Tutorial on data assimilation

Marc Bocquet

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

(marc.bocquet@enpc.fr)
Outline

1. Data assimilation: principles
   - Definition
   - Mathematical framework

2. Main techniques
   - 3D-Var and optimal interpolation
   - The Kalman filter
   - The ensemble Kalman filter
   - 4D-Var

3. Advanced techniques
   - Ensemble variational methods
   - Particle filters

4. Uncertainty quantification of the best estimate
Data assimilation: definition

- **Data assimilation** is the set of techniques that allow to **optimally combine** observations of a physical system with numerical models and **prior information** of that system, so as to get an estimate of the state of the system.

- In the geosciences: Numerical models are often computationally costly. They are often dynamical.

- In the geosciences: The state space and observations space are huge (up to $10^9/10^7$ for operational systems, up to $10^7/10^5$ for research systems). A big data problem with costly dynamical models.

- What for?: estimate initial state of chaotic systems for forecasting, re-analysis, estimate parameters ($\sim$ inverse modelling).

- Example: Data assimilation for prediction.
Data assimilation system

Data assimilation system = observation and evolution models + statistics of the errors. Typically:

\[
\begin{align*}
    x_k &= M_{k:k-1}(x_{k-1}) + \eta_k \\
    y_k &= H_k(x_k) + \varepsilon_k
\end{align*}
\]

with \( \eta_k \sim \mathcal{N}(0, Q_k) \) and \( \varepsilon_k \sim \mathcal{N}(0, R_k) \).

Denoting \( x_{K:1} = x_1, x_2, \ldots, x_K, y_{K:1} = y_1, y_2, \ldots, y_K \):

- **Prediction**: Estimate \( x_k \) for \( k > K \), knowing \( y_{K:1} \);
- **Filtering**: Estimate \( x_K \), knowing \( y_{K:1} \);
- **Smoothing**: Estimate \( x_{K:1} \), knowing \( y_{K:1} \).
The ideal mathematical framework

- Bayes/Laplace approach:

\[ p(x|y) = \frac{p(y|x)p(x)}{p(y)} \]

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

\[ p(y) = \int dx \, p(y|x)p(x). \]

- 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 et al., 2003]. Most DA methods can be derived or comply with Bayes’ rule.

- But it does not lend to a closed form analytically tractable solution.
Gaussian approximation

A key to obtain a (approximate) solution is to truncate the errors to second-order moments $\sim$ 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 (for Best Linear Unbiased Estimator) analysis. Time is considered fixed. $H$ is assumed linear.

$$y = Hx + \varepsilon^o \quad x = x^b + \varepsilon^b$$

where $\varepsilon^o \sim \mathcal{N}(0, R)$, and $\varepsilon^b \sim \mathcal{N}(0, B)$.

Solution:

$$x^a = x^b + K( y - Hx^b)$$

$$K = BH^T (R + HBH^T)^{-1}$$

$$P^a = (I - KH)B.$$
Outline

1 Data assimilation: principles
   • Definition
   • Mathematical framework

2 Main techniques
   • 3D-Var and optimal interpolation
   • The Kalman filter
   • The ensemble Kalman filter
   • 4D-Var

3 Advanced techniques
   • Ensemble variational methods
   • Particle filters

4 Uncertainty quantification of the best estimate
Variational formulation of the same problem

\[ J(x) = \frac{1}{2} \| x - x^b \|^2_B + \frac{1}{2} \| y - Hx \|^2_R \]

where \( \| x \|^2_A = x^T A x \), which is equivalent to BLUE.

Probabilistic/Bayesian interpretation:

\[ p(x|y) \propto e^{-J(x)} \]

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

\[ J(x) = \frac{1}{2} \| x - x^b \|^2_B + \frac{1}{2} \| y - H(x) \|^2_R. \]
Chaining the analyses in time

- Chaining the BLUE/3D-Var cycles:
  1. Analysis with a forecast at $t_k$: $x^f_k$ and with static information $B$: $x^a_k$;
  2. Forecast to $t_{k+1}$: $x^f_{k+1} = M_{k+1:k}(x^a_k)$.

- Also known as optimal interpolation (if the analysis step is BLUE).

- Relatively cheap. Used in oceanography, atmospheric chemistry. Requires a smart construction of $B$.

- But the information about the errors is not propagated in time ...
The Kalman filter

▶ Similar to optimal interpolation. But, now, we want to replace the static $B$ with a dynamic $P^f$ which needs updating and propagating.

