Lecture 1 on data assimilation: Elementary principles of geophysical data assimilation

Marc Bocquet

With help from Alberto Carrassi, Alban Farchi

CEREA, joint lab École des Ponts ParisTech and EdF R&D, Université Paris-Est, France
Institut Pierre-Simon Laplace

(marc.bocquet@enpc.fr)

Synopsis of the course

- Monday, October 28 10:30-12:30
  Lecture 1: Elementary principles of geophysical data assimilation. The Bayesian standpoint. Classical methods of data assimilation: 3D-Var, the Kalman filter, 4D-Var.
- Tuesday, October 29, 10:30-12:30
  Lecture 2: The ensemble Kalman filter and its variants (focus on the algorithmic/mathematical aspects.)
- Thursday, October 31, 10:30-12:30
  Lecture 3: Recent advances: hybrid and ensemble variational techniques. Discussion on what to expect from machine learning/deep learning.

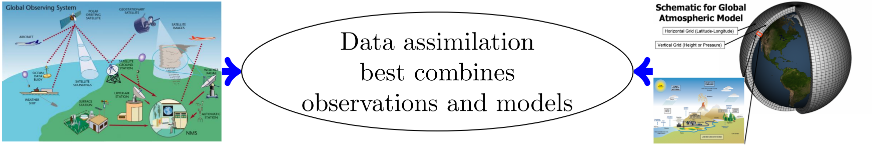
Followed next week by:

- A course on data assimilation and stochastic filtering, particle filters by Dan Crisan (Imperial College, London)
- A course on big data and uncertainty quantification by Omar Ghattas (Uni. of Texas, Austin)

Outline

1. Data assimilation: principles
   - Introduction
   - Bayesian framework
   - Goals and practical tools of data assimilation
2. Focus on a key elementary derivation
3. Main techniques
   - 3D-Var and optimal interpolation
   - The Kalman filter
   - 4D-Var
   - Particle filters
4. References

Data assimilation (DA) in the geosciences



Data assimilation
best combines
observations and models

An ongoing expansion from numerical weather prediction to the climate science/geosciences:

- Oceanography
- Atmospheric chemistry
- Climate prediction and assessment
- Glaciology
- Hydrology and hydraulics
- Geology
- Space weather
- and many other fields

Data assimilation: an inference problem

▶ Inference is the process of taking a decision based on limited information.

▶ Information comes from

- an approximate knowledge about the laws (if any) governing the time evolution of the dynamical system
- imperfect (partial, noisy, indirect) observations of this system

▶ Sequential inference is the problem of updating our knowledge about the system each time a new batch of observations becomes available.

First ingredient: the dynamical model

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

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

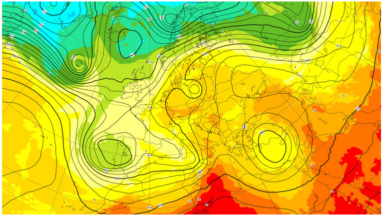
- $\mathbf{x}_k \in \mathbb{R}^{N_x}$ and $\boldsymbol{\lambda} \in \mathbb{R}^{N_\mathrm{p}}$ are the model state and parameter vectors respectively.

- $\mathcal{M}_{k:k-1} : \mathbb{R}^{N_x} \to \mathbb{R}^{N_x}$ is usually a nonlinear, possibly chaotic, map from t_{k-1} to t_k.

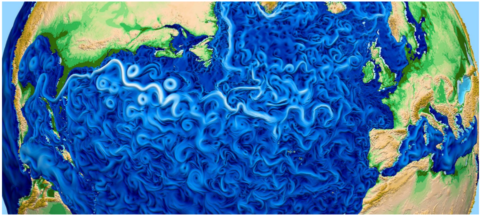
- $\boldsymbol{\eta}_k \in \mathbb{R}^{N_x}$ is the model error, represented as a stochastic additive term (more general representations are possible).

First ingredient: the dynamical model

- In the geosciences:
  - The state space dimension is huge (up to 10^9 for operational systems, up to 10^7 for research systems). A big data problem with costly models to integrate.
  - Numerical models (i.e. implementation of $\mathcal{M}$) are often computationally very costly.
  - The unstable dynamics of chaotic geofluids has implicit consequences on the design of DA algorithms: One key reason why we use sequential inference.



ECMWF IFS: Geopotential at 500hPa and temperature at 850hPa



E3SM Earth system model

Second ingredient: the observations

- Noisy observations, $\mathbf{y}_k \in \mathbb{R}^{N_y}$, are available at discrete times and are related to the model state vector through

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

with $\mathcal{H} : \mathbb{R}^{N_x} \to \mathbb{R}^{N_y}$ being the (generally nonlinear) observation operator mapping from the model to the observational space.

- The observation error, $\boldsymbol{\epsilon}_k$, is represented as a stochastic term. It account for the instrumental error, for deficiencies in the formulation of $\mathcal{H}$, and for the representation error.

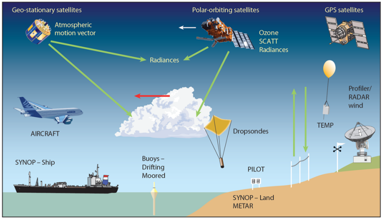
- The representation error arises from the presence of unresolved scales and represents their effect on the resolved scales – it is ubiquitous in physical science and inherent to the discretisation procedure [Janjić et al. 2018].

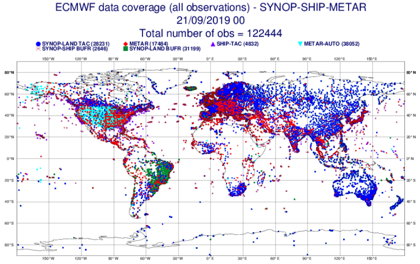
- We assume that the observation dimension is constant, so that $N_y(k) \equiv N_y$ (the generalisation is simple). Remark: often $N_y \ll N_x$, i.e. the amount of available data is insufficient to fully describe the system.

Second ingredient: the observations

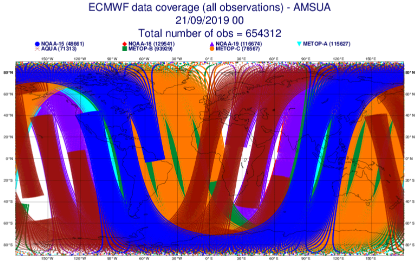
- In the geosciences: The observation space dimension is huge (up to 10^7 for operational systems, up to 10^6 for research systems). A big data problem.

- The Earth observations gather measurements of many sources: conventional and space-borne.





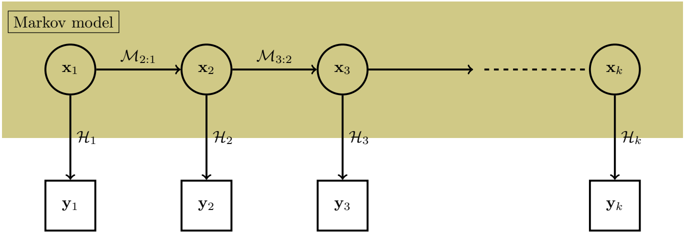
Conventional observations coverage used at ECMWF



AMSUA observations used at ECMWF

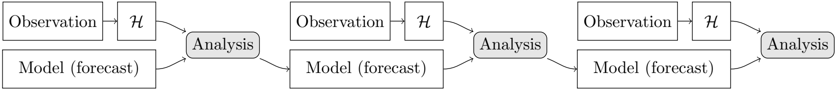
Hidden Markov model

- Considering the states and observations as random variables, the dynamical model, together with the observation model, define a Hidden Markov model:



- This is an inverse problem: Estimate the state **x** given the observation **y**.

