

Ensemble Kalman filter applied to Lorenz's 1963 model

A. Fournier fournier@ipgp.fr

Université Paris Cité, Institut de physique du globe de Paris, France

May 28th, 2026 – NuTS-W4, Strasbourg



Link to online wooclap event: <https://app.wooclap.com/SSOSAM>

Preliminary question

How familiar are you with data assimilation / inverse problem theory concepts?

1st question of the wooclap series

What is data assimilation?

The basic purpose of data assimilation is to combine different sources of information in order to produce the best possible estimate of the state of a system.

- ▶ observations of the system
- ▶ physical laws describing its behaviour (numerical model)

Why not simply use observations?

- ▶ too sparse or partial in geo- and astrophysics
- ▶ interpolation: numerical model
- ▶ combining (by means of the model) several noised data: filter out part of the noise & more accurate estimate ("accuracies are added")

Different mathematical approaches

The problem of data assimilation can be tackled using different mathematical approaches: signal processing, control theory, estimation theory, ...

- ▶ Stochastic methods (eg Kalman filter): estimation theory.
- ▶ Variational methods (3D-Var, 4D-Var...): control theory.

Fields of application

Meteorology (initialization of a forecast): sole field of application until early 1990ies.

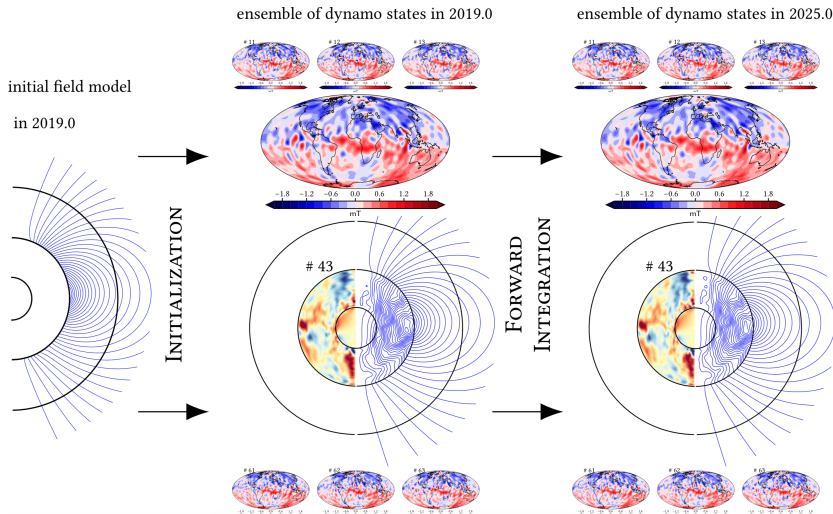
Today:

- ▶ atmospheric chemistry
- ▶ oceanic biochemistry
- ▶ glaciology
- ▶ physical oceanography
- ▶ geomagnetism
- ▶ solar magnetism
- ▶ seismology
- ▶ ...

A variety of purposes

- ▶ the estimation of the trajectory of a system to forecast its subsequent evolution
- ▶ the identification of systematic errors in numerical models
- ▶ the estimation of unobserved field variables (e.g. the magnetic field inside Earth's core)
- ▶ the estimation of parameters (e.g. a structural Earth model in seismology)
- ▶ the optimization of observation networks

An example in geomagnetism: A candidate secular variation model for IGRF-13



Fournier et al. [2021]

Question break

A scalar example

Assume we have two distinct measurements,

$$y_1 = 1$$

and

$$y_2 = 2,$$

of the same unknown quantity x .

What estimate of its true value can we make?

Answer online on wooclap

First approach

We seek x which minimizes $(x - 1)^2 + (x - 2)^2$, and we find the estimate

$$\hat{x} = 3/2 = 1.5$$

(this is the least-squares solution).

This solution has the following problems:

- ▶ it is sensitive to any change of units. If $y_1 = 1$ is a measurement of x and $y_2 = 4$ is a measurement of $2x$, then minimizing $(x - 1)^2 + (2x - 4)^2$ leads to $\hat{x} = 9/5 = 1.8$.
- ▶ it does not reflect the quality of the various measurements.