▶ Analysis step:

$$x^a_k = x^f_k + K_k \left( y_k - H_k x^f_k \right),$$

$$K_k = P^f_k H^T_k \left( R_k + H_k P^f_k H^T_k \right)^{-1},$$

$$P^a_k = \left( I - K_k H_k \right) P^f_k.$$  

▶ Forecast step:

$$x^f_{k+1} = M_{k+1:k} x^a_k,$$

$$P^f_{k+1} = M_{k+1:k} P^a_k M^T_{k+1:k} + Q_{k+1}.$$
Main techniques

The Kalman filter

- Optimal if the models are linear and if all the initial and observations errors are Gaussian: it gives the perfect Gaussian solution of Bayes’ rule.

- Can be extended to nonlinear models: then

\[
x_{k+1}^f = M_{k+1:k}(x_{k}^a),
\]

\[
P_{k+1}^f = M'_{k+1:k}P_{k}^fM'^T_{k+1:k} + Q_{k+1},
\]

where \( M'_{k+1:k} \) is the tangent linear model.

- Extremely costly for large geophysical models: storage space (storage of \( P^f \)) and computations (\( M'_{k+1:k}P_{k}^fM'^T_{k+1:k} \) requires \( 2n \) integrations of the model).
The ensemble Kalman filter

- The idea [Evensen, 1994; Houtekamer and Mitchell, 1998] is to make the KF work in high dimensions and replace $\mathbf{P}$ ($\mathbf{P}^a$ or $\mathbf{P}^f$) with an ensemble of states $x_1, x_2, \ldots, x_m$. The moments of the error could theoretically be approximated by the sample/empirical moments:

$$x^f = \frac{1}{m} \sum_{i=1}^{m} \bar{x}, \quad \mathbf{P}^f = \frac{1}{m-1} \sum_{i=1}^{m} (x(i) - \bar{x})(x(i) - \bar{x})^T.$$ 

- Analysis step: Similar to the KF but $\mathbf{P}^f$ explicitly or implicitly taken as the sample covariance estimator.

- Forecast step: The ensemble is propagated using the full nonlinear model (not the tangent linear model!)

$$x_{k+1}^{(i),f} = M_{k+1:k} \left( x_{k}^{(i),a} \right).$$
Main techniques

The ensemble Kalman filter

Two main flavors of EnKFs: stochastic and deterministic, but many variants.

The stochastic EnKF is the closest to traditional KF, but adds stochastic perturbations to the observations of each members to properly account for the observation errors [Burgers et al., 1998]:

\[ x_{a(i)} = x_{f(i)} + K \left( y + \varepsilon_{(i)} - Hx_{f(i)} \right). \]

The deterministic EnKF avoids the introduction of the stochastic perturbations by updating the square root of \( P^f = X_f X_f^T \), i.e. \( X_f \). One of the variant (ETKF, [Hunt et al., 2007]) operates the linear algebra in the space of the perturbations:

\[ x^a = x^f + X_f w^a. \]

The analysis in the perturbation space is given by

\[ w^a = \left( I_m + Y_f^T R^{-1} Y_f \right)^{-1} Y_f^T R^{-1} \left( y - Hx^f \right) \]

where \( Y_f = Hx^f \). This updates the mean state via \( x^a = x^f + X_f w^a \). The perturbations around it are updated via

\[ X_a = X_f \left( I_m - Y_f^T (Y_f Y_f^T + R)^{-1} Y_f \right)^{1/2} U, \quad \text{where} \quad U \in O(N) \quad \text{and} \quad U1 = 1. \]
The downside of the EnKF: rank-deficiency

There is a heavy price to pay for replacing the $P^f_{n \times n}$ covariance matrix of the KF with the $X_f_{m \times n}$ anomaly matrix: spurious correlations for distant state components. If $P = X_f X_f^T$ and $B$ is the true error covariance matrix of a Gaussian process:

$$\text{Cov} ([P]_{ii}, [P]_{jj}) = \frac{2}{N-1} [B]_{ij}^2, \quad \text{Cov} ([P]_{ij}, [P]_{ij}) = \frac{1}{N-1} \left( [B]_{ij}^2 + [B]_{ii} [B]_{jj} \right).$$

But, for geophysical systems, we know that most long-range correlations are dampened exponentially. Consequently, the covariances are misestimated (too low variances, too high long-range covariances) and leads to divergence of the EnKF. Practically, this is solved using two fixes: inflation and localisation.

