Data assimilation in the geosciences

An overview

Alberto Carrassi[1,2], Marc Bocquet[3] and **Laurent Bertino**[1]

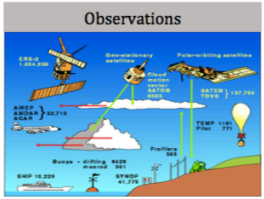
(1): Nansen Environmental and Remote Sensing Center - Norway
(2): Geophysical Institute, University of Bergen - Norway
(3): CEREA, joint lab École des Ponts ParisTech and EdF R&D, IPSL, France

UNIVERSITETET I BERGEN
Geofysisk institutt

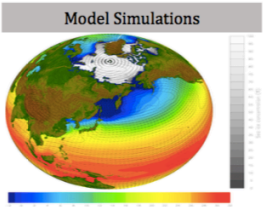
École des Ponts
ParisTech

OceanPredict'19, Halifax, May 2019

Data Assimilation

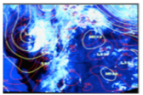


Data Assimilation best combines model and observations and brings synergy

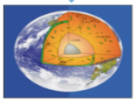


An on-going rapid expansion from **Weather Science (NWP)** into **Climate Science/Geophysics** in general:

- Oceanography
- Climate Prediction
- Climate Assessment
- Hydrology
- Geology
- Climatology
- Detection & Attribution
- ... and many more beyond geosciences ...



NWP



Geophysics

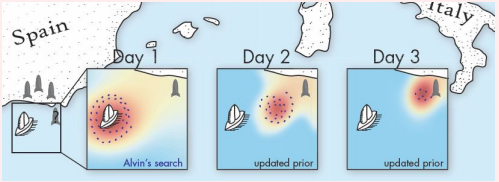
Outline of part I

Posing the problem: sequential inference

- *Inference* is the process of taking a decision based on limited information.
- Limitations arise by incomplete and noisy data and by an approximate knowledge about the laws (if any) governing the system evolution.
- The problem we intend to solve is the estimation of the state of a system, at any arbitrary past, present and/or future times.
- *Sequential inference* is the problem of updating our knowledge about the system each time new data becomes available.

Posing the problem

An example: *Palomares (Spain) incident*



From *Berkeley science review* at
http://berkeleysciencereview.com/article/toolbox/

- On January 17$^{\text{th}}$, 1966, a US Air Force bomber flying over the south of Spain, with four hydrogen bombs, exploded in midair.
- Three bombs were recovered, undetonated, on land, while the fourth was lost.
- According to a local fisherman, it splashed down somewhere in the Mediterranean Sea.
- How would you go about looking for the bomb in a way that maximizes the chance you will find it?

▶ This lecture treats the case of the *discrete-model/discrete-observation* estimation problem (relevant in many practical cases such as climate science, biology among others).

▶ Complete treatments of the *continuous-continuous* and *discrete-continuous* cases can be found in many textbooks on estimation theory (*e.g.*, *Jazwinski*, 1970; *Bain and Crisan*, 2009).

The model: *what we know about the system physical-dynamical laws*

We will assume that a model of the natural process of interest is available as a discrete stochastic-dynamical system,

$$\mathbf{x}_k = \mathcal{M}_{k:k-1}(\mathbf{x}_{k-1}, \boldsymbol{\lambda}) + \boldsymbol{\eta}_k. \tag{1}$$

- $\mathbf{x}_k \in \mathbb{R}^m$ and $\boldsymbol{\lambda} \in \mathbb{R}^p$ are the model state and parameter vectors respectively.
- The model parameters may include the external forcings or the boundary conditions.
- $\mathcal{M}_{k:k-1} : \mathbb{R}^m \to \mathbb{R}^m$ is usually a nonlinear, possibly *chaotic*, function from time t_{k-1} to t_k.
- $\boldsymbol{\eta}_k \in \mathbb{R}^m$ is the *model error*, represented as a stochastic additive term.