Question break

Reformulation in a statistical framework

We define

$$Y_i = x + \epsilon_i, \quad (1)$$

where the observation errors ϵ_i satisfy the following hypotheses

- ▶ $E(\epsilon_i) = 0$ (unbiased measurements)
- ▶ $\text{Var}(\epsilon_i) = \sigma_i^2$ (accuracy is known)
- ▶ $\text{Covar}(\epsilon_1, \epsilon_2) = 0$, i.e. $E(\epsilon_1 \epsilon_2) = 0$, errors are independent.

We next seek an estimator (i.e. a random variable) $\widehat{X}$ which is

- ▶ linear: $\widehat{X} = \alpha_1 Y_1 + \alpha_2 Y_2$
- ▶ unbiased: $E(\widehat{X}) = x$
- ▶ of minimum variance: $\text{Var}(\widehat{X})$ minimal (optimal accuracy)

This estimator is called the **BLUE**: Best Linear Unbiased Estimator. To compute the α_i we use the unbiased hypothesis

$$\mathbb{E}(\widehat{X}) = x = (\alpha_1 + \alpha_2)x + \alpha_1 \mathbb{E}(\epsilon_1) + \alpha_2 \mathbb{E}(\epsilon_2) = (\alpha_1 + \alpha_2)x, \quad (2)$$

so that $\alpha_1 + \alpha_2 = 1$, or $\alpha_2 = 1 - \alpha_1$. Next we compute the variance of $\widehat{X}$.

$$\begin{aligned} \text{Var}(\widehat{X}) &= \mathbb{E} \left[\left(\widehat{X} - x \right)^2 \right] = \mathbb{E} \left[(\alpha_1 \epsilon_1 + \alpha_2 \epsilon_2)^2 \right] \\ &= \alpha_1^2 \mathbb{E}(\epsilon_1^2) + 2\alpha_1 \alpha_2 \mathbb{E}(\epsilon_1 \epsilon_2) + \alpha_2^2 \mathbb{E}(\epsilon_2^2) \\ &= \alpha_1^2 \sigma_1^2 + \alpha_2^2 \sigma_2^2 \\ &= \alpha_1^2 \sigma_1^2 + (1 - \alpha_1)^2 \sigma_2^2. \end{aligned}$$

Our estimator $\widehat{X}$ has to minimize this quantity.

Computing α_1 such that

$$\frac{d}{d\alpha_1} \text{Var}(\widehat{X}) = 0 \quad (3)$$

yields

$$\alpha_1 = \frac{\sigma_2^2}{\sigma_2^2 + \sigma_1^2}. \quad (4)$$

It follows that

$$\widehat{X} = \frac{\sigma_2^2}{\sigma_1^2 + \sigma_2^2} y_1 + \frac{\sigma_1^2}{\sigma_1^2 + \sigma_2^2} y_2. \quad (5)$$

Note that we get the same result if we try to minimize the functional

$$\mathcal{J}(x) = \frac{1}{2} \left[\frac{(x - y_1)^2}{\sigma_1^2} + \frac{(x - y_2)^2}{\sigma_2^2} \right]. \quad (6)$$

- ▶ This statistical approach solves the problem of sensitivity to units and it incorporates measurement accuracies.
- ▶ The accuracy of the estimator is given by the second derivative of $\mathcal{J}$

$$\left. \frac{d^2 \mathcal{J}}{dx^2} \right|_{x=\hat{X}} = \frac{1}{\text{Var}(\hat{X})} = \frac{1}{\sigma_1^2} + \frac{1}{\sigma_2^2}, \quad (7)$$

so that “accuracies are added”.

- ▶ If we consider that $y_1 = x^b$ is a first guess of x (with standard deviation $\sigma_b = \sigma_1$) and $y_2 = y$ is an additional observation (with std dev $\sigma = \sigma_2$), then we can rearrange Eq. (5) as

$$\hat{X} = x^b + \frac{\sigma_b^2}{\sigma^2 + \sigma_b^2} (y - x^b). \quad (8)$$

The quantity $y - x^b$ is called the **innovation**. It contains the additional information provided by y with respect to x^b .

Question break

Data assimilation methods

Two classes of methods

- ▶ statistical methods: direct computation of the BLUE thanks to algebraic computations (the Kalman filter and its variants);
- ▶ variational methods: minimization of the functional $\mathcal{J}$ (4D-Var).

Shared properties

- ▶ they provide the same result (in the linear case);
- ▶ their optimality can only be demonstrated in the linear case;

Shared difficulties

- ▶ accounting for non-linearities
- ▶ dealing with large problems
- ▶ error statistics are required but sometimes only poorly known

Notations

There exists some sort of standard notations, summarized by [Ide et al. \[1997\]](#).