- Data assimilation for forecasting chaotic geofluids: sequential schemes



Bayesian inference

- When making inference we have to decide how much we trust the uncertain information. We need to quantify the uncertainty.

- Given the random nature of the problem,

  uncertainty quantification is achieved using probabilities.

- The Bayesian approach offers a natural mathematical framework to understand and formalise this problem.

- In particular, the goal of Bayesian inference is to estimate the uncertainty in $\mathbf{x}$ given $\mathbf{y}$, i.e compute the conditional probability density function (pdf) $p(\mathbf{x}|\mathbf{y})$.

Bayesian inference

- Bayes/Laplace's rule:

$$p(\mathbf{x}|\mathbf{y}) = \frac{p(\mathbf{y}|\mathbf{x})p(\mathbf{x})}{p(\mathbf{y})}$$

with $p(\mathbf{y}|\mathbf{x})$ the likelihood of the observations, $p(\mathbf{x})$ the prior/background on the system's state, and $p(\mathbf{y})$ the evidence. The evidence is a normalisation factor that does not depend on $\mathbf{x}$:

$$p(\mathbf{y}) = \int \mathrm{d}\mathbf{x}\, p(\mathbf{y}|\mathbf{x})p(\mathbf{x})\,.$$

- This is a probabilistic approach. It quantifies the uncertainty/the information. It does not provide a deterministic estimator. This would require to make a choice on top of Bayes'rule.

- The Bayesian approach is very satisfactorily [Jaynes 2003]. Most DA methods can be derived or comply with Bayes'rule.

Sequential Bayesian estimation

▶ Recall our HMM given by the dynamical model and observation model:

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

▶ The model and the observational errors, $\{\boldsymbol{\eta}_k\}_{k=1,\ldots,K}$, $\{\boldsymbol{\epsilon}_k\}_{k=0,\ldots,K}$ 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 $[t_0, \cdots, t_K]$ as $\mathbf{x}_{K:0} = \{\mathbf{x}_K, \mathbf{x}_{K-1}, \cdots, \mathbf{x}_0\}$ and $\mathbf{y}_{K:0} = \{\mathbf{y}_K, \mathbf{y}_{K-1}, \cdots, \mathbf{y}_0\}$ respectively.

We wish to estimate the posterior $p(\mathbf{x}_{K:0}|\mathbf{y}_{K:0})$ for increasing K. Using Bayes'rule:

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

Sequential Bayesian estimation

- Since the observational errors are assumed to be uncorrelated in time we have $p(\mathbf{y}_k|\mathbf{x}_{K:0}) = p(\mathbf{y}_k|\mathbf{x}_k)$ and we can split the global likelihood:

$$p(\mathbf{y}_{K:0}|\mathbf{x}_{K:0}) = \prod_{k=0}^{K} p(\mathbf{y}_k|\mathbf{x}_k) = \prod_{k=0}^{K} p_{\boldsymbol{\epsilon}}\left(\mathbf{y}_k - \mathcal{H}_k(\mathbf{x}_k)\right).$$

- Also, in virtue of the Markov property we have $p(\mathbf{x}_{k+1}|\mathbf{x}_{k:0}) = p(\mathbf{x}_{k+1}|\mathbf{x}_k)$ (prediction at t_{k+1} only depends on the state at t_k), and we can split the global prior as

$$p(\mathbf{x}_{K:0}) = p(\mathbf{x}_0)\prod_{k=1}^{K} p(\mathbf{x}_k|\mathbf{x}_{k-1}) = p(\mathbf{x}_1)\prod_{k=0}^{K} p_{\boldsymbol{\eta}}\left(\mathbf{x}_k - \mathcal{M}_{k:k-1}(\mathbf{x}_{k-1})\right).$$

Sequential Bayesian estimation

- By combining these equations using Bayes'rule we get the posterior distribution

$$\begin{aligned} p(\mathbf{x}_{K:0}|\mathbf{y}_{K:0}) &\propto p(\mathbf{x}_0)p(\mathbf{y}_0|\mathbf{x}_0)\prod_{k=1}^{K} p(\mathbf{y}_k|\mathbf{x}_k)p(\mathbf{x}_k|\mathbf{x}_{k-1}) \\ &\propto p(\mathbf{x}_0)p_{\boldsymbol{\epsilon}}\left(\mathbf{y}_0 - \mathcal{H}_0(\mathbf{x}_0)\right)\prod_{k=1}^{K} p_{\boldsymbol{\epsilon}}\left(\mathbf{y}_k - \mathcal{H}_k(\mathbf{x}_k)\right)p_{\boldsymbol{\eta}}\left(\mathbf{x}_k - \mathcal{M}_{k:k-1}(\mathbf{x}_{k-1})\right). \end{aligned}$$

- This equation is of central importance: it states that a new update can be obtained as soon as new observations are available.

- Sequential inference can be obtained by recursively estimating $p(\mathbf{y}_k|\mathbf{x}_k)p(\mathbf{x}_k|\mathbf{x}_{k-1})$.

- The Bayesian formalism has all the qualities we wish for except that it does not lend to a closed form, analytically tractable solution.

Sequential Bayesian estimation

- Thanks to the main result on the HMM:

$$p(\mathbf{x}_{K:0}|\mathbf{y}_{K:0}) \propto p(\mathbf{x}_0)p(\mathbf{y}_0|\mathbf{x}_0)\prod_{k=1}^{K} p(\mathbf{y}_k|\mathbf{x}_k)p(\mathbf{x}_k|\mathbf{x}_{k-1})$$

we can define the following sequential algorithm to iteratively compute it:

$$p(\mathbf{x}_{k:0}|\mathbf{y}_{k:0}) \propto p(\mathbf{y}_k|\mathbf{x}_k)p(\mathbf{x}_k|\mathbf{x}_{k-1})p(\mathbf{x}_{k-1:0}|\mathbf{y}_{k-1:0}). \tag{1}$$

- An analysis step, in which the conditional pdf $p(\mathbf{x}_k|\mathbf{y}_{k:0})$ is updated using the latest

observation vector, $\mathbf{y}_k$,

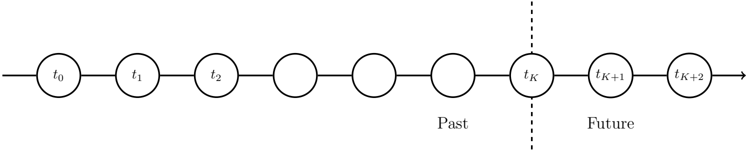
$$p(\mathbf{x}_k|\mathbf{y}_{k:0}) \propto p_{\boldsymbol{\eta}}\left(\mathbf{y}_k - \mathcal{H}_k(\mathbf{x}_k)\right)p(\mathbf{x}_k|\mathbf{y}_{k-1:0}),$$

- which alternates with a forecast step that propagates this pdf, using the Chapman-Kolmogorov equation, forward in time until the new observation batch:

$$p(\mathbf{x}_{k+1}|\mathbf{y}_{k:0}) = \int \mathrm{d}\mathbf{x}\, p_{\boldsymbol{\eta}}\left(\mathbf{x}_k - \mathcal{M}_{k:k-1}(\mathbf{x}_{k-1})\right)p(\mathbf{x}_k|\mathbf{y}_{k:0})$$

to get $p(\mathbf{x}_{k+1}|\mathbf{y}_{k:0})$.