Remark: $\boldsymbol{\eta}_k$ could be included into the parentheses without loss of generality.

Remark: The *stochastic difference model*, Eq. (1), has a continuous-time counterpart ($\Delta t \to 0$); it is known as the *Itô stochastic differential equation* (see, *e.g. Jazwinski*, 1970; *Reich and Cotter*, 2015)

The model: *why stochastic?*

▶ **Imperfect model** - $\mathcal{M}_{k:k-1}$ embeds our knowledge about the laws governing the process $\Longrightarrow$ Such a knowledge is always (in realistic cases) partial and/or incorrect.

▶ **Numerical discretization** - $\mathcal{M}_{k:k-1}$ is often a spatio-temporal discretization of physical laws (*e.g.*, the Navier Stokes equations for fluids) expressed as partial differential equations on a continuous media $\Longrightarrow$ The finite resolution induces errors.

▶ **Chaos** - Many natural systems are chaotic and exhibit extreme sensitivity to initial conditions $\Longrightarrow$ Any (inevitable) error in the system state contaminates the prediction.

The sources of error are accounted for using a stochastic model

The observations and their relation with the quantities of interest

- Noisy observations, $\mathbf{y}_k \in \mathbb{R}^d$, are related to the model state vector through

$$\mathbf{y}_k = \mathcal{H}_k(\mathbf{x}_k) + \boldsymbol{\epsilon}_k. \tag{2}$$

with $\mathcal{H} : \mathbb{R}^m \to \mathbb{R}^d$ the, generally nonlinear, *observation operator* mapping from model to observational space.

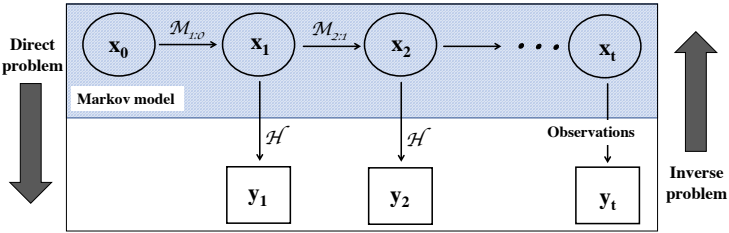
- Observational error, $\boldsymbol{\epsilon}_k$, is a stochastic term, accounting for the instrumental error, deficiencies in the formulation of $\mathcal{H}$ and the *representativity error*.

- The latter arises from the presence of unresolved scales and represents their effect on the resolved scales - it is ubiquitous in physical science even when observations and model have the same resolution (why?).

Remark: often $d \ll m$, *i.e.*, the amount of available data is insufficient to fully describe the system.

The observations and their relation with the quantities of interest

- The stochastic model dynamics, Eq. (1), together with the stochastic observation model, Eq. (2) define an **Hidden Markov model (HMM)**



- A stochastic model is said *Markov* if its future state depends only on the current state and not on any states the models has attained before.

- We focus here on the **inverse problem** $\Leftrightarrow$ Estimate $\mathbf{x}$ by observing $\mathbf{y}$.

Bayesian inference for the inverse problem

- With the two complementary pieces of information in hand, *model* and *data*, we can move forward and formalize their fusion.
- When making inference we have to decide how much we trust the uncertain information. $\Rightarrow$ We need to **quantify the uncertainty**.
- Given the stochastic nature of the problem

uncertainty quantification is done using probabilities.

- The Bayesian approach offers a natural mathematical framework to understand and formalize this problem.
- In particular, the goal of Bayesian inference is to estimate the uncertainty in $\mathbf{x}$ given $\mathbf{y} \Leftrightarrow$ Compute the conditional **probability density function** (PDF) $p(\mathbf{x}|\mathbf{y})$.