- ▶ $\mathbf{x}$ state vector
- ▶ $\mathbf{x}^t$ true state
- ▶ $\mathbf{x}^b$ background state
- ▶ $\mathbf{x}^a$ analyzed state

Superscripts denote vector types, subscripts refer to space or time. In the following: unless otherwise noted, all vectors will be column vectors. If $\mathbf{a}$ and $\mathbf{b}$ are two column vectors of equal size n , with the superscript T denoting transposition, then

$$\mathbf{a}^T \mathbf{b} \quad \text{is their scalar product} = \sum a_i b_i, \quad (9)$$

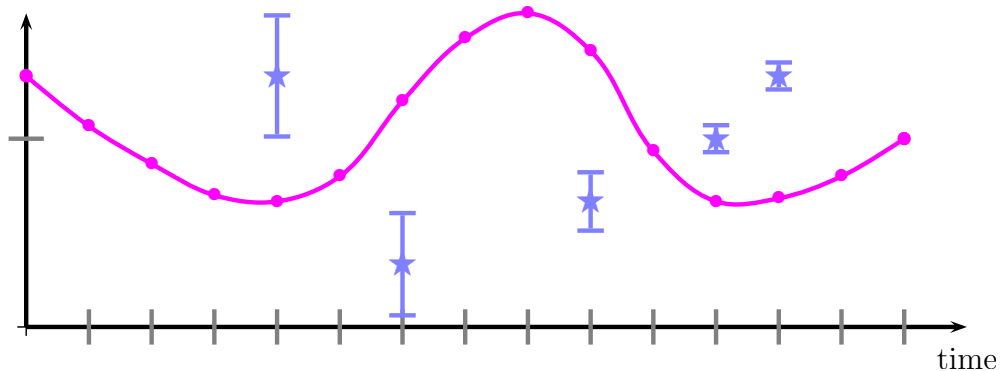
$$\mathbf{a} \mathbf{b}^T \quad \text{is a matrix of coefficients } a_i b_j, (i, j) \in \{1, \dots, n\}^2. \quad (10)$$

