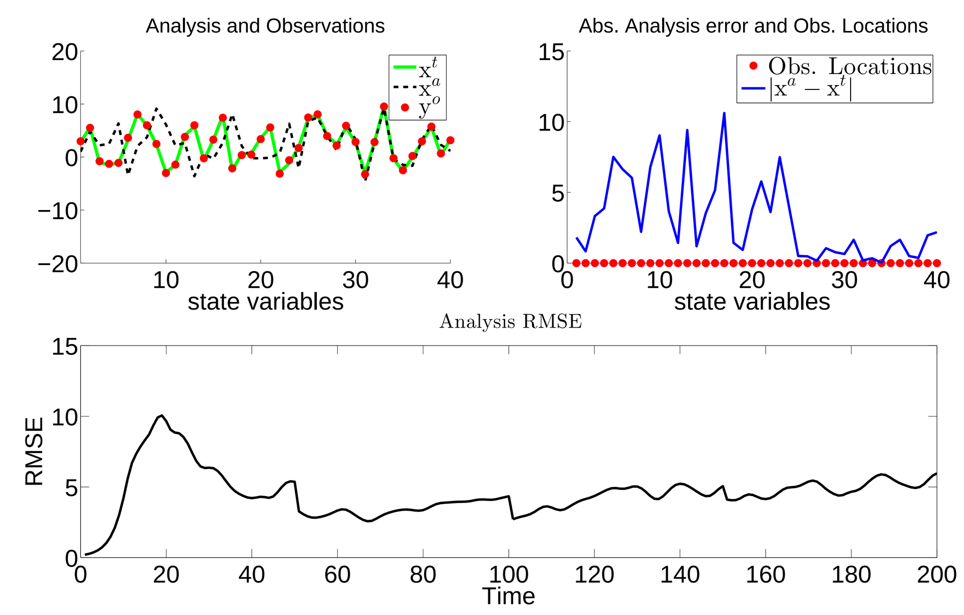

Figure 2: *Upper left plot shows the analysis estimate, the true state and the observations, during the last assimilation step. Upper right plot shows the absolute analysis error against the observation locations. Bottom plot shows the evolution in time of the analysis RMSE. The average RMSE is 4.5171.*

Ensemble Square Root Filter

EnSRF is a deterministic Ensemble Kalman Filter (EnKF). It is a recursive algorithm for estimating the true state of a dynamical system from a set of noisy observations and consists of the *forecast* and the *analysis* step. EnSRF provides the optimal analysis estimate of the true state and also its uncertainty.

We begin with an ensemble of K members at time t_{i-1} $\{\mathbf{x}_k^a \in X,\ k = 1, \dots, K\}$. Using the model we obtain K forecasts at time t_i:

$$\mathbf{x}_k^f(t_i) = M_{i-1}[\mathbf{x}_k^a(t_{i-1})], \qquad k = 1, \dots, K.$$

We define the forecast ensemble perturbations matrix

$$\mathbf{X}^f = \frac{1}{\sqrt{K-1}} \sum_{k=1}^{K} \left(\mathbf{x}_k^f - \bar{\mathbf{x}}^f\right), \quad \text{where} \quad \bar{\mathbf{x}}^f = \frac{1}{K}\sum_{k=1}^{K} \mathbf{x}_k^f$$

is the forecast ensemble mean. Then, we form the forecast error covariance matrix $\mathbf{P}^f = \mathbf{X}^f(\mathbf{X}^f)^T$.
A set of observations $\mathbf{y}^o$ becomes available and we proceed to the analysis step. We first compute the *Kalman Gain* matrix

$$\mathbf{K} = \mathbf{X}^f(\mathbf{X}^f)^T\,\mathbf{H}^T\left[\mathbf{R} + \mathbf{H}\,\mathbf{X}^f(\mathbf{X}^f)^T\mathbf{H}^T\right]^{-1}$$

and then, the analysis ensemble mean is given as

$$\bar{\mathbf{x}}^a = \bar{\mathbf{x}}^f + \mathbf{K}[\mathbf{y}^o - H(\bar{\mathbf{x}}^f)].$$

The analysis error covariance matrix, which is defined as $\mathbf{P}^a = [\mathbf{I} - \mathbf{K}\mathbf{H}]\mathbf{P}^f$, can be written in the form

$$\mathbf{P}^a = \mathbf{X}^f\mathbf{T}(\mathbf{X}^f)^T,$$

where $\mathbf{T}$ is a Hermitian, positive-definite matrix. Then, the analysis ensemble perturbations matrix is calculated as $\mathbf{X}^a = \mathbf{X}^f\mathbf{S}$, where $\mathbf{S}$ is the square root matrix of $\mathbf{T}$. Finally, adding to $\mathbf{X}^a$ the analysis ensemble mean $\bar{\mathbf{x}}^a$, we obtain the desired analysis ensemble.
After completing the analysis step at time t_i, we use the calculated analysis to update the forecast ensemble at the next time-step t_{i+1}.