Bayes' theorem

Bayes' theorem

Let $\mathbf{x}$ and $\mathbf{y}$ be jointly distributed random vectors with joint PDF, $p(\mathbf{x}, \mathbf{y})$. Then

$$p(\mathbf{x}|\mathbf{y}) = \frac{p(\mathbf{y}|\mathbf{x})p(\mathbf{x})}{p(\mathbf{y})}\,. \tag{3}$$

- An algebraic equation for conditional probabilities $\Leftrightarrow$ The probability that the event, $\mathbf{x}$, occurs, knowing that another one, $\mathbf{y}$, has occurred.
- The output of the estimation process is the **posterior distribution** $p(\mathbf{x}|\mathbf{y})$
- $p(\mathbf{x})$ is the **prior PDF** that gathers all the knowledge before assimilating the new observations. It is a distinctive feature of the Bayesian approach.

Bayes' theorem

- $p(\mathbf{y}|\mathbf{x})$ is the **likelihood of the data** conditioned on the state $\mathbf{x}$ (*i.e.*, it quantifies the likelihood of observing $\mathbf{y}$ given a particular value of $\mathbf{x}$).
- $p(\mathbf{y})$ is the marginal distribution of the data, $p(\mathbf{y}) = \int \mathrm{d}\mathbf{x}\, p(\mathbf{y}|\mathbf{x})p(\mathbf{x})$. It integrate to one and is treated as a normalization coefficient.

Remark: The factor $p(\mathbf{y})$ is relevant, for instance, in model selection problems and is also referred to as **model evidence**.

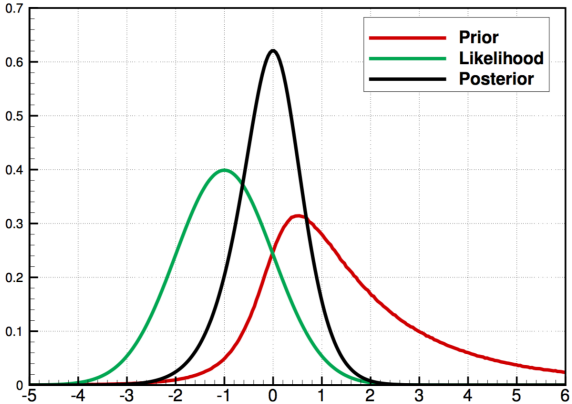


Figure: Courtesy of Geir Evensen

Sequential Bayesian estimate

- Recall our HMM given by the dynamical model, Eq. (1), data model, Eq. (2)

$$\mathbf{x}_k = \mathcal{M}_{k:k-1}(\mathbf{x}_{k-1}) + \boldsymbol{\eta}_k\,, \qquad \mathbf{y}_k = \mathcal{H}_k(\mathbf{x}_k) + \boldsymbol{\epsilon}_k$$

- The model and the observational errors, $\{\boldsymbol{\eta}_k, \boldsymbol{\epsilon}_k : k = 1, \ldots, K\}$ are assumed to be uncorrelated in time, mutually independent, and distributed according to the PDFs $p_{\boldsymbol{\eta}}$ and $p_{\boldsymbol{\epsilon}}$

- Let us define the sequences of system states and observations within the interval $[t_0, t_K]$ as $\mathbf{x}_{K:0} = \{\mathbf{x}_K, \mathbf{x}_{K-1}, ..., \mathbf{x}_0\}$ and $\mathbf{y}_{K:1} = \{\mathbf{y}_K, \mathbf{y}_{K-1}, ..., \mathbf{y}_1\}$ respectively.

We wish to estimate the posterior $p(\mathbf{x}_{K:0}|\mathbf{y}_{K:1})$ for any arbitrary, sequentially increasing, t_K.