Inflation consists in inflating the covariances by a scalar in the hope to compensate for the underestimation of the error statistics [Pham et al., 1998, Anderson et al., 1999]:

$$x_{(i)} \leftarrow x_{(i)} + \lambda \left( x_{(i)} - \bar{x} \right).$$

Can be avoided in a perfect-model context: finite-size EnKF (EnKF-N) [Bocquet et al., 2011-2018].
Localisation

- Two flavors of localisation: domain localisation and covariance localisation.

- Domain localisation: divide and conquer. The DA analysis is performed in parallel in local domains. The outcomes of these analyses are later sewed together. This is applicable only if the long-range error correlations are negligible.

- Covariance localisation: killing off spurious correlation explicitly: $P_f = \rho \circ (X_fX_f^T)$.

- These strategies have successfully been applied to the EnKF [Hamill et al, 2001; Houtekamer and Mitchell, 2001; Evensen, 2003; Hunt et al., 2007].
Nonlinear chaotic models: the Lorenz-95 low-order model

- It represents a mid-latitude zonal circle of the global atmosphere.
- Set of $n = 40$ ordinary differential equations [Lorenz and Emmanuel 1998]:

$$\frac{dx_i}{dt} = (x_{i+1} - x_{i-2})x_{i-1} - x_i + F,$$

where $F = 8$, and the boundary is cyclic.
- Conservative system except for a forcing term $F$ and a dissipation term $-x_i$.
- Chaotic dynamics, 13 positive and 1 neutral Lyapunov exponents, a doubling time of about 0.42 time units.
Illustration with the Lorenz-95 model

▶ Performance of the EnKF in the absence/presence of inflation/localisation.
What about smoothing?

There are smoothing variants of the Kalman filter [Anderson & Moore, 1979], the Kalman smoother used in the geosciences [Cohn et al., 1994].

And they have been adapted to the EnKF and variants [Evensen & van Leeuwen, 2000], [Evensen, 2009], [Cosme et al., 2012], [Bocquet & Sakov, 2014], etc.

Sometimes called asynchronous data assimilation [Sakov et al., 2010; Sakov & Bocquet, 2018].

With the notable exception of the IEnKS, these smoothers relies on Gaussian assumptions within the DAW.

4D-Var is a more natural method to handle nonlinearity within the DAW.
4D-Var

- Strongly constrained 4D-Var, i.e. assuming the model is perfect

\[
J(x_0) = \frac{1}{2} \|x_0 - x_0^b\|^2_B^{-1} + \frac{1}{2} \sum_{k=1}^{K} \|y_k - H_k(x_k)\|^2_R^{-1} + \frac{1}{2} \sum_{k=1}^{K} \|x_k - M_{k+1:k-1}(x_k-1)\|^2_Q^{-1}
\]

under the constraints that \(x_{k+1} = M_{k+1:k}(x_k)\) for \(k = 0, \ldots, K - 1\).

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

- In high-dimensional spaces, requires \(\nabla_{x_0} J\) for an efficient minimisation. But \(\nabla_{x_0} J\) depends on the adjoint of \(M_{k+1:k}\) and \(H_k\). This can be a very difficult technical task if the model is a huge piece of code for a nonlinear high-dimensional model.

- Weakly constrained 4D-Var, i.e. assuming the model is imperfect

\[
J(x_{K:0}) = \frac{1}{2} \|x_0 - x_0^b\|^2_B^{-1} + \frac{1}{2} \sum_{k=0}^{K} \|y_k - H_k(x_k)\|^2_R^{-1} + \frac{1}{2} \sum_{k=1}^{K} \|x_k - M_{k:k-1}(x_k-1)\|^2_Q^{-1}
\]
Outline

1. Data assimilation: principles
   - Definition
   - Mathematical framework

2. Main techniques
   - 3D-Var and optimal interpolation
   - The Kalman filter
   - The ensemble Kalman filter
   - 4D-Var

3. Advanced techniques
   - Ensemble variational methods
   - Particle filters

4. Uncertainty quantification of the best estimate
Hybridising ensemble and variational methods

- **Hybrid**: Use flow-dependent statistics from an EnKF into 3D-Var [Hamil & Snyder 2000; Wang et al. 2007].

- **4D-LETKF** [Hunt et al., 2004; Fertig et al., 2007]

- **EDA**: ensemble of 4D-Var (ECMWF, Météo-France) [Raynaud et al., 2009; Bonavita et al., 2012; Berre et al., 2015; Jardak & Talagrand 2018]

- **4DEnVar**: Adjoint-less 4D-Var [Liu et al., 2008; Buehner et al. 2010; Zhang and Zhang, 2012; Fairbairn et al. 2014, Desroziers et al. 2014], but ensemble update and nonlinearity still not completely addressed.