Main goals of data assimilation

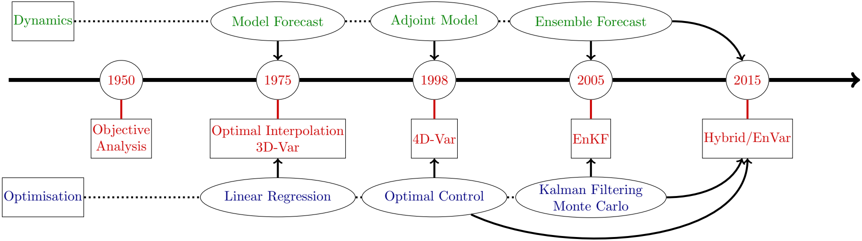


- Recall $\mathbf{x}_{K:0} = \{\mathbf{x}_0, \mathbf{x}_1, \ldots, \mathbf{x}_K\}$, $\mathbf{y}_{K:0} = \{\mathbf{y}_0, \mathbf{y}_1, \ldots, \mathbf{y}_K\}$:
  - Prediction: Estimate $\mathbf{x}_k$ for $k > K$, knowing $\mathbf{y}_{K:0}$,
  - Filtering: Estimate $\mathbf{x}_K$, knowing $\mathbf{y}_{K:0}$,
  - Smoothing: Estimate $\mathbf{x}_{K:0}$, knowing $\mathbf{y}_{K:0}$.

- Less formal names:
  - nowcasting and forecasting,
  - reanalysis,
  - parameter estimation.

Mathematical methods in DA

- Introduction of mathematical methods in operational numerical weather prediction:



- Using increasingly complex mathematical methods and increasingly resolved high-dimensional models.

Outline

1. Data assimilation: principles
   - Introduction
   - Bayesian framework
   - Goals and practical tools of data assimilation

2. Focus on a key elementary derivation

3. Main techniques
   - 3D-Var and optimal interpolation
   - The Kalman filter
   - 4D-Var
   - Particle filters

4. References

Gaussian approximation

▶ A key to obtain a (approximate) solution is to truncate the errors to second-order moments ∼ the Gaussian approximation. Most of DA methods are fully or partially based on this assumption.

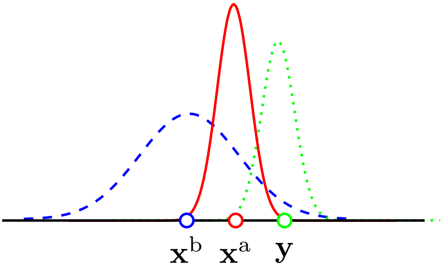
▶ The elementary building block of DA schemes is the statistical BLUE (Best Linear Unbiased Estimator) analysis. Time is considered fixed. **H** is assumed linear.

$$\mathbf{y} = \mathbf{H}\mathbf{x} + \boldsymbol{\epsilon}^{\mathrm{o}}, \qquad \mathbf{x}^{\mathrm{b}} = \mathbf{x} + \boldsymbol{\epsilon}^{\mathrm{b}},$$

where $\boldsymbol{\epsilon}^{\mathrm{o}} \sim \mathcal{N}(\mathbf{0}, \mathbf{R})$, and $\boldsymbol{\epsilon}^{\mathrm{b}} \sim \mathcal{N}(\mathbf{0}, \mathbf{B})$.

▶ Solution:

$$\begin{cases} \mathbf{x}^{\mathrm{a}} & = & \mathbf{x}^{\mathrm{b}} + \mathbf{K}\left(\mathbf{y} - \mathbf{H}\mathbf{x}^{\mathrm{b}}\right) \\ \mathbf{K} & = & \mathbf{B}\mathbf{H}^{\top}\left(\mathbf{R} + \mathbf{H}\mathbf{B}\mathbf{H}^{\top}\right)^{-1} \\ \mathbf{P}^{\mathrm{a}} & = & \left(\mathbf{I} - \mathbf{K}\mathbf{H}\right)\mathbf{B}. \end{cases}$$



Error statistics – Assumptions and definitions

- $\mathbf{x}^\mathrm{t}$ is defined as the true unknown state.

- Observation error statistics:

$$\boldsymbol{\epsilon}^\mathrm{o} = \mathbf{y} - \mathbf{H}\mathbf{x}^\mathrm{t} \qquad \text{with} \qquad \mathbb{E}[\boldsymbol{\epsilon}^\mathrm{o}] = \mathbf{0}, \qquad \mathbb{E}\left[\boldsymbol{\epsilon}^\mathrm{o}{\boldsymbol{\epsilon}^\mathrm{o}}^\top\right] = \mathbf{R},$$

which is in particular satisfied if $\boldsymbol{\epsilon}^\mathrm{o} \sim \mathcal{N}(\mathbf{0}, \mathbf{R})$.

- Background error statistics:

$$\boldsymbol{\epsilon}^\mathrm{b} = \mathbf{x}^\mathrm{b} - \mathbf{x}^\mathrm{t} \qquad \text{with} \qquad \mathbb{E}[\boldsymbol{\epsilon}^\mathrm{b}] = \mathbf{0}, \qquad \mathbb{E}\left[\boldsymbol{\epsilon}^\mathrm{b}{\boldsymbol{\epsilon}^\mathrm{b}}^\top\right] = \mathbf{B}, \qquad \mathbb{E}\left[\boldsymbol{\epsilon}^\mathrm{b}{\boldsymbol{\epsilon}^\mathrm{o}}^\top\right] = \mathbf{0}.$$

- Analysis error statistics:

$$\boldsymbol{\epsilon}^\mathrm{a} = \mathbf{x}^\mathrm{a} - \mathbf{x}^\mathrm{t} \qquad \text{with} \qquad \mathbb{E}[\boldsymbol{\epsilon}^\mathrm{a}] = \mathbf{0}, \qquad \mathbb{E}\left[\boldsymbol{\epsilon}^\mathrm{a}{\boldsymbol{\epsilon}^\mathrm{a}}^\top\right] = \mathbf{P}^\mathrm{a}.$$

Linear unbiased Ansatz for the estimate

- General Ansatz, linear in the observation and the first guess:

$$\mathbf{x}^{\mathrm{a}} = \mathbf{L}\mathbf{x}^{\mathrm{b}} + \mathbf{K}\mathbf{y}.$$

- Writing it in terms of errors:

$$\begin{aligned}\mathbf{x}^{\mathrm{a}} - \mathbf{x}^{\mathrm{t}} &= \mathbf{L}\left(\mathbf{x}^{\mathrm{b}} - \mathbf{x}^{\mathrm{t}} + \mathbf{x}^{\mathrm{t}}\right) + \mathbf{K}\left(\mathbf{H}\mathbf{x}^{\mathrm{t}} + \boldsymbol{\epsilon}^{\mathrm{o}}\right) - \mathbf{x}^{\mathrm{t}},\\ \boldsymbol{\epsilon}^{\mathrm{a}} &= \mathbf{L}\boldsymbol{\epsilon}^{\mathrm{b}} + \mathbf{K}\boldsymbol{\epsilon}^{\mathrm{o}} + \left(\mathbf{L} + \mathbf{K}\mathbf{H} - \mathbf{I}\right)\mathbf{x}^{\mathrm{t}}.\end{aligned}$$