Using Bayes's law we have

$$p(\mathbf{x}_{K:0}|\mathbf{y}_{K:1}) \propto p(\mathbf{y}_{K:1}|\mathbf{x}_{K:0})p(\mathbf{x}_{K:0}) \tag{4}$$

Prediction, *filtering* and *smoothing*

Depending on which time period is needed for state estimation, it is possible to define *three estimation problems*:

1. **Prediction**: estimate $p(\mathbf{x}_l|\mathbf{y}_{k:1})$ with $l > k$.
2. **Filtering**: estimate $p(\mathbf{x}_k|\mathbf{y}_{k:1})$.
3. **Smoothing**: estimate $p(\mathbf{x}_{K:0}|\mathbf{y}_{K:1})$.

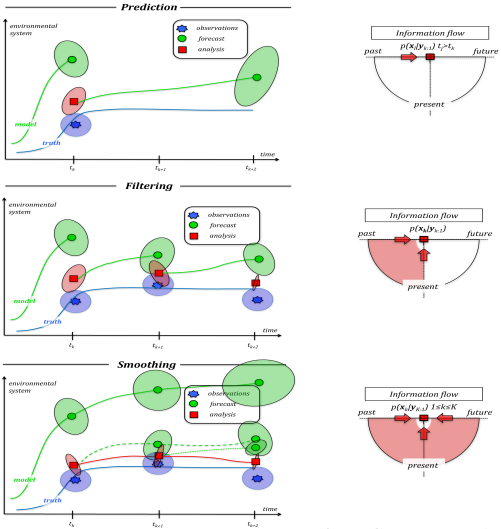


Figure from Carrassi *et al.*, 2018

Formal Bayesian solutions - **Prediction**

▶ Given the conditional PDF $p(\mathbf{x}_k|\mathbf{y}_{k:0})$, we seek a law to propagate it forward in time under the effect of the model dynamics, Eq. (1).

This leads to the **Chapman-Kolmogorov equation** for the propagation of a PDF under the model dynamics, Eq. (1), as

$$p(\mathbf{x}_l|\mathbf{y}_{k:1}) = \int_{\mathbb{R}^m} \mathrm{d}\mathbf{x}_k \, p(\mathbf{x}_l|\mathbf{x}_k) p(\mathbf{x}_k|\mathbf{y}_{k:1}), \tag{5}$$

where the initial PDF at t_k is given by $p(\mathbf{x}_k|\mathbf{y}_{k:1})$, and $p(\mathbf{x}_l|\mathbf{x}_k) = p_{\boldsymbol{\eta}}[\mathbf{x}_l - \mathcal{M}_{l:k}(\mathbf{x}_k)]$ in our HMM model.

▶ The **Prediction** problem is addressed by solving the Chapman-Kolmogorov equation (5), given the conditional PDF at time t_k, $p(\mathbf{x}_k|\mathbf{y}_{k:1})$.

Remark: In case of model dynamics given as a stochastic differential equation, instead of a stochastic difference equation as in Eq. (1), the Chapman-Kolmogorov equation becomes the *Fokker-Planck* equation.

Formal Bayesian solutions - **Filtering**

▶ **Filtering problem** is the most common in applications, and is characterized by sequential processing, in which measurements are utilized as they become available.

▶ An analysis step, in which the conditional PDF $p(\mathbf{x}_k|\mathbf{y}_{k:1})$ is updated using the latest observation, $\mathbf{y}_k$,

Analysis

$$p(\mathbf{x}_k|\mathbf{y}_{k:1}) \propto p_{\boldsymbol{\epsilon}}[\mathbf{y}_k - \mathcal{H}_k(\mathbf{x}_k)]p(\mathbf{x}_k|\mathbf{y}_{k-1:1}), \tag{6}$$

▶ alternates with a forecast step which propagates this PDF, using Chapman-Kolmogorov equation, forward until the time of a new observation,

Forecast

$$p(\mathbf{x}_{k+1}|\mathbf{y}_{k:1}) = \int_{\mathbb{R}^m} \mathrm{d}\mathbf{x}_k \, p_{\boldsymbol{\eta}}[\mathbf{x}_{k+1} - \mathcal{M}_{k+1:k}(\mathbf{x}_k)]p(\mathbf{x}_k|\mathbf{y}_{k:1}), \tag{7}$$

to get $p(\mathbf{x}_{k+1}|\mathbf{y}_{k:1})$.