Assimilation in a dynamical framework

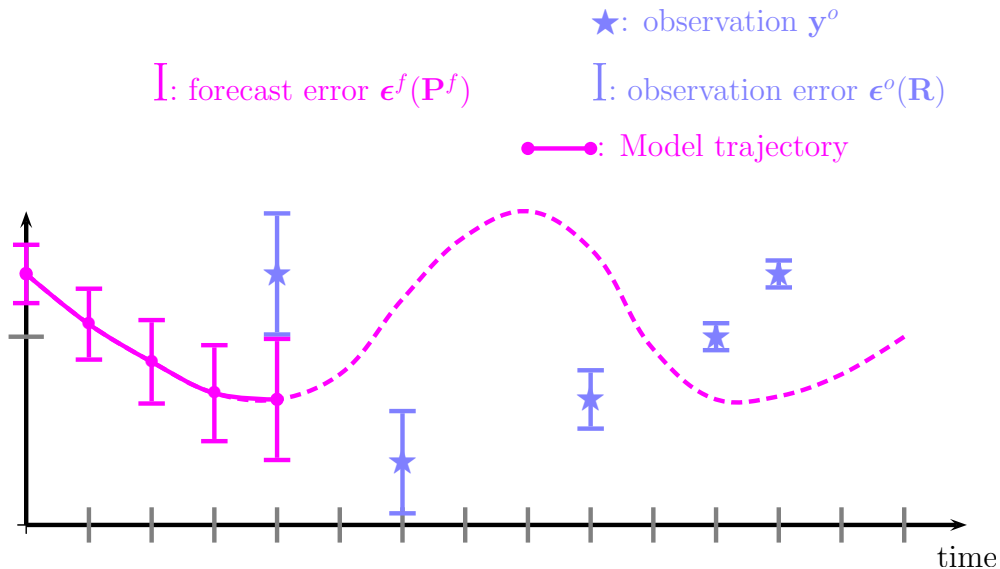
★: observation y^o

I: observation error $\epsilon^o(\mathbf{R})$

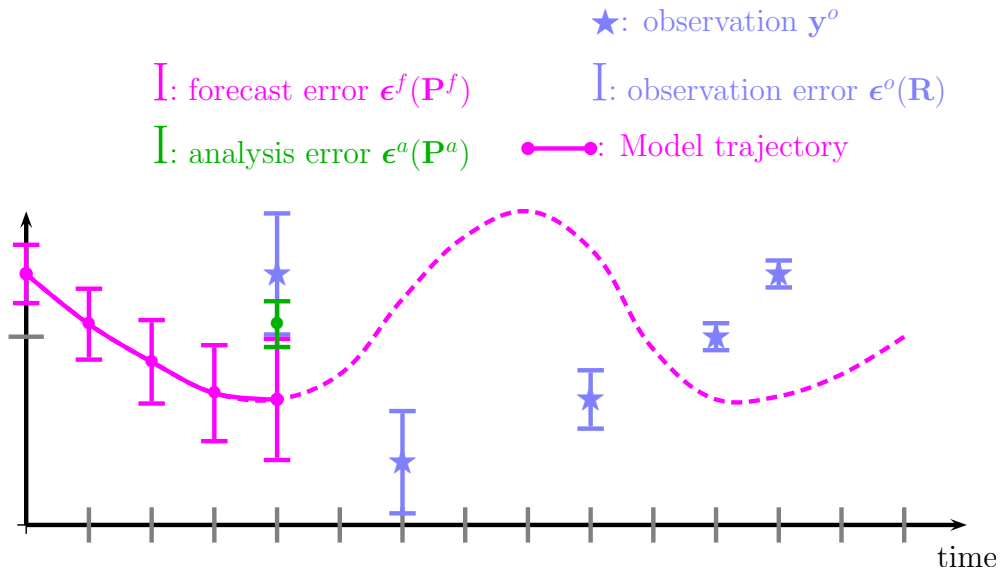
●—●: Model trajectory



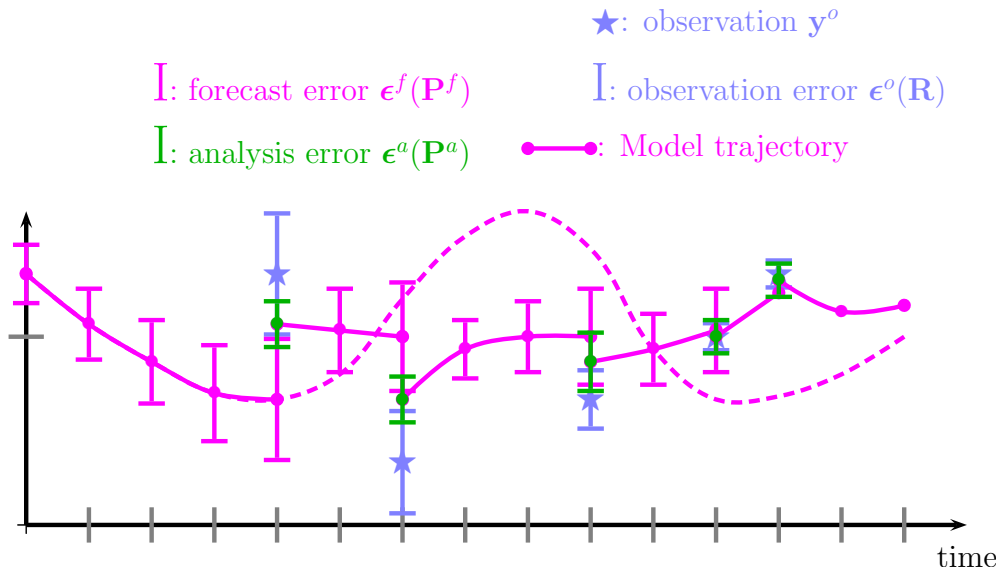
Sequential assimilation



Sequential assimilation



Sequential assimilation

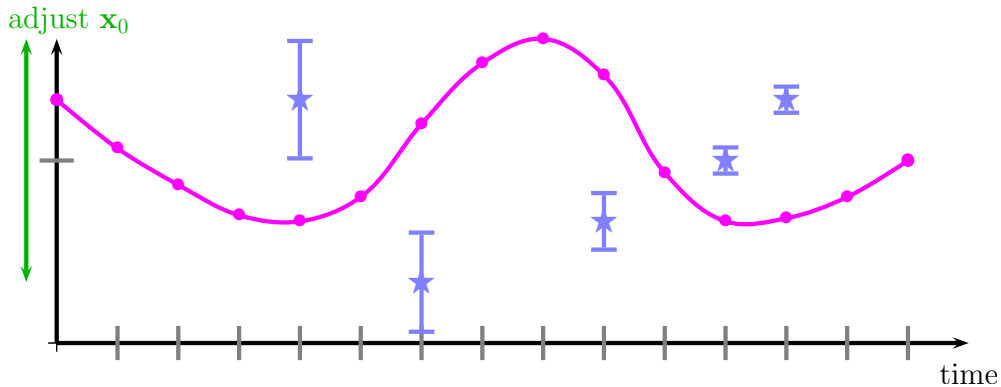


Variational assimilation

★: observation y^o

I: observation error $\epsilon^o(\mathbf{R})$

●—●: Model trajectory

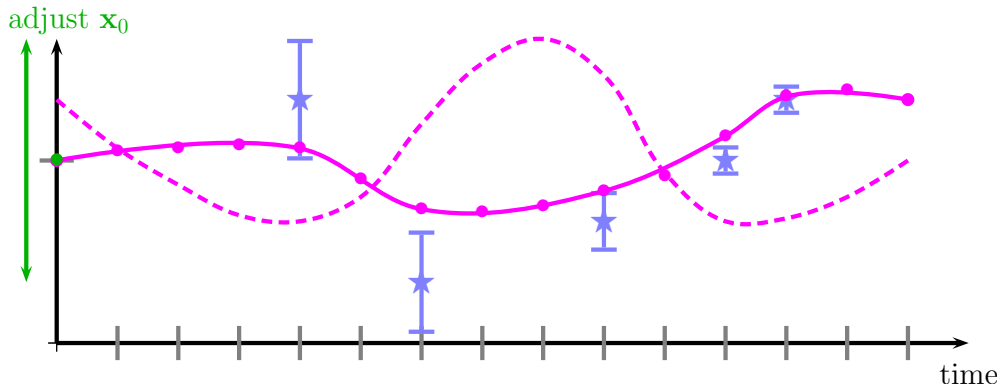


Variational assimilation, 4D-Var (adjoint)

★: observation y^o

I: observation error $\epsilon^o(\mathbf{R})$

●—●: Model trajectory



Pros & cons

Pros

- ▶ KF: error estimate after analysis
- ▶ 4D-Var: all observations in one sweep

Cons

- ▶ KF: Nonlinearities & problem size. EnKF
- ▶ 4D-Var: Implementation of adjoint model (Automatic differentiation)
- ▶ 4D-Var: Nonlinearities (tangent linear model) & perfect model assumption

The ensemble Kalman filter

Evensen [1994]

Concept

1. Have an ensemble of dynamical states evolve concurrently
2. Use this ensemble to generate (on-the-fly) the statistics needed for the analysis of the stream of observations