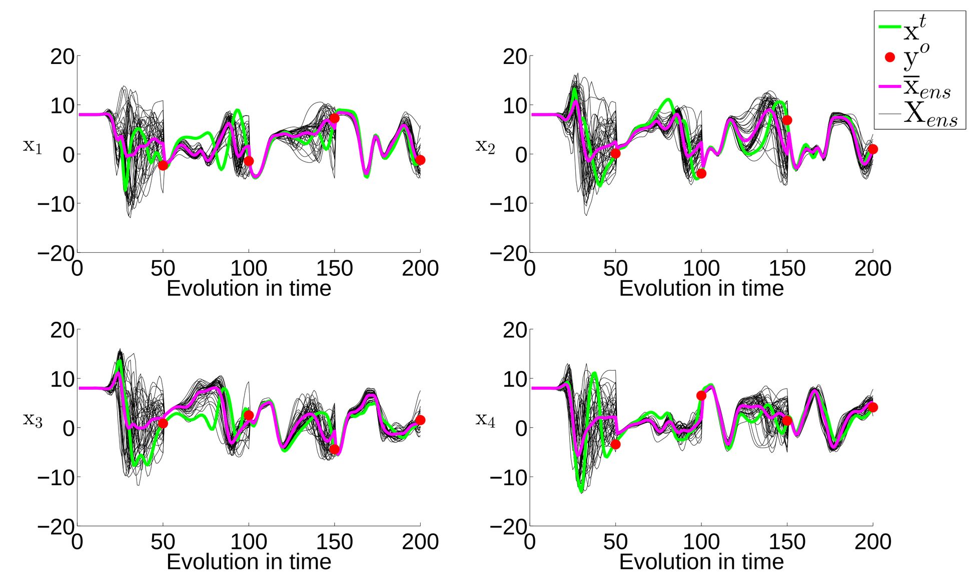

Figure 3: *EnSRF results for the Lorenz-96 model. We use an ensemble of 40 members. Observations are available at each site, with an error $\varepsilon_o \sim N(0, \sigma_o^2)$, $\sigma_o = 0.20$. Assimilation is performed every 50 integration steps. As soon as an observation (red dot) becomes available, the ensemble (black lines) is adjusted to fit the data. The ensemble mean (magenta line) is the analysis estimate of the true state (green line).*

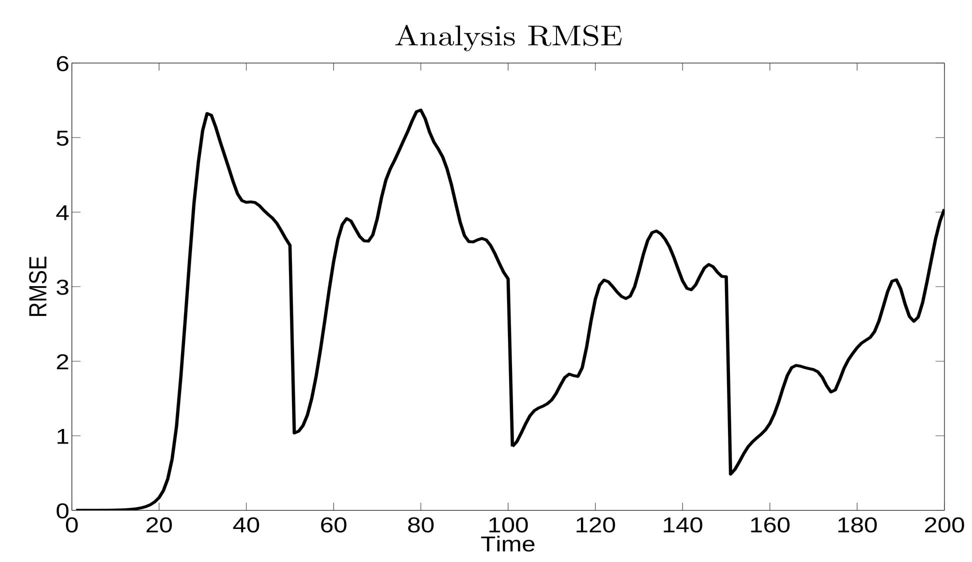

Figure 4: *EnSRF analysis RMSE. After the assimilation of each observation, there is a significant reduction of the analysis RMSE. The average RMSE is* 2.6344.

This work is based on my Master Thesis entitled "Data Assimilation Methods" presented in June 2016 at the University of Crete.