Formal Bayesian solutions - **Smoothing**

- Smoothing is relevant when, for instance, one is interested in a retrospective analysis after the observations have been collected.
- The goal is to estimate the conditional PDF, $p(\mathbf{x}_k|\mathbf{y}_{K:1})$ of the state at any time t_k, $0 \leq k \leq K$, based on all observations (past, present and future).
- First write the smoothing PDF at time t_k by marginalizing over $\mathbf{x}_{k+1}$

$$p(\mathbf{x}_k|\mathbf{y}_{K:1}) = \int_{\mathbb{R}^m} \mathrm{d}\mathbf{x}_{k+1}\, p(\mathbf{x}_k|\mathbf{x}_{k+1}, \mathbf{y}_{K:1}) p(\mathbf{x}_{k+1}|\mathbf{y}_{K:1}). \tag{8}$$

Note that, using Bayes' rule, the integrand in Eq. (8) can be written as

$$p(\mathbf{x}_k|\mathbf{x}_{k+1}, \mathbf{y}_{K:1}) = p(\mathbf{x}_k|\mathbf{x}_{k+1}, \mathbf{y}_{k:1}) \propto p(\mathbf{x}_{k+1}|\mathbf{x}_k) p(\mathbf{x}_k|\mathbf{y}_{k:1}), \tag{9}$$

given that the observations $\{\mathbf{y}_{k+1}, \ldots, \mathbf{y}_K\}$ are independent of $\mathbf{x}_k$ when $\mathbf{x}_{k+1}$ is known.

- Finally remark that $p(\mathbf{x}_k|\mathbf{y}_{k:1})$ in Eq. (9) is the filter solution at time t_k.

Formal Bayesian solutions - Smoothing

- This implies that the smoothing PDFs can be obtained using a

  forward-backward recursive algorithm.

- *Forward phase* - Start from $p(\mathbf{x}_0)$ then from $k = 1$ to $k = K$
  - Estimate and store the filter PDFs, $p(\mathbf{x}_k|\mathbf{y}_{k:1})$.
- *Backward phase* - From $k = K - 1$ to $k = 1$
  - Compute $p(\mathbf{x}_k|\mathbf{x}_{k+1}, \mathbf{y}_{k:1})$ with Eq. (9) using the stored filter PDFs, $p(\mathbf{x}_k|\mathbf{y}_{k:1})$ and $p(\mathbf{x}_{k+1}|\mathbf{x}_k)$.
  - Obtain the smoothing PDFs, $p(\mathbf{x}_k|\mathbf{y}_{K:1})$, from Eq. (8) making use of the smoothing PDF at time t_{k+1}, $p(\mathbf{x}_{k+1}|\mathbf{y}_{K:1})$ estimated at the previous iteration.

Comments on Bayesian filter and smoother

- The filter solution t_k $(0 \leq k \leq K)$ is obtained by sequential updating until $t_k \Rightarrow$ it thus accounts for all observations until t_k.
- In contrast, the smoothing solution also accounts for future observations until $t_K \Rightarrow$ it is thus generally more accurate than the filtering one.
- At the final time t_K both solutions have incorporated the same amount of data $\Rightarrow$ in the absence of approximations, they will coincide.
- In general the filtering and smoothing do not possess analytical solutions.
- However, when the dynamical and observational model are linear and all error PDFs are Gaussian an analytic solutions exists: the famous **Kalman filter** and **Kalman smoother**.

The Gaussian and linear approximation

- The huge dimension of models and datasets hampers the use of a fully Bayesian approach in geosciences (curse of dimensionality).
- It is usually assumed that observation and model error are Gaussian distributed $\Longrightarrow$ PDFs can be described completely in terms of their mean and covariance.
- The Gaussian approximation is at the core of most of the DA procedures successfully used in the geosciences.

