BE: Data Assimilation

Selime Gürol
CERFACS
selime.gurol@cerfacs.fr

Emilien Flayac
ISAE - SUPAERO
emilien.flayac@isae-supaero.fr

For the following practical we encourage the students to work in pairs. A report is to submit on the LMS platform. **Deadline : Friday February** 24^{th}**, 2023**.

1 Statistical Analysis

The aim of this hands-on is to understand the analysis sensitivity with respect to different settings of the statistical parameters in a simple 1D data assimilation system. Assume that we have noisy observations :

$$\mathbf{y} = \mathbf{G}\mathbf{x}_{true} + \epsilon_o$$

with $\mathbf{G} \in \mathbb{R}^{m\times n}$ being an observation operator and ϵ_o follows a Gaussian distribution with zero mean and observation error covariance matrix ($\mathbf{R}$), i.e. $\epsilon_o \sim \mathcal{N}(\mathbf{0}, \mathbf{R})$. Assume also that we have noisy a priori (background) knowledge :

$$\mathbf{x}^b = \mathbf{x}_{true} + \epsilon_b$$

with $\epsilon_b \sim \mathcal{N}(\mathbf{0}, \mathbf{B})$, with $\mathbf{B}$ being the background error covariance matrix. We assume that the observation and background errors are unbiased. We also assume that the background and observation errors are uncorrelated with each other. Then, the Best Linear Unbiased Estimater (BLUE) is given by

$$\mathbf{x}^a = \mathbf{x}^b + \mathbf{K}(\mathbf{y} - \mathbf{G}\mathbf{x}^b) \tag{1}$$

where

$$\mathbf{K} = \mathbf{B}\mathbf{G}^T(\mathbf{G}\mathbf{B}\mathbf{G}^T + \mathbf{R})^{-1}$$

1.1 Theoretical analysis on the influence of observations at different points

Question 1 *Assume that we have temperature estimates (background information) at Toulouse and Montpellier (x^b_T, x^b_M) and one observation available only at Montpellier (y_M). Assume that the observation error has variance σ_o^2. The background error has the same variance σ_b^2 at both locations. The correlation of background error between two cities is given by ρ. The observation operator* $\mathbf{G}$ *is defined as a selection operator, i.e.*

$$\mathbf{G} = \begin{bmatrix} 0 & 1 \end{bmatrix}$$

which leads to

$$\mathbf{G}\begin{bmatrix}x^b_T\\x^b_M\end{bmatrix}=\begin{bmatrix}0 & 1\end{bmatrix}\begin{bmatrix}x^b_T\\x^b_M\end{bmatrix}=x^b_M$$

Write the equations for the temperature estimates both at Toulouse (x^a_T) and at Montpellier (x^a_M) based on the BLUE analysis (Eq. 1).
Hint : *First write out the matrix expressions for* $\mathbf{G}$, $\mathbf{B}$, $\mathbf{BG}^{\mathrm{T}}$, $\mathbf{GBG}^{\mathrm{T}}$ *and* $\mathbf{R}$ *from which the analysis can be derived from the BLUE equation.*

Question 2 *Let us now consider that we have two observations (one at grid point* 1 *(y_1) and one at grid point* 2 *(y_2)) coming from the same instrument type, and having the same error variance σ_o^2. The observation errors are assumed to be uncorrelated. Assume also that we have a background estimate at both observation locations, and that their error variances are assumed to be the same and given by σ_b^2. The correlation of background error between grid point i and j is given by ρ_{ij}. The analysis, x_0^a, is located at grid point* 0.

(Q2.1) *Write out the matrix expressions for* $\mathbf{G}$, $\mathbf{B}$, $\mathbf{BG}^{\mathrm{T}}$, $\mathbf{GBH}^{\mathrm{T}}$ *and* $\mathbf{R}$ *from which the analysis can be derived from the BLUE equation.*

Using these matrices the BLUE at grid point 0 can be written as

$$x_0^a = x_0^b + \begin{bmatrix}w_1 & w_2\end{bmatrix}\begin{bmatrix}y_1 - x_1^b\\ y_2 - x_2^b\end{bmatrix} \tag{2}$$

where

$$w_1 = \frac{\rho_{10}(1+\alpha)-\rho_{12}\rho_{20}}{(1+\alpha)^2-\rho_{12}^2}$$
$$w_2 = \frac{\rho_{20}(1+\alpha)-\rho_{12}\rho_{10}}{(1+\alpha)^2-\rho_{12}^2}$$

and $\alpha = \sigma_o^2/\sigma_b^2$.