- **IEnKS**: has it all [Sakov et al. 2012, Bocquet & Sakov 2012-2016].

As ensemble methods, they all require localisation, which is more difficult to implement in a 4D context [Bocquet, 2016] except if the adjoint is available.

→ For a review on EnVar methods, see Chapter 7 of the new book [Asch et al., 2016].
The iterative ensemble Kalman smoother (IEnKS)

- Reduced scheme in ensemble space, \( x_0 = \bar{x}_0 + X_0 w \), where \( X_0 \) is the ensemble anomaly matrix:
  \[
  \tilde{J}(w) = J(\bar{x}_0 + X_0 w).
  \]

- Analysis IEnKS cost function in ensemble space:
  \[
  \tilde{J}(w) = \frac{1}{2} \sum_{k=1}^{L} \| y_k - H_k \circ M_{k:0} (\bar{x}_0 + X_0 w) \|^2_{\beta_k R_k^{-1}} + \frac{1}{2} (N - 1) \| w \|^2.
  \]

\( \{ \beta_0, \beta_1, \ldots, \beta_L \} \) weight the observations impact within the window.

- As a variational reduced method, one can use Gauss-Newton [Sakov et al., 2012], Levenberg-Marquardt [Bocquet & Sakov, 2012], quasi-Newton, trust region, etc., minimisation schemes.

- Perturbation update: same as the ETKF
  \[
  E_0^* = x_0^* 1^T + \sqrt{N-1} X_0 \left[ \nabla_w^2 \tilde{J} \right]_{\star}^{-1/2} U \text{ where } U \in O(N) \text{ and } U1 = 1.
  \]

→ Cécile Defforge’s talk.
The IEnKS opens up new perspectives on the chaining of DA cycles which was little relevant for either the EnKF or 4D-Var.

$L$: length of the data assimilation window,

$S$: shift of the data assimilation window in between two updates.

Variational analysis in ens. space $\rightarrow$ Posterior ens. generation $\rightarrow$ Ens. forecast
Comparing 4D-Var, the EnKF, the EnKS and the IEnKS.
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 \( \{x_1, x_2, \ldots, x_M\} \).

Pdf at time \( t_k \): \( p_k(x) \approx \sum_{i=1}^{M} \omega_i^k \delta(x - x_i^k) \).

**Forecast:** Particles propagated by

\[
p_{k+1}(x) \approx \sum_{i=1}^{M} \omega_i^k \delta(x - x_i^{k+1})
\]

with \( x_i^{k+1} = M_{k+1}(x_k) \).

**Analysis:** Weights updated according to

\[
\omega_{k+1}^{a,i} \propto \omega_{k+1}^{f,i} p(y_{k+1}|x_i^{k+1})
\]

» Analysis is carried out with only a few multiplications. No matrix inversion!
Taking the bull by the horns: the particle filter

► These normalised statistical weights have a potentially large amplitude of fluctuation. One particle (one trajectory of the model) will stand out among the others. Its weight will largely dominate the others ($\omega_i \lesssim 1$). Then the particle filter becomes very inefficient as an estimating tool since it has lost its variability. This phenomenon is called degeneracy of the particle filter [Kong et al. 1994].

Resampling One way to mitigate this phenomenon is to resample the particles by redrawing a sample with uniform weights from the degenerate distribution. After resampling, all particles have the same weight: $\omega_i^k = 1/M$.

► 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. This is a manifestation of the curse of dimensionality.
Advanced techniques
Particle filters

Application of the particle filter in the geosciences

▶ The applicability of particle filters to high-dimensional models has been investigated in the geosciences [van Leeuwen, 2009; Bocquet, 2010]. The impact of the curse of dimensionality has been quantitatively studied in [Snyder et al., 2008]. It was known [Mackay et al., 2003] that using an importance proposal to guide the particles towards regions of high probability will not change this trend, albeit with a reduced exponential scaling, which was confirmed by [Snyder et al., 2015]: optimal importance sampling particle filter [Doucet et al., 2000; Bocquet, 2010; Snyder, 2011].

▶ Particle smoother over a data assimilation window, alternative and more efficient particle filters can be designed, such as the implicit particle filter [Morzfeld et al., 2012].

▶ Particle filters can nevertheless be useful for high-dimensional models if the significant degrees of nonlinearity are confined to a small subspace of the state space, e.g. Lagrangian data assimilation [Slivinski et al., 2015].

▶ It is possible possible to design nonlinear filters for high-dimensional models such as the equal-weight particle filter [van Leeuwen & Ades, 2010-2017].