Let assume that the dynamical and observational models are both linear

$$\mathbf{x}_k = \mathbf{M}_{k:k-1}\mathbf{x}_{k-1} + \boldsymbol{\eta}_k, \qquad \boldsymbol{\eta}_k \sim \mathcal{N}(\mathbf{0}, \mathbf{Q}_k), \tag{10}$$

$$\mathbf{y}_k = \mathbf{H}_k\mathbf{x}_k + \boldsymbol{\epsilon}_k, \qquad \boldsymbol{\epsilon}_k \sim \mathcal{N}(\mathbf{0}, \mathbf{R}_k), \tag{11}$$

with $\mathbf{M}_{k:k-1}$ and $\mathbf{H}_k$ being matrices in $\mathbb{R}^{m\times m}$ and $\mathbb{R}^{d\times m}$, respectively.

- The observational and model noises are assumed to be white-in-time, unbiased, and Gaussian distributed with covariances $\mathbf{R}_k \in \mathbb{R}^{d\times d}$ and $\mathbf{Q}_k \in \mathbb{R}^{m\times m}$ respectively.

The Kalman filter

The **KF recursive equations** reads

Forecast Step

$$\mathbf{x}_k^{\mathrm{f}} = \mathbf{M}_{k:k-1}\mathbf{x}_{k-1}^{\mathrm{a}}, \tag{12}$$

$$\mathbf{P}_k^{\mathrm{f}} = \mathbf{M}_{k:k-1}\mathbf{P}_{k-1}^{\mathrm{a}}\mathbf{M}_{k:k-1}^{\mathrm{T}} + \mathbf{Q}_k. \tag{13}$$

Analysis step

$$\mathbf{K}_k = \mathbf{P}_k^{\mathrm{f}}\mathbf{H}_k^{\mathrm{T}}(\mathbf{H}_k\mathbf{P}_k^{\mathrm{f}}\mathbf{H}_k^{\mathrm{T}} + \mathbf{R}_k)^{-1}, \tag{14}$$

$$\mathbf{x}_k^{\mathrm{a}} = \mathbf{x}_k^{\mathrm{f}} + \mathbf{K}_k(\mathbf{y}_k - \mathbf{H}_k\mathbf{x}_k^{\mathrm{f}}), \tag{15}$$

$$\mathbf{P}_k^{\mathrm{a}} = (\mathbf{I}_k - \mathbf{K}_k\mathbf{H}_k)\mathbf{P}_k^{\mathrm{f}}. \tag{16}$$

- Given $\mathbf{Q}_k$, $\mathbf{R}_k$, $\mathbf{H}_k$ and $\mathbf{M}_k$, for $k \geq 1$, and initial condition for the mean, $\mathbf{x}_0^{\mathrm{a}} = \mathbf{x}_0$, and error covariance, $\mathbf{P}_0^{\mathrm{a}} = \mathbf{P}_0$, Eqs. (12)–(16) estimate sequentially the state and the associated error covariance at any future time $k > 1$.
- The matrix $\mathbf{K}_k \in \mathbb{R}^{m\times d}$ is the Kalman gain and contains the coefficients of the linear combination between the forecast $\mathbf{x}_k^{\mathrm{f}}$, and the observations.
- The resulting state estimate, the analysis $\mathbf{x}_k^{\mathrm{a}}$, has minimum variance and is unbiased.

The Kalman smoother

▶ A recursive estimate of the PDFs $p(\mathbf{x}_k|\mathbf{y}_{K:1})$, can be obtained using a forward and backward recursions, in which a forward-in-time filter is followed by a backward-in-time smoother.

▶ Assume that a forward in time KF has been implemented and the forecast and analysis means and covariances, $\mathbf{x}_k^{\mathrm{f/a}}$ and $\mathbf{P}_k^{\mathrm{f/a}}$, have been computed and stored.