Then $\mathbb{E}[\boldsymbol{\epsilon}^{\mathrm{o}}] = \mathbf{0}$ and $\mathbb{E}[\boldsymbol{\epsilon}^{\mathrm{b}}] = \mathbf{0}$ imply $\mathbb{E}[\boldsymbol{\epsilon}^{\mathrm{a}}] = (\mathbf{L} + \mathbf{K}\mathbf{H} - \mathbf{I})\,\mathbb{E}[\mathbf{x}^{\mathrm{t}}]$.
Hence, we wish to impose

$$\mathbf{L} = \mathbf{I} - \mathbf{K}\mathbf{H}.$$

- As a result, we obtain a linear and unbiased Ansatz:

$$\begin{aligned}\mathbf{x}^{\mathrm{a}} &= (\mathbf{I} - \mathbf{K}\mathbf{H})\mathbf{x}^{\mathrm{b}} + \mathbf{K}\mathbf{y},\\ \mathbf{x}^{\mathrm{a}} &= \mathbf{x}^{\mathrm{b}} + \mathbf{K}\underbrace{\left(\mathbf{y} - \mathbf{H}\mathbf{x}^{\mathrm{b}}\right)}_{\text{innovation}}.\end{aligned}$$

Best linear unbiased estimator

- Posterior error:

$$\boldsymbol{\epsilon}^{\mathrm{a}} = \boldsymbol{\epsilon}^{\mathrm{b}} + \mathbf{K}(\boldsymbol{\epsilon}^{\mathrm{o}} - \mathbf{H}\boldsymbol{\epsilon}^{\mathrm{b}}),$$

so that

$$\begin{aligned}
\mathbf{P}^{\mathrm{a}} &= \mathbb{E}\left[(\boldsymbol{\epsilon}^{\mathrm{a}})(\boldsymbol{\epsilon}^{\mathrm{a}})^\top\right] = \mathbb{E}\left[\left(\boldsymbol{\epsilon}^{\mathrm{b}} + \mathbf{K}(\boldsymbol{\epsilon}^{\mathrm{o}} - \mathbf{H}\boldsymbol{\epsilon}^{\mathrm{b}})\right)\left(\boldsymbol{\epsilon}^{\mathrm{b}} + \mathbf{K}(\boldsymbol{\epsilon}^{\mathrm{o}} - \mathbf{H}\boldsymbol{\epsilon}^{\mathrm{b}})\right)^\top\right] \\
&= \mathbb{E}\left[\left(\mathbf{L}\boldsymbol{\epsilon}^{\mathrm{b}} + \mathbf{K}\boldsymbol{\epsilon}^{\mathrm{o}}\right)\left(\mathbf{L}\boldsymbol{\epsilon}^{\mathrm{b}} + \mathbf{K}\boldsymbol{\epsilon}^{\mathrm{o}}\right)^\top\right] = \mathbb{E}\left[\mathbf{L}\boldsymbol{\epsilon}^{\mathrm{b}}(\boldsymbol{\epsilon}^{\mathrm{b}})^\top\mathbf{L}^\top)\right] + \mathbb{E}\left[\mathbf{K}\boldsymbol{\epsilon}^{\mathrm{o}}(\boldsymbol{\epsilon}^{\mathrm{o}})^\top\mathbf{K}^\top\right] \\
&= \mathbf{L}\mathbf{B}\mathbf{L}^\top + \mathbf{K}\mathbf{R}\mathbf{K}^\top,
\end{aligned}$$

In summary:

$$\mathbf{P}^{\mathrm{a}} = (\mathbf{I} - \mathbf{K}\mathbf{H})\mathbf{B}(\mathbf{I} - \mathbf{K}\mathbf{H})^\top + \mathbf{K}\mathbf{R}\mathbf{K}^\top.$$

- We look for a metric as a global measure of the error. For instance $\mathrm{Tr}(\mathbf{P}^{\mathrm{a}})$. Let us find the optimal $\mathbf{K}$ that minimises this metric.

Best linear unbiased estimator

- Variation of the metric with respect to a variation of $\mathbf{K}$, i.e. $\delta\mathbf{K}$:

$$\begin{aligned}
\delta(\mathrm{Tr}(\mathbf{P}^{\mathrm{a}})) &= \mathrm{Tr}\left((-\delta\mathbf{K}\mathbf{H})\mathbf{B}\mathbf{L}^\top + \mathbf{L}\mathbf{B}(-\delta\mathbf{K}\mathbf{H})^\top + \delta\mathbf{K}\mathbf{R}\mathbf{K}^\top + \mathbf{K}\mathbf{R}\delta\mathbf{K}^\top\right) \\
&= \mathrm{Tr}\left((-\mathbf{L}\mathbf{B}^\top\mathbf{H}^\top - \mathbf{L}\mathbf{B}\mathbf{H}^\top + \mathbf{K}\mathbf{R}^\top + \mathbf{K}\mathbf{R})(\delta\mathbf{K})^\top\right) \\
&= 2\mathrm{Tr}\left((-\mathbf{L}\mathbf{B}\mathbf{H}^\top + \mathbf{K}\mathbf{R})(\delta\mathbf{K})^\top\right).
\end{aligned}$$

- At optimality, one infers that $-(\mathbf{I}-\mathbf{K}^\star\mathbf{H})\mathbf{B}\mathbf{H}^\top + \mathbf{K}^\star\mathbf{R} = \mathbf{0}$, from which we obtain

$$\mathbf{K}^\star = \mathbf{B}\mathbf{H}^\top(\mathbf{R}+\mathbf{H}\mathbf{B}\mathbf{H}^\top)^{-1},$$

from which we get the BLUE solution:

$$\begin{cases}
\mathbf{x}^{\mathrm{a}} = \mathbf{x}^{\mathrm{b}} + \mathbf{K}\left(\mathbf{y} - \mathbf{H}\mathbf{x}^{\mathrm{b}}\right) \\
\mathbf{K} = \mathbf{B}\mathbf{H}^\top\left(\mathbf{R}+\mathbf{H}\mathbf{B}\mathbf{H}^\top\right)^{-1} \\
\mathbf{P}^{\mathrm{a}} = (\mathbf{I}-\mathbf{K}\mathbf{H})\mathbf{B}.
\end{cases}$$

Outline

1. Data assimilation: principles
   - Introduction
   - Bayesian framework
   - Goals and practical tools of data assimilation

2. Focus on a key elementary derivation

3. Main techniques
   - 3D-Var and optimal interpolation
   - The Kalman filter
   - 4D-Var
   - Particle filters

4. References

3D-Var and BLUE in the linear case: derivation

- 3D-Var cost function:

$$J(\mathbf{x}) = \frac{1}{2}\|\mathbf{x}-\mathbf{x}^{\mathbf{b}}\|^2_{\mathbf{B}^{-1}} + \frac{1}{2}\|\mathbf{y}-\mathbf{H}\mathbf{x}\|^2_{\mathbf{R}^{-1}}, \qquad \text{with} \quad \|\mathbf{x}\|^2_{\mathbf{A}} = \mathbf{x}^\top\mathbf{A}\mathbf{x}.$$