▶ Localisation can be (should be?) used in conjunction with the particle filter [Reich et al. 2013; Potterjoy, 2016; Penny & Miyoshi, 2016; Farchi & Bocquet, 2018].

→ Alban Farchi’s talk.

▶ It has been applied in hydrology, nivology, climate, etc [Goosse, Dubinkina et al.].
Outline

1. Data assimilation: principles
   - Definition
   - Mathematical framework

2. Main techniques
   - 3D-Var and optimal interpolation
   - The Kalman filter
   - The ensemble Kalman filter
   - 4D-Var

3. Advanced techniques
   - Ensemble variational methods
   - Particle filters

4. Uncertainty quantification of the best estimate
Case study: Chernobyl and Fukushima accidents

- 30 deaths in the first days of the accident
- 200,000 evacuees
- 30 km exclusion zone
- Mid and long term sanitary impact: thyroid cancer (thousands of cases).
Case study: Chernobyl and Fukushima accidents

Caesium-137 deposition [IRSN database]

Air quality monitoring network
Case study: Chernobyl and Fukushima accidents

- Modelled by PDEs of the transport processes and physical and chemical parametrisations.
- Source term usually unknown.
- Parameters of the physical parametrisations often poorly know (effective turbulent diffusion, scavenging and dry deposition parameters).
Case study: Cost function

- Source-receptor relationship: $\mathbf{H}$. Linear model.
- Problem usually solved using 4D-Var [Bocquet, 2012] or methods equivalent to the representer technique. Here, study focused on UQ of the best estimate [Liu et al., 2017].
- Log-normal errors for the prior and for the observations. Non-Gaussian statistics.
- Cost function from Bayes’ rule:

$$
\mathcal{L}(\mathbf{z}; \theta) = -\ln p(\mathbf{z}|\mathbf{y}, \theta) = -\ln p(\mathbf{y}|\mathbf{z}, \theta) - \ln p(\mathbf{z}|\theta) + \ln p(\mathbf{y}|\theta)
$$

$$
= \frac{1}{2} \| \ln \mathbf{y} - \ln \mathbf{H} \mathbf{x} \mathbf{e} \mathbf{z} \|_R^2 + \frac{1}{2} \| \mathbf{z} \|_B^{-1} + \frac{1}{2} \ln |\mathbf{R}| + \frac{1}{2} \ln |\mathbf{B}| + \xi.
$$

- Two strategies to quantify the uncertainty of the best estimate:
  - Bayesian hierarchy (HB):
    $$
    p(\mathbf{x}, \theta|\mathbf{y}) = \frac{p(\mathbf{y}|\mathbf{x}, \theta)p(\mathbf{x}|\theta)p(\theta)}{p(\mathbf{y})}, \quad p(\mathbf{x}|\mathbf{y}) = \int d\theta \ p(\mathbf{x}, \theta|\mathbf{y}).
    $$
  - Empirical Bayes (EB):
    $$
    p(\mathbf{x}|\mathbf{y}) \approx p(\mathbf{x}|\mathbf{y}, \theta^*). \quad (3)
    $$

$\theta^*$ here estimated by the Expectation-Maximisation (EM) algorithm.
Case study: Inversions (EB)

- Uniform hyperparameters: \( R = r^* I, \ B = b^* I, \) where \( r^* \) and \( b^* \) are obtained from EM.

- Chernobyl and Fukushima-Daiichi source terms with Gaussian and lognormal assumptions on the observation errors. Comparison with the Unscear reference source term.
Case study: UQ of the retrieved total radioactivity (EB)

- Probability density function of the total released activity for Chernobyl and Fukushima-Daiichi.

 EB: optimal hyperparameters are first determined. Followed by nonlinear sampling of the total activity using three methods: with a Laplace proposal, a random-then-optimise sampling, an unbiased random-then-optimise sampling and a basic MCMC.
Case study: Inversion (HB)

- Full solution of the Bayesian hierarchy (HB)
- Obtained from a Monte Carlo Markov Chain (MCMC)
- Transdimensional analysis (adaptive grid). Here using only 20 grid cells for Chernobyl and 40 grid cells for Fukushima.
Case study: UQ of the retrieved total radioactivity (HB)

- Probability density function of the total released activity for Chernobyl and Fukushima-Daiiichi.

- Full solution of the Bayesian hierarchy; obtained from an MCMC.

- Transdimensional analysis (adaptive grid). Here using only 20 grid cells for Chernobyl and 40 grid cells for Fukushima.
Final word

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.