▶ We can then run the **KS recursion** backward in time, for $k = K-1, \ldots, 1$, to compute the smoothing mean and covariance, $\mathbf{x}_k^{\mathrm{sm}}$ and $\mathbf{P}_k^{\mathrm{sm}}$, according to

$$\mathbf{S}_k = \mathbf{P}_k^{\mathrm{a}}\mathbf{M}_{k+1:k}^{\mathrm{T}}(\mathbf{M}_{k+1:k}\mathbf{P}_k^{\mathrm{a}}\mathbf{M}_{k+1:k}^{\mathrm{T}} + \mathbf{Q}_{k+1})^{-1} \quad = \mathbf{P}_k^{\mathrm{a}}\mathbf{M}_{k+1:k}^{\mathrm{T}}\mathbf{P}_{k+1}^{-\mathrm{f}}, \tag{17}$$

KS Mean
$$\mathbf{x}_k^{\mathrm{sm}} = \mathbf{x}_k^{\mathrm{a}} + \mathbf{S}_k(\mathbf{x}_{k+1}^{\mathrm{sm}} - \mathbf{x}_{k+1}^{\mathrm{f}}), \tag{18}$$

KS Covariance
$$\mathbf{P}_k^{\mathrm{sm}} = \mathbf{P}_k^{\mathrm{a}} + \mathbf{S}_k(\mathbf{P}_{k+1}^{\mathrm{sm}} - \mathbf{P}_{k+1}^{\mathrm{f}})\mathbf{S}_k^{\mathrm{T}} \tag{19}$$

Some properties of the Kalman filter and smoother

- **Time dependent prior** - The KF/KS recursions provide a time-dependent estimate of the prior that is highly desirable in chaotic systems, so that $\mathbf{P}_k^{\mathrm{f}}$ is itself strongly time-dependent.
- **Filter divergence** - It happens when the solution of the KF deviates dramatically from the true signal. When $\mathbf{H}_k\mathbf{P}_k^{\mathrm{f}}\mathbf{H}_k^{\mathrm{T}} << \mathbf{R}_k$ the filter solution may start to ignore the observations
- **A diagnostic tool** - The innovation vector sequence, $\mathbf{v}_k = \mathbf{y}_k - \mathbf{H}_k\mathbf{x}_k^{\mathrm{f}}$, is Gaussian and white-in-time: one can thus keep monitoring the innovations to asses the KF optimality and, possibly, to implement corrections.
- **Not a *learning algorithm*** - Large innovations do not increase the uncertainties.
- **Dependency on the initial condition** - The dependence on the initial error covariance, $\mathbf{P}_0$, is more critical for some models than others.

Which method for what?

- The **extended Kalman filter** (EKF) is a first-order expansion of the KF to nonlinear dynamics.
- The EKF has been successful for models showing no dynamical instabilities (soil models), but diverges for a Quasi-Geostrophic model (*Evensen*, 1992).
- Both 4DVAR and the Ensemble Kalman Filter are better suited for chaotic dynamics.

Essential bibliography

- Asch, M., M. Bocquet, and M. Nodet, *Data Assimilation: Methods, Algorithms, and Applications*, Fundamentals of Algorithms, SIAM, Philadelphia, 2016. (Chapter 3).
- Carrassi, A., M. Bocquet, L. Bertino and G. Evensen. *Data assimilation in the geosciences - An overview on methods, issues and perspectives.* WIREs Climate Change, 9:e535, 2018.
- Jazwinski, A.H., *Stochastic Processes and Filtering Theory*, Academic Press, New York, 1970. (Chapters 2 and 3).
- Reich, S., and C. Cotter, *Probabilistic Forecasting and Bayesian Data Assimilation*, Cambridge University Press, Cambridge, 2015. (Chapters 2, 4 and 5).
- Wikle and Berliner, *A Bayesian tutorial for data assimilation.* Physica D, 203, 1-16, 2007.