With these optimal weights, the expected analysis error variance is

$$\langle\,(\epsilon_0^a)^2\,\rangle = \sigma_b^2\left[1-\frac{(1+\alpha)(\rho_{10}^2+\rho_{20}^2)-2\rho_{10}\rho_{20}\rho_{12}}{(1+\alpha)^2-\rho_{12}^2}\right] \tag{3}$$

By using Eqs. (2) and (3), we now look at the influence of two observations on the analysis for different scenarios as shown in Figure 1.

<u>Two isolated observations :</u>
Consider the case where the observations are located on either side of the analysis grid point. Let us assume that $\rho_{12} = 0$, i.e. the two observations are so far from each other that the background error correlation between the two locations is zero. Assume that $\rho_{10} = \rho_{20} = \rho$, i.e., the background error correlation between grid points 1 and 0, and grid points 2 and 0 is the same.
(Q2.2) *What is the analysis and its error variance at grid point* 0 *? (Use Eqs. (2) and (3))*

<u>Two collocated observations :</u>
What happens if the two observations are collocated rather than being located on either side

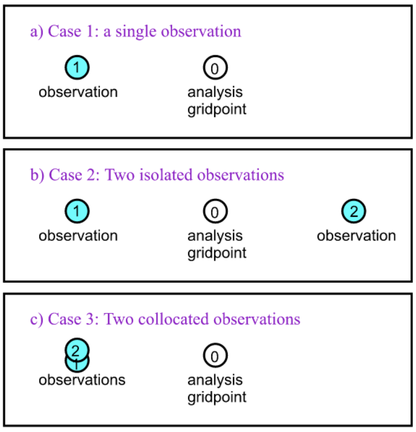

FIGURE 1 – The influence of two observations on the analysis at the grid point 0

of the analysis grid point. In this case, take $\rho_{12} = 1$ and $\rho_{10} = \rho_{20} = \rho$.
(Q2.3) *What is the analysis and its variance at gridpoint 0 ? (Use Eqs. (2) and (3))*
(Q2.4) *Do independent or collocated observations provide more information ?*

1.2 Implementation

The python code `statistical_analysis.py` allows solving a simple statistical analysis problem to understand fundamental concepts. In this code, the observation error covariance matrix is defined as a diagonal matrix :

$$\mathbf{R} = \sigma_o^2 \mathbf{I}$$

where σ_o is the observation error standard deviation. The background error covariance matrix can be chosen either as a diagonal matrix, i.e. $\mathbf{B} = \sigma_b^2 \mathbf{I}$ or as a dense matrix where the correlations between the grid points are considered. In the latter case, the elements of the background error covariance matrix $\mathbf{B}$ are given by

$$\mathbf{B}_{ij} = \sigma_b^2 \, \rho_{ij}, \tag{4}$$

where σ_b is the background error standard deviation. The ρ_{ij} are given by a second-order autoregressive (SOAR) correlation function of the distance r_{ij} between the grid-points i and j

$$\rho_{ij} = \left(1 + \frac{r_{ij}}{L}\right) \exp\left(-\frac{r_{ij}}{L}\right), \tag{5}$$

where L is a correlation length scale parameter. The observation operator $\mathbf{G}(\cdot)$ is chosen as a selection operator.

Question 3 *Complete the* `statistical_analysis.py` *code. Explore the sensitivity of the analysis with respect to given parameters :*
(Q3.1) *What is the influence of the observations on the analysis by changing their locations*

and σ_o ;
(Q3.2) *Compare the influence of the B matrix on the analysis at the grid points with an observation and without an observation when (i) B is chosen as a diagonal matrix and (ii) B is chosen as a soar function.*
(Q3.3) *Fix the parameter "btype" = 'soar' in the code. What is the influence of the background on the analysis by changing the correlation length scale parameter and σ_b.*

2 Variational Data Assimilation

In order to analyse the different data assimilation algorithms, simulated data ("identical twin") assimilation experiments are conducted with *Python*. In the identical twin experiments, the Lorenz-95 model is used to generate a true model state, $\mathbf{x}_t$, at time t. The background (*a priori*) state and observations are then generated by adding Gaussian noise to fields drawn from the true model state :

$$\begin{aligned} \mathbf{x}_b &= \mathbf{x}_t + \epsilon_b, \quad \epsilon_b \sim \mathcal{N}(\mathbf{0}, \mathbf{B}) \\ \mathbf{y} &= \mathcal{H}(\mathbf{x}_t) + \epsilon_o, \quad \epsilon_o \sim \mathcal{N}(\mathbf{0}, \mathbf{R}) \end{aligned}$$

Since the true model state is known, we can compare solutions obtained from different data assimilation algorithm with the truth.