3. Well-suited for moderately nonlinear problems

The ensemble Kalman filter: Theory

1. Forecast:

$$\mathbf{x}_{i,e}^f = \mathcal{M}_{i-1,i}(\mathbf{x}_{i-1,e}^a) + \boldsymbol{\eta}_{i,e}, \quad e = \{1, \dots, N_e\}.$$

2. Analysis:

$$\langle \mathbf{x}_i^f \rangle = \frac{1}{N_e} \sum_{e=1}^{N_e} \mathbf{x}_{i,e}^f,$$

$$\mathbf{P}_i^f = \frac{1}{N_e - 1} \sum_{e=1}^{N_e} \left(\mathbf{x}_{i,e}^f - \langle \mathbf{x}_i^f \rangle \right) \left(\mathbf{x}_{i,e}^f - \langle \mathbf{x}_i^f \rangle \right)^T,$$

$$\mathbf{H}_i \mathbf{P}_i^f = \frac{1}{N_e - 1} \sum_{e=1}^{N_e} \left[\mathcal{H}_i \left(\mathbf{x}_{i,e}^f \right) - \mathcal{H}_i \left(\langle \mathbf{x}_i^f \rangle \right) \right] \left[\mathbf{x}_{i,e}^f - \langle \mathbf{x}_i^f \rangle \right]^T$$

$$\mathbf{H}_i \mathbf{P}_i^f \mathbf{H}_i^T = \frac{1}{N_e - 1} \sum_{e=1}^{N_e} \left[\mathcal{H}_i \left(\mathbf{x}_{i,e}^f \right) - \mathcal{H}_i \left(\langle \mathbf{x}_i^f \rangle \right) \right] \left[\mathcal{H}_i \left(\mathbf{x}_{i,e}^f \right) - \mathcal{H}_i \left(\langle \mathbf{x}_i^f \rangle \right) \right]^T$$