- Let us minimise J and compute the variation of $J(\mathbf{x})$ with respect to a variation of $\mathbf{x}$:

$$\begin{aligned}
\delta J(\mathbf{x}) &= \frac{1}{2}(\delta\mathbf{x})^\top\mathbf{B}^{-1}\left(\mathbf{x}-\mathbf{x}^{\mathrm{b}}\right) + \frac{1}{2}\left(\mathbf{x}-\mathbf{x}^{\mathrm{b}}\right)^\top\mathbf{B}^{-1}\delta\mathbf{x} \\
&\quad + \frac{1}{2}(-\mathbf{H}\delta\mathbf{x})^\top\mathbf{R}^{-1}(\mathbf{y}-\mathbf{H}\mathbf{x}) + \frac{1}{2}\left(\mathbf{x}^{\mathrm{b}}-\mathbf{H}\mathbf{x}\right)\mathbf{R}^{-1}(-\mathbf{H}\delta\mathbf{x}) \\
&= (\delta\mathbf{x})^\top\mathbf{B}^{-1}\left(\mathbf{x}-\mathbf{x}^{\mathrm{b}}\right) - (\delta\mathbf{x})^\top\mathbf{H}^\top\mathbf{R}^{-1}(\mathbf{y}-\mathbf{H}\mathbf{x}) \\
&= (\delta\mathbf{x})^\top\nabla J.
\end{aligned}$$

- The extremum condition is $\nabla J = \mathbf{B}^{-1}(\mathbf{x}^\star-\mathbf{x}^{\mathrm{b}}) - \mathbf{H}^\top\mathbf{R}^{-1}(\mathbf{y}-\mathbf{H}\mathbf{x}^\star) = \mathbf{0}$, which yields:

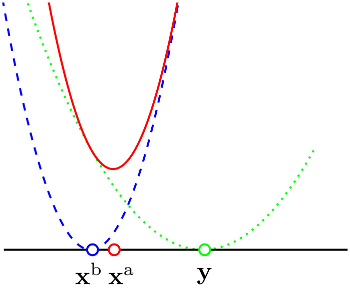
$$\mathbf{x}^\star = \mathbf{x}^{\mathrm{b}} + \underbrace{(\mathbf{B}^{-1}+\mathbf{H}^\top\mathbf{R}^{-1}\mathbf{H})^{-1}\mathbf{H}^\top\mathbf{R}^{-1}}_{\mathbf{K}^\star}(\mathbf{y}-\mathbf{H}\mathbf{x}^{\mathrm{b}}).$$

Thanks to the Sherman-Morrison-Woodbury identity,

$$\mathbf{K}^\star = (\mathbf{B}^{-1}+\mathbf{H}^\top\mathbf{R}^{-1}\mathbf{H})^{-1}\mathbf{H}^\top\mathbf{R}^{-1} = \mathbf{B}\mathbf{H}^\top\left(\mathbf{R}+\mathbf{H}\mathbf{B}\mathbf{H}^\top\right)^{-1}.$$

$\longrightarrow$ $\mathbf{x}^\star$ coincides with the BLUE optimal analysis $\mathbf{x}^{\mathrm{a}}$.

3D-Var and optimal interpolation



- Variational formulation of the same problem

$$J(\mathbf{x}) = \frac{1}{2}\|\mathbf{x}-\mathbf{x}^{\mathbf{b}}\|^2_{\mathbf{B}^{-1}} + \frac{1}{2}\|\mathbf{y}-\mathbf{H}\mathbf{x}\|^2_{\mathbf{R}^{-1}},$$

which is equivalent to BLUE.

- Probabilistic/Bayesian interpretation:

$$p(\mathbf{x}|\mathbf{y}) \propto e^{-J(\mathbf{x})}$$

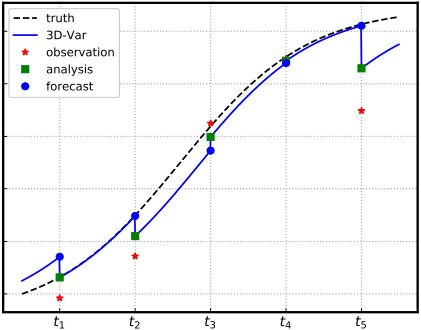
- Capable of handling a nonlinear observation operator using standard nonlinear optimisation methods:

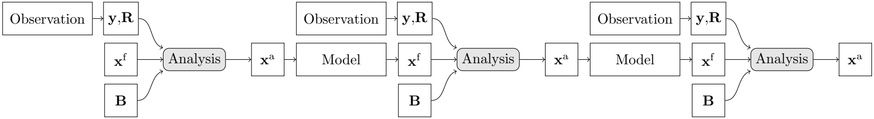
$$J(\mathbf{x}) = \frac{1}{2}\|\mathbf{x}-\mathbf{x}^{\mathrm{b}}\|^2_{\mathbf{B}^{-1}} + \frac{1}{2}\|\mathbf{y}-\mathcal{H}(\mathbf{x})\|^2_{\mathbf{R}^{-1}}.$$

Chaining the analyses in time

- Chaining the BLUE/3D-Var cycles:
  1. Analysis with a forecast at t_k: $\mathbf{x}^{\mathrm{f}}_k$ and with static information $\mathbf{B}$: $\mathbf{x}^{\mathrm{a}}_k$,
  2. Forecast to t_{k+1}: $\mathbf{x}^{\mathrm{f}}_{k+1} = \mathcal{M}_{k+1:k}(\mathbf{x}^{\mathrm{a}}_k)$.

- Also known as optimal interpolation (if the analysis step is BLUE).
- Relatively cheap. Used in oceanography, atmospheric chemistry. Requires a smart construction of $\mathbf{B}$.
- But the information about the errors is not propagated in time…




Observation → $\mathbf{y},\mathbf{R}$; $\mathbf{x}^{\mathrm{f}}$ → Analysis ; $\mathbf{B}$ → Analysis → $\mathbf{x}^{\mathrm{a}}$ → Model → $\mathbf{x}^{\mathrm{f}}$

Observation → $\mathbf{y},\mathbf{R}$ → Analysis ; $\mathbf{B}$ → Analysis → $\mathbf{x}^{\mathrm{a}}$ → Model → $\mathbf{x}^{\mathrm{f}}$

Observation → $\mathbf{y},\mathbf{R}$ → Analysis ; $\mathbf{B}$ → Analysis → $\mathbf{x}^{\mathrm{a}}$

The Kalman filter

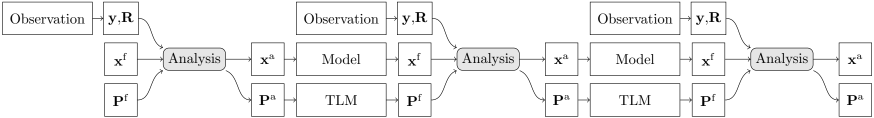
- Similar to optimal interpolation. But, now, we want to replace the static $\mathbf{B}$ with a dynamic $\mathbf{P}^{\mathrm{f}}$ which needs updating and propagating.