2.1 Dynamical Model : The Lorenz-95 system

The Lorenz-95 system contains K variables : $X_1, X_2, \ldots, X_K$, and is governed by the K equations :

$$\frac{dX_k}{dt} = \underbrace{-X_{k-2}X_{k-1} + X_{k-1}X_{k+1}}_{\text{advection}} - \underbrace{X_k}_{\text{internal dissipation}} + \underbrace{F}_{\text{forcing}} . \tag{6}$$

In Eq. (6) the quadratic terms represent the advection that conserves the total energy, the linear term represents the damping through which the energy decreases, and the constant term represents external forcing keeping the total energy away from zero. The K variables may be thought of as *values of some atmospheric quantity in K sectors of a latitude circle.* (Figure 2). For small values of F, all solutions decay to the steady solution $X_1 = .. = X_k =$

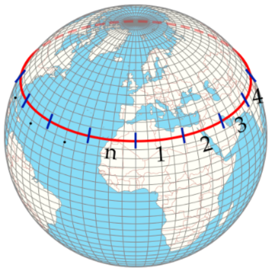

Figure 2 – Example of a latitude circle of the Earth, divided into K equal sized sectors.

F, while when F is somewhat larger, most solutions are periodic. For still larger values of $F \approx 8$, the system is chaotic. The system has similar error growth characteristics to an operational Numerical Weather Prediction (NWP) system if one time unit of the Lorenz-95 system is associated with 5 days of an NWP system.

For this exercise, $F = 8$, $K = 40$ and the boundary conditions are cyclic, i.e., $x_0 = x_{40}$, $x_{-1} = x_{39}$ and $x_{41} = x_1$. The equations are solved using a fourth-order Runge-Kutta scheme, using $\Delta t = 0.025$ (a 3 hour time step).

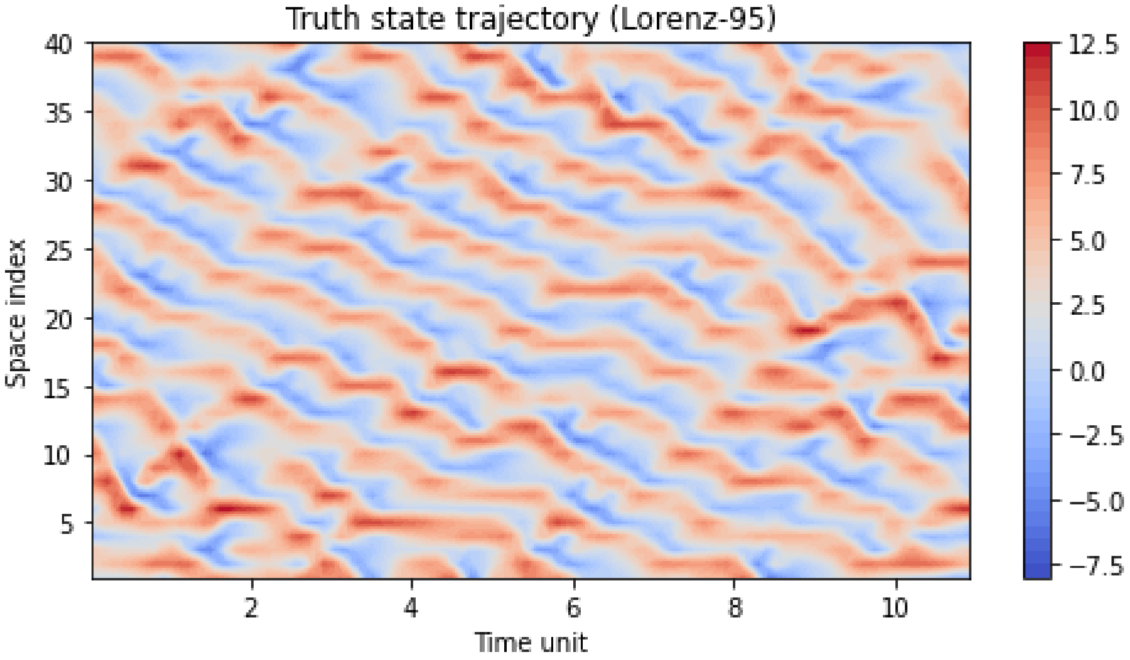

FIGURE 3 – Space-time diagram of a model run in the configuration of the exercise.

2.2 3D-Var

The python code `3dvar.py` allows solving a 3D variational data assimilation (3D-Var) problem. In this code, unlike `statistical_analysis.py`, the matrices are coded as *operators*. This is necessary when dealing with large-scale systems. All the operators are coded in `operators.py`.