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

$$\mathbf{y}_{i,e}^o = \mathbf{y}_i^o + \boldsymbol{\epsilon}_e^o, \quad e = \{1, \dots, N_e\},$$

$$\mathbf{x}_{i,e}^a = \mathbf{x}_{i,e}^f + \mathbf{K}_i \left[\mathbf{y}_{i,e}^o - \mathcal{H}_i \left(\mathbf{x}_{i,e}^f \right) \right], \quad e = \{1, \dots, N_e\}.$$

The ensemble Kalman filter: Practice with the Lorenz model

Lorenz [1963]

This model is the canonical example of a set of coupled deterministic, nonlinear, ordinary differential equations (ode) able to exhibit chaotic behaviour. It is an admittedly simplified, 3-variable representation of atmospheric cellular convection, based on the earlier work of Saltzman [1962]. The state vector $\mathbf{x}$ has only three components $\mathbf{x} \equiv [X, Y, Z]^T$. Its time evolution is governed by the following set of nondimensional equations

$$\frac{dX}{dt} = -\text{Pr}(X - Y), \quad (11)$$

$$\frac{dY}{dt} = -XZ + rX - Y, \quad (12)$$

$$\frac{dZ}{dt} = XY - bZ, \quad (13)$$

which has to be supplemented with a (column) vector of initial conditions

$$\mathbf{x}_0 = [X(t=0), Y(t=0), Z(t=0)]^T.$$

The ensemble Kalman filter: Practice with the Lorenz model

The variable X is connected with the streamfunction describing atmospheric flow, while both variables Y and Z are connected with the temperature deviation responsible for convection (for further details, consult [Lorenz \[1963\]](#) – a great read). So the state of atmospheric quiescence (no convection) is described by $\mathbf{x} = [0, 0, 0]^T$.

Three nondimensional numbers define the parameter space – they are the **control parameters**:

- ▶ Pr , the Prandtl number, which is the ratio of kinematic viscosity to thermal diffusivity
- ▶ r , which is the ratio of the Rayleigh number Ra to the critical value of the Rayleigh number Ra_c (in convection parlance, r tells you how many times supercritical the system is).
- ▶ b , which is a geometrical factor.

(Note that time had been non-dimensionalized using the thermal diffusion timescale as reference timescale.) Lorenz's original choice of parameters was:

$$Pr = 10, \tag{14}$$

$$r = 28, \tag{15}$$

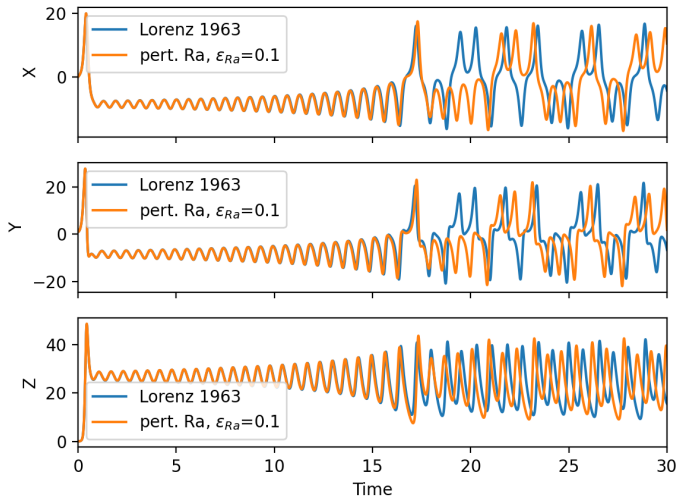
$$b = 8/3. \tag{16}$$

The ensemble Kalman filter: Practice with the Lorenz model

Outline of the lab

1. Get familiar with the Lorenz model: sensitivity to control parameters and initial condition
2. Decide on a set of control parameters and an initial condition to define the true dynamical state $\mathbf{x}^t(t)$
3. Generate a database of observations: specify how often obs are performed, which variables are observed, the level of error that impacts the observation
4. Run the EnKF starting with an ensemble centered on a zero initial condition and assess how the success (or failure) of the assimilation depends on
 - ▶ the observations (how often, which variables, noise level)
 - ▶ the EnKF (size of the ensemble, initial spread)
 - ▶ the control parameters (assuming that you do not have them right)