- Analysis step:

$$\mathbf{x}_k^{\mathrm{a}} = \mathbf{x}_k^{\mathrm{f}} + \mathbf{K}_k \left(\mathbf{y}_k - \mathbf{H}_k \mathbf{x}_k^{\mathrm{f}} \right),$$
$$\mathbf{K}_k = \mathbf{P}_k^{\mathrm{f}} \mathbf{H}_k^\top \left(\mathbf{R}_k + \mathbf{H}_k \mathbf{P}^{\mathrm{f}} \mathbf{H}_k^\top \right)^{-1},$$
$$\mathbf{P}_k^{\mathrm{a}} = \left(\mathbf{I} - \mathbf{K}_k \mathbf{H}_k \right) \mathbf{P}_k^{\mathrm{f}}.$$

- Forecast step:

$$\mathbf{x}_{k+1}^{\mathrm{f}} = \mathbf{M}_{k+1:k} \mathbf{x}_k^{\mathrm{a}},$$
$$\mathbf{P}_{k+1}^{\mathrm{f}} = \mathbf{M}_{k+1:k} \mathbf{P}_k^{\mathrm{a}} \mathbf{M}_{k+1:k}^\top + \mathbf{Q}_{k+1}.$$


Observation
$\mathbf{y},\mathbf{R}$
$\mathbf{x}^{\mathrm{f}}$
$\mathbf{P}^{\mathrm{f}}$
Analysis
$\mathbf{x}^{\mathrm{a}}$
$\mathbf{P}^{\mathrm{a}}$
Model
TLM
Observation
$\mathbf{y},\mathbf{R}$
$\mathbf{x}^{\mathrm{f}}$
$\mathbf{P}^{\mathrm{f}}$
Analysis
$\mathbf{x}^{\mathrm{a}}$
$\mathbf{P}^{\mathrm{a}}$
Model
TLM
Observation
$\mathbf{y},\mathbf{R}$
$\mathbf{x}^{\mathrm{f}}$
$\mathbf{P}^{\mathrm{f}}$
Analysis
$\mathbf{x}^{\mathrm{a}}$
$\mathbf{P}^{\mathrm{a}}$

The extended Kalman filter

▶ Optimal if the model and observation operators are linear and if all the initial and observations errors are Gaussian: it gives the exact Gaussian solution of Bayes' rule.

▶ Can be extended to nonlinear models with:

$$\begin{aligned} \mathbf{x}^{\mathrm{f}}_{k+1} &= \mathcal{M}_{k+1:k}(\mathbf{x}^{\mathrm{a}}_k), \\ \mathbf{P}^{\mathrm{f}}_{k+1} &= \mathbf{M}_{k+1:k}\mathbf{P}^{\mathrm{a}}_k\mathbf{M}^{\top}_{k+1:k} + \mathbf{Q}_{k+1}, \end{aligned}$$

where $\mathbf{M}_{k+1:k}$ is the tangent linear model (linearisation at $\mathbf{x}^{\mathrm{a}}_k$) of $\mathcal{M}_{k+1:k}$.

▶ Extremely costly for large geophysical models: storage space (storage of $\mathbf{P}^{\mathrm{f}}$) and computations ($\mathbf{M}_{k+1:k}\mathbf{P}^{\mathrm{f}}_k\mathbf{M}^{\top}_{k+1:k}$ requires $2N_x$ integrations of the model).

▶ Solutions: The reduced-rank / ensemble Kalman filters. Wait for lecture 2!

The extended Kalman filter: numerical illustration

- Anharmonic oscillator:

$$\frac{\mathrm{d}^2 x}{\mathrm{d}\, t^2} - \Omega^2 x + \Lambda^2 x^3 = 0,$$

whose numerical implementation is

$$x_0 = 0\,, \quad x_1 = 1 \quad \text{and for } 1 \leqslant k \leqslant N: \quad x_{k+1} - 2x_k + x_{k-1} = \omega^2 x_k - \lambda^2 x_k^3.$$

$\longrightarrow$ Equations for a material dot in a double well potential $V(x) = -\frac{1}{2}\Omega^2 x^2 + \frac{1}{4}\Lambda^2 x^4$.

- Markovian dynamics with an augmented state vector:

$$\mathbf{u}_k = \begin{bmatrix} x_k \\ x_{k-1} \end{bmatrix},$$

with the augmented dynamics

$$\mathcal{M}_{k+1:k} = \begin{bmatrix} 2 + \omega^2 - \lambda^2 x_k^2 & -1 \\ 1 & 0 \end{bmatrix},$$

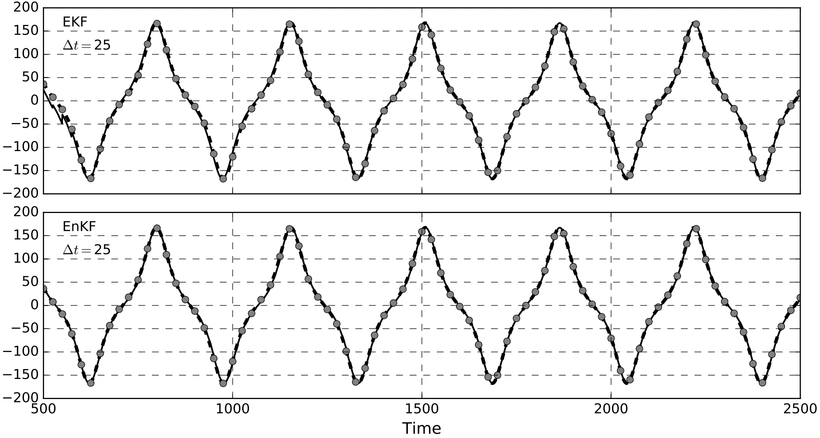
yields

$$\mathbf{u}_{k+1} = \mathcal{M}_{k+1:k}(\mathbf{u}_k).$$

- $\mathbf{H}_k = [1, 0]$. The observation equation is $y_k = \mathbf{H}_k \mathbf{u}_k + \epsilon_k$.

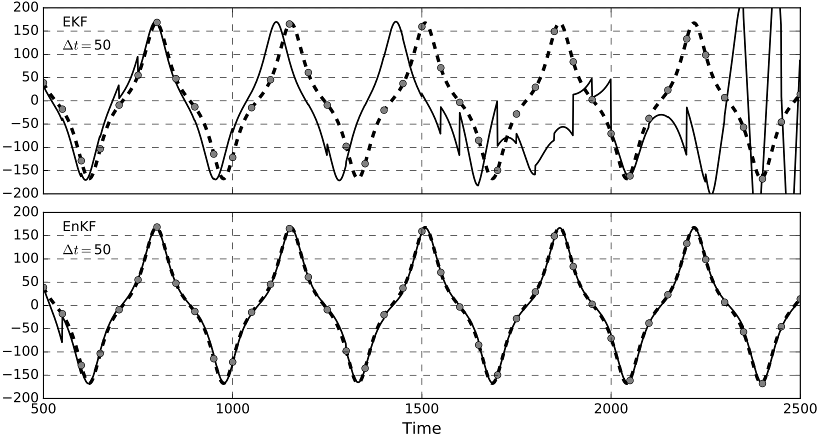
The extended Kalman filter: numerical illustration

- Comparison with the EnKF that does not rely on the tangent linear approximation.



The extended Kalman filter: numerical illustration

- Comparison with the EnKF that does not rely on the tangent linear approximation.



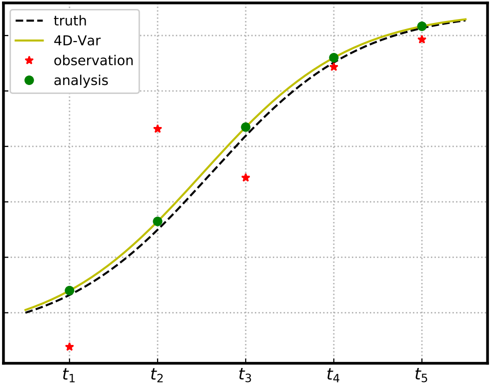
4D-Var

- Strongly constrained 4D-Var, i.e. assuming the model is perfect (no model error)

$$J(\mathbf{x}_0) = \frac{1}{2}\|\mathbf{x}_0 - \mathbf{x}_0^{\mathbf{b}}\|^2_{\mathbf{B}^{-1}} + \frac{1}{2}\sum_{k=0}^{K}\|\mathbf{y}_k - \mathcal{H}_k(\mathbf{x}_k)\|^2_{\mathbf{R}_k^{-1}},$$

under the constraints that $\mathbf{x}_{k+1} = \mathcal{M}_{k+1:k}(\mathbf{x}_k)$ for $k = 0, \ldots, K-1$.