In the `3dvar.py` code, the observation error covariance matrix is defined as a diagonal matrix :

$$\mathbf{R} = \sigma_o^2 \mathbf{I}$$

where σ_o is the observation error standard deviation. Two types of background error covariance operators are available. The first models the product of a vector with a diagonal $\mathbf{B}$ and is equivalent to a product with σ_b^2. The second is a diffusion operator, which models the product of a vector with a dense covariance matrix without building explicitly the matrix in question. With diffusion operators, two parameters control the nature of the correlation ρ_{ij}

between the different points : a length scale D, and a smoothness parameter M, chosen as an even integer. The ρ_{ij} correspond to an auto-regressive function of order M and depend on the distance r_{ij} :

$$\rho_{ij} = \sum_{j=0}^{M-1} \beta_j \left(\frac{r_{ij}\sqrt{2M-3}}{D} \right)^j e^{-\frac{r_{ij}\sqrt{2M-3}}{D}}, \tag{7}$$

where

$$\beta_j = \frac{2^j (M-1)!(2M-j-2)!}{j!(M-j-1)!(2M-2)!}. \tag{8}$$

When $M = 2$, this function is equivalent to the SOAR function seen in the previous section.

The diffusion and diagonal operators, and their respective inverse and square root operators are implemented in `operators.py`.

Question 4 *First complete the* ***3dvar.py*** *code. Then let us understand the effect of* $\mathbf{B}$ *:*

(Q4.1) *Random correlated background errors can be generated by applying the square root of* $\mathbf{B}$ *to a standard normally distributed vector.Run the code 3dvar.py with differents settings (diagonal, diffusion with small/large length scale), and look at the difference between the truth* $\mathbf{x}_t$ *and the background* $\mathbf{x}_b$ *(i.e. the background error* ϵ_b*). How is it affected by the different settings of* $\mathbf{B}$ *? In particular, how does a spatially correlated error differ from an uncorrelated one ?*

(Q4.2) *How is the analysis affected when correlations are accounted for in* $\mathbf{B}$ *? Try to reduce the number of observations and increase* σ_b*. What is the effect of the correlations on the analysis ? In particular, what happens between the observation points ?*

(Q4.3) *Two parameters are available for the diffusion operator :* D *and* M*. How does* D *affect the background error covariance and the analysis ?*
Bonus *: How does* M *affect the background error covariance and the analysis ?*

2.3 4D-Var

The python code `4dvar.py` allows solving a 4D variational data assimilation (4D-Var) problem. Identical twin experiments are conducted by using Lorenz-95 model. This code structure is quite similar to that of `3dvar.py` in the sense that all the matrices are available as operators. These operators are coded in `operators.py`.

Question 5 *Understand the effect of observation accuracy and frequency :*

— *Start with many observations that are noisy, for instance :*
→ $n = 100$
→ $nt = 10$
→ $\mathbf{B}$ *as a diffusion operator*
→ *frequency of observations in space :* 30
→ *frequency of observations in time :* 10
→ $\sigma_o = 10^{-2}$

- *Compare these results with (thinning the data)*
  → *frequency of observations in space :* 10
  → *frequency of observations in time :* 5
- *Compare these results with more accurate observations :*
  → *frequency of observations in space :* 10
  → *frequency of observations in time :* 5
  → $\sigma_o = 10^{-4}$
  What is the absolute error for each case ? How many iterations are required for convergence ? What is the relation with the conditioning of the minimization problem and the accuracy/frequency of the observations ?

Question 6 *Understand the effect of number of outer and inner iterations. Change the number of inner and outer iterations (keep the total number of inner iterations fixed). For instance :*

- *Start with the following parameters :*
  → $n = 100$
  → $nt = 10$
  → **B** *as a diffusion operator* $(D = 5, M = 4)$
  → *frequency of observations in space :* 30
  → *frequency of observations in time :* 10
  → $\sigma_o = 10^{-4}$
  → *max_outer* = 5 *and max_inner* = 200
- *Compare these results with*
  → *max_outer* = 10 *and max_inner* = 100

Question 7 *Understand the effect of preconditioning. For instance,*

- *Start with the following parameters*
  → $n = 100$
  → $nt = 10$
  → **B** *as a diffusion operator* $(D = 5, M = 4)$
  → *frequency of observations in space :* 10
  → *frequency of observations in time :* 5
  → $\sigma_o = 10^{-2}$
  → *max_outer* = 10 *and max_inner* = 500
  → $\mathbf{F} = \mathbf{B}$
- *Compare these results with*
  → $\mathbf{F} = \mathbf{I}$