Sensitivity to control parameters



Useful references

- ▶ “Discrete Inverse and State Estimation Problems”, by Wunsch [2006], provides a very personal and powerful account of adjoint methods and their application in geophysical fluid dynamics (oceanography).
- ▶ In her book entitled “Atmospheric Modelling, Data Assimilation and Predictability”, E. Kalnay [2003] has two comprehensive and very well-written chapters on the basics and applications of data assimilation techniques to atmospheric dynamics.
- ▶ In addition, Evensen [2009] provides a very complete treatment of data assimilation techniques, with a strong and useful emphasis on the basics and applications of the ensemble Kalman filter he invented.
- ▶ Last, but not least, Blayo et al. [2014] is an excellent compilation covering theoretical and practical aspects of data assimilation in geosciences.

More references

For a start, I would highly recommend the review paper by [Talagrand \[1997\]](#), “Assimilation of observations, an introduction” which provides an extremely concise and well-written overview of the topic. In addition, if you are looking for references related to the geophysical inverse problem in general, [Parker \[1994\]](#) and [Tarantola \[2005\]](#) provide two very personal and insightful views on how we should go about making inference on the Earth based on a finite number of noisy observations and on physical laws governing its behaviour.

References

- Eric Blayo, Marc Bocquet, Emmanuel Cosme, and Leticia F Cugliandolo. Advanced data assimilation for geosciences. In *International Summer School-Advanced Data Assimilation for Geosciences*, page 608. Oxford University Press, 2014.
- G. Evensen. Sequential data assimilation with a nonlinear quasi-geostrophic model using Monte Carlo methods to forecast error statistics. *Journal of Geophysical Research*, 99 (C5):10143–10162, 1994. doi:[10.1029/94JC00572](https://doi.org/10.1029/94JC00572).
- G. Evensen. *Data assimilation: The ensemble Kalman filter*. Springer, Berlin, 2 edition, 2009. doi:[10.1007/978-3-642-03711-5](https://doi.org/10.1007/978-3-642-03711-5).
- Alexandre Fournier, Julien Aubert, Vincent Lesur, and Guillaume Ropp. A secular variation candidate model for IGRF-13 based on Swarm data and ensemble inverse geodynamo modelling. *Earth, Planets and Space*, 73:43, 2021. doi:[10.1186/s40623-020-01309-9](https://doi.org/10.1186/s40623-020-01309-9).
- K. Ide, P. Courtier, M. Ghil, and A. C. Lorenc. Unified notation for data assimilation: Operational, sequential and variational. *Journal of the meteorological society of Japan*, 75: 181–189, 1997. doi:[10.2151/jmsj1965.75.1B_181](https://doi.org/10.2151/jmsj1965.75.1B_181).
- E. Kalnay. *Atmospheric modeling, data assimilation, and predictability*. Cambridge University Press, Cambridge, 2003.
- E. N. Lorenz. Deterministic Nonperiodic Flow. *Journal of the Atmospheric Sciences*, 20(2):130–141, 1963. doi:[10.1175/1520-0469\(1963\)020<0130:DNF>2.0.CO;2](https://doi.org/10.1175/1520-0469(1963)020<0130:DNF>2.0.CO;2).
- R. L. Parker. *Geophysical inverse theory*. Princeton University Press, Princeton, NJ, 1994.
- B. Saltzman. Finite amplitude free convection as an initial value problem–I. *Journal of the Atmospheric Sciences*, 19(7):329–341, 1962. doi:[10.1175/1520-0469\(1962\)019<0329:FAFCAA>2.0.CO;2](https://doi.org/10.1175/1520-0469(1962)019<0329:FAFCAA>2.0.CO;2).
- O. Talagrand. Assimilation of observations, an introduction. *Journal of the Meteorological Society of Japan*, 75(1B):191–209, 1997. doi:[10.2151/jmsj1965.75.1B_191](https://doi.org/10.2151/jmsj1965.75.1B_191).
- A. Tarantola. *Inverse problem theory and methods for model parameter estimation*. Society for Industrial Mathematics, Philadelphia, PA, 2005. doi:[10.1137/1.9780898717921](https://doi.org/10.1137/1.9780898717921).
- C. Wunsch. *Discrete inverse and state estimation problems*. Cambridge University Press, Cambridge, 2006.