- Fits a model trajectory through the 4D data points.



4D-Var: algorithm

- Lagrangian for 4D-Var:

$$L(\mathbf{x}_{K:0},\boldsymbol{\lambda}_{k:0}) = \frac{1}{2}\|\mathbf{x}_0-\mathbf{x}_0^{\mathbf{b}}\|^2_{\mathbf{B}^{-1}} + \frac{1}{2}\sum_{k=0}^{K}\|\mathbf{y}_k-\mathcal{H}_k(\mathbf{x}_k)\|^2_{\mathbf{R}_k^{-1}} + \sum_{k=1}^{K}\boldsymbol{\lambda}_k^{\top}\left(\mathbf{x}_k-\mathcal{M}_{k:k-1}(\mathbf{x}_{k-1})\right).$$

- Gradient of the Lagrangian with respect to $\mathbf{x}_{K:0}$:

$$\begin{aligned}
\nabla_{\mathbf{x}_0}L(\mathbf{x}_0) &= \mathbf{B}^{-1}\left(\mathbf{x}_0-\mathbf{x}_0^{\mathbf{b}}\right)-\mathbf{H}_0^{\top}\mathbf{R}_0^{-1}\left(\mathbf{y}_0-\mathbf{H}_0(\mathbf{x}_0)\right)-\mathbf{M}_{1:0}^{\top}\boldsymbol{\lambda}_1,\\
\nabla_{\mathbf{x}_k}L(\mathbf{x}_0) &= -\mathbf{H}_k^{\top}\mathbf{R}_k^{-1}\left(\mathbf{y}_k-\mathbf{H}_k(\mathbf{x}_k)\right)-\mathbf{M}_{k+1:k}^{\top}\boldsymbol{\lambda}_{k+1}+\boldsymbol{\lambda}_k,\\
\nabla_{\mathbf{x}_K}L(\mathbf{x}_0) &= -\mathbf{H}_K^{\top}\mathbf{R}_K^{-1}\left(\mathbf{y}_K-\mathbf{H}_K(\mathbf{x}_K)\right)+\boldsymbol{\lambda}_K.
\end{aligned}$$

- Requires the computation of the tangent linear and adjoint of $\mathcal{H}_k$ and $\mathcal{M}_{k+1:k}$.

- No perfect (general purpose) automatic differentiation tool: developing and maintaining the adjoint codes is computationally very costly!

4D-Var: algorithm

- Algorithm: one outer loop
  1. Given the initial condition $\mathbf{x}_0$, compute the trajectory $\mathbf{x}_{K:0}$ with the dynamical model $\mathcal{M}$.
  2. Compute the adjoint trajectory backwards in time:
$$\begin{aligned}\boldsymbol{\lambda}_K &= \mathbf{H}_K^\top \mathbf{R}_K^{-1}\left(\mathbf{y}_K - \mathbf{H}_K(\mathbf{x}_K)\right),\\ \boldsymbol{\lambda}_k &= \mathbf{H}_k^\top \mathbf{R}_k^{-1}\left(\mathbf{y}_k - \mathbf{H}_k(\mathbf{x}_k)\right) - \mathbf{M}_{k+1:k}^\top \boldsymbol{\lambda}_{k+1},\\ \boldsymbol{\lambda}_0 &= \mathbf{H}_0^\top \mathbf{R}_0^{-1}\left(\mathbf{y}_0 - \mathbf{H}_0(\mathbf{x}_0)\right) - \mathbf{M}_{1:0}^\top \boldsymbol{\lambda}_1.\end{aligned}$$
  3. This finally yields:
$$\nabla_{\mathbf{x}_0} J(\mathbf{x}_0) = \mathbf{B}^{-1}\left(\mathbf{x}_0 - \mathbf{x}_0^{\mathbf{b}}\right) - \boldsymbol{\lambda}_0.$$

- Can be used to feed any gradient-based minimisation scheme (Newton, Gauss-Newton, L-BFGS, conjugate-gradient, Levenberg-Marquardt, trust region methods).

4D-Var: algorithm

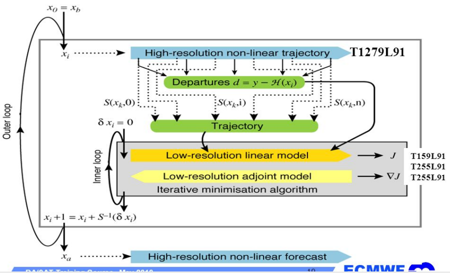
▶ For high-dimensional systems: incremental strategy with outer/inner loops. The inner-loop Lagrangian, which is quadratic in $\delta\mathbf{x}_{K:0}$, is

$$L^{(p)}(\delta\mathbf{x}_{K:0},\boldsymbol{\lambda}_{k:0}) = \frac{1}{2}\|\mathbf{x}_0^{(p)} - \mathbf{x}_0^{\mathrm{b}} + \delta\mathbf{x}_0\|^2_{\mathbf{B}^{-1}} + \frac{1}{2}\sum_{k=0}^{K}\|\mathbf{y}_k - \mathcal{H}_k(\mathbf{x}_k^{(p)}) + \mathbf{H}^{(p)}(\delta\mathbf{x}_k)\|^2_{\mathbf{R}_k^{-1}}$$

$$+\sum_{k=1}^{K}\boldsymbol{\lambda}_k^{\top}\left(\mathbf{x}_{k+1}^{(p)} - \mathcal{M}_{k+1:k}(\mathbf{x}_k^{(p)}) - \mathbf{M}_{k:k-1}^{(p)}(\delta\mathbf{x}_{k-1})\right).$$

It can efficiently be solved using a conjugate-gradient algorithm.

Multi-incremental quadratic 4D-Var at ECMWF


$x_0 = x_b$
x_i
High-resolution non-linear trajectory
T1279L91
Departures $d = y - \mathcal{H}(x_i)$
$S(x_k,0)$
$S(x_k,\mathrm{i})$
$S(x_k,\mathrm{n})$
$\delta x_i = 0$
Trajectory
Outer loop
Inner loop
Low-resolution linear model
J
Low-resolution adjoint model
∇J
Iterative minimisation algorithm
T159L91
T255L91
T255L91
$x_i + 1 = x_i + S^{-1}(\delta x_i)$
x_a
High-resolution non-linear forecast
ECMWF

4D-Var: algorithm

- Let us assume Gaussian model error:

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

- Weakly constrained 4D-Var, i.e. assuming the model is imperfect [Trémolet 2006]

$$J(\mathbf{x}_{K:0}) = \frac{1}{2}\|\mathbf{x}_0 - \mathbf{x}_0^\mathbf{b}\|^2_{\mathbf{B}^{-1}} + \frac{1}{2}\sum_{k=0}^{K}\|\mathbf{y}_k - \mathcal{H}_k(\mathbf{x}_k)\|^2_{\mathbf{R}_k^{-1}} + \frac{1}{2}\sum_{k=1}^{K}\|\mathbf{x}_k - \mathcal{M}_{k:k-1}(\mathbf{x}_{k-1})\|^2_{\mathbf{Q}_k^{-1}}.$$

- Adds much flexibility to trajectory fitting.

- Huge control variables (K times bigger) for a very specific form of model error. . .

Taking the bull by the horns: the particle filter

- The particle filter is the Monte-Carlo solution of the Bayes'equation. This is a sequential Monte Carlo method.

- The most simple algorithm of Monte Carlo type that solves the Bayesian filtering equations is called the bootstrap particle filter [Gordon et al. 1993].

Sampling: Particles $\{\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_M\}$.
Pdf at time t_k: $p_k(\mathbf{x}) \simeq \sum_{i=1}^{M} \omega_i^k \delta(\mathbf{x} - \mathbf{x}_k^i)$.

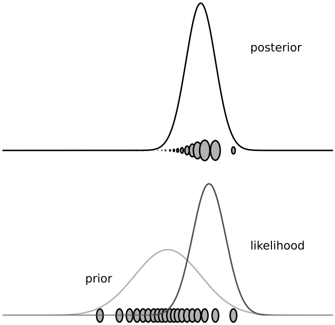
Forecast: Particles propagated by

$$p_{k+1}(\mathbf{x}) \simeq \sum_{i=1}^{M} \omega_k^i \delta(\mathbf{x} - \mathbf{x}_{k+1}^i)$$

with $\mathbf{x}_{k+1}^i = \mathcal{M}_{k+1}(\mathbf{x}_k)$.

Analysis: Weights updated according to

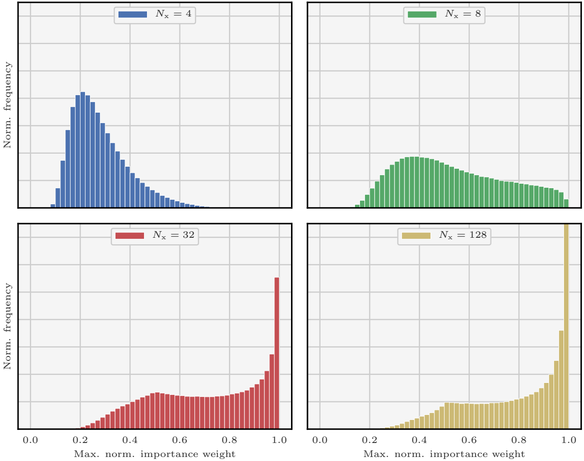
$$\omega_{k+1}^{\mathrm{a},i} \propto \omega_{k+1}^{\mathrm{f},i} p(\mathbf{y}_{k+1}|\mathbf{x}_{k+1}^i) .$$



- Analysis is carried out with only a few multiplications. No matrix inversion!

The particle filter: degeneracy

▶ These normalised statistical weights have a potentially large amplitude of fluctuation. One particle will stand out among the others. Its weight will largely dominate the others ($\omega_i \lesssim 1$). This phenomenon is called degeneracy of the particle filter [Kong et al. 1994].



The particle filter: the curse of dimensionality

▶ Handles very well, very nonlinear low-dimensional systems. But, without modification, very inefficient for high-dimensional models. Avoiding degeneracy requires a great number of particles that scales exponentially with the size of the system [Snyder et al. 2008]. This is a manifestation of the curse of dimensionality.

▶ Are there solutions to circumvent this curse of dimensionality?

- Resampling the particles to reset the weights.
- Introduce diversity by adding jitter to the particles.
- Localisation can be (should be?) used in conjunction with the particle filter [Reich 2013; Poterjoy 2016; Penny and Miyoshi 2016; Farchi and Bocquet 2018].

⟶ Much more on particle filters in Dan Crisan's lectures next week!

References I

[1] M. Asch, M. Bocquet, and M. Nodet. *Data Assimilation: Methods, Algorithms, and Applications*. Fundamentals of Algorithms. SIAM, Philadelphia, 2016, p. 324.

[2] A. Carrassi et al. "Data Assimilation in the Geosciences: An overview on methods, issues, and perspectives". In: *WIREs Climate Change* 9 (2018), e535.

[3] R. Daley. *Atmospheric Data Analysis*. Cambridge University Press, New-York, 1991, p. 472.

[4] G. Evensen. *Data Assimilation: The Ensemble Kalman Filter*. Second. Springer-Verlag Berlin Heildelberg, 2009, p. 307.

[5] A. Farchi and M. Bocquet. "Review article: Comparison of local particle filters and new implementations". In: *Nonlin. Processes Geophys.* 25 (2018), pp. 765–807.

[6] S. J. Fletcher. *Data assimilation for the geosciences: From theory to application*. Elsevier, 2017.

[7] M. Ghil and P. Malanotte-Rizzoli. "Data assimilation in meteorological and oceanography". In: *Advanc. in Geophys.* 33 (1991), pp. 141–266.

[8] N. J. Gordon, D. J. Salmond, and A. F. M. Smith. "Novel approach to nonlinear/non-Gaussian Bayesian state estimation". In: *IEE Proc.-F* 140 (1993), pp. 107–113.

[9] T. Janjić et al. "On the representation error in data assimilation". In: *Q. J. R. Meteorol. Soc.* 144 (2018), pp. 1257–1278.

[10] E. T. Jaynes. *Probability theory: The logic of science*. Cambridge university press, 2003, p. 753.

[11] E. Kalnay. *Atmospheric Modeling, Data Assimilation and Predictability*. Cambridge University Press, Cambridge, 2002, p. 357.

[12] A. Kong, J. S. Liu, and W. H. Wong. "Sequential imputations and Bayesian missing data problems". In: *Journal of the American statistical association* 89 (1994), pp. 278–288.

[13] S. G. Penny and T. Miyoshi. "A local particle filter for high dimensional geophysical systems". In: *Nonlin. Processes Geophys.* 23 (2016), pp. 391–405.

[14] J. Poterjoy. "A localized particle filter for high-dimensional nonlinear systems". In: *Mon. Wea. Rev.* 144 (2016), pp. 59–76.

[15] S. Reich. "A nonparametric ensemble transform method for Bayesian inference.". In: *SIAM J. Sci. Comput.* 35 (2013), A2013–A2014.

[16] S. Reich and C. Cotter. *Probabilistic Forecasting and Bayesian Data Assimilation*. Cambridge University Press, 2015, p. 306.

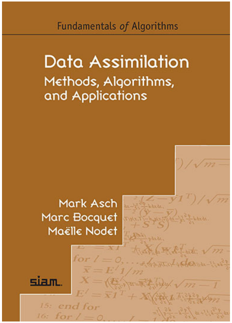
[17] C. Snyder et al. "Obstacles to High-Dimensional Particle Filtering". In: *Mon. Wea. Rev.* 136 (2008), pp. 4629–4640.

[18] Y. Trémolet. "Accounting for an imperfect model in 4D-Var". In: *Q. J. R. Meteorol. Soc.* 132 (2006), pp. 2483–2504.

Looking for a textbook in data assimilation?

Thank you for your attention!

- Part I: A gentle introduction to DA.
- Part II: More advanced topics including EnKF and EnVar.
- Part III: Applications of DA including emerging ones such as: glaciology, biology, geomagnetism, medicine, imaging and acoustics, economics and finance, traffic control, etc.


Fundamentals *of* Algorithms
Data Assimilation
Methods, Algorithms, and Applications
Mark Asch
Marc Bocquet
Maëlle Nodet
siam