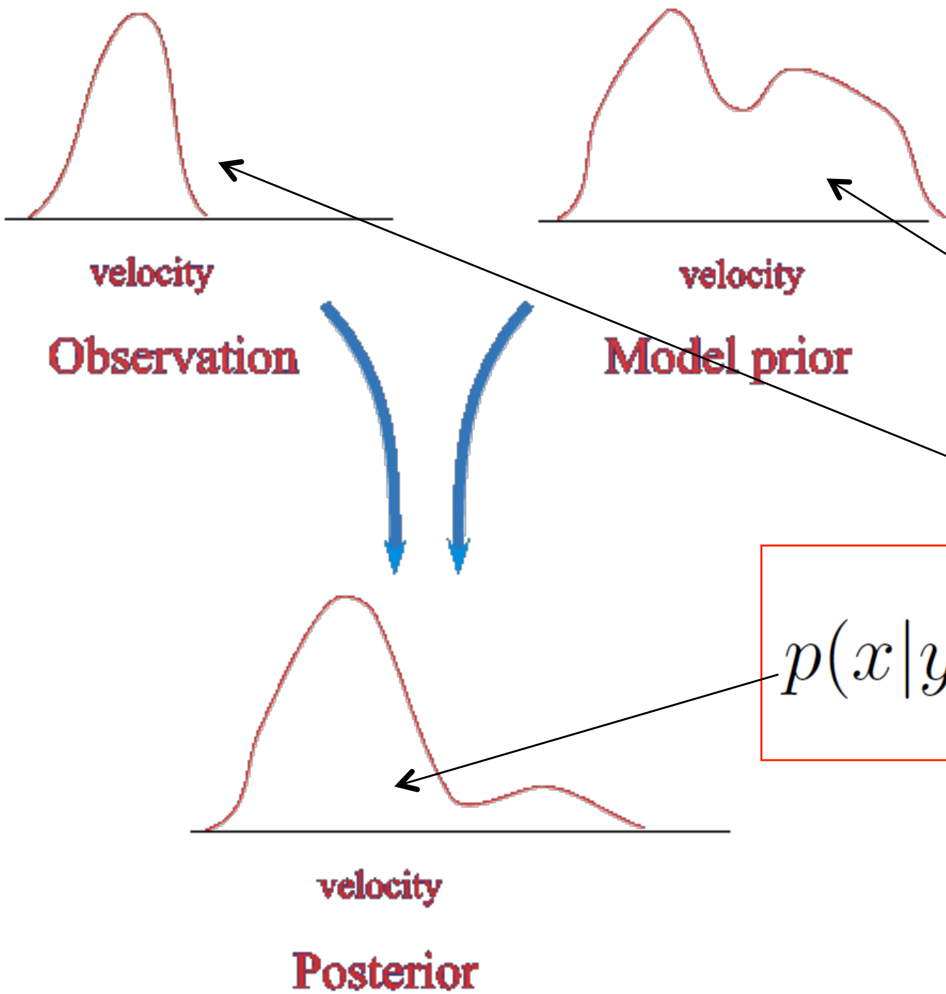
Overview of data-assimilation methods
(and what to use when…)

Peter Jan van Leeuwen
Data-Assimilation Research Centre
University of Reading
United Kingdom
p.j.vanleeuwen@reading.ac.uk

Data assimilation: general formulation



Bayes theorem:

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

The solution is a pdf!

3DVar and Optimal Interpolation

Assumptions:

- Prior is Gaussian
- Observation errors are Gaussian
    - a) H linear -> Optimal interpolation
    - b) H nonlinear -> 3DVar

$$p(x|y) \propto \exp\left[-\frac{1}{2}J\right]$$

$$J = (x - x_b)^T B^{-1} (x - x_b) + (y - H(x))^T R^{-1} (y - H(x))$$

3DVar and Optimal Interpolation

Characteristics:

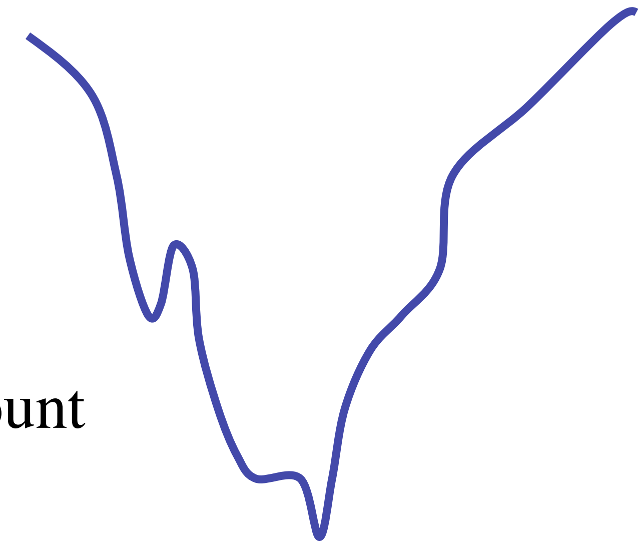
- Both find the mode of the posterior pdf
- Both typically do not provide an error estimate
- Extensively used in real systems
- Strong theoretical background

Ingredients:

- B error covariance of the model state
- H observation operator
- R Observation error covariance

Potential problems:

- Rely heavily on correct B matrix
- Doesn't take system evolution into account
- Can end up in local minima:



4DVar

Assumptions:

- Prior is Gaussian
- Observation errors are Gaussian
- H can be nonlinear
- strong and weak constraint

$$p(x|y) \propto \exp\left[-\frac{1}{2}J\right]$$

$$J = (x - x_b)^T B^{-1}(x - x_b) + (y - H(x))^T R^{-1}(y - H(x))$$

in which H contains the model operator, and H^T its adjoint.

4DVar

Characteristics:

- Finds the mode of the posterior pdf joint in time
- Needs adjoint equations
- Extensively used in real systems
- Strong theoretical background

Ingredients:

- B error covariance of the model state
- H_k observation operator at each observation time k
- R Observation error covariance
- (Q model evolution error covariance)
- Tangent-linear model and adjoint

Potential problems:

- Relies heavily on correct B matrix
- Typically no error estimate
- Difficult to make parallel
- Can end up in local minima:

Kalman Filter

Assumptions:

- Prior is Gaussian
- Observation errors are Gaussian
- H is linear (nonlinear extension: Extended KF)

$$x_a = x_b + BH^T(HBH^T + R)^{-1}(y - Hx_b)$$

$$P = (1 - KH)B$$

Kalman Filter

Characteristics:

- Propagates model and error covariance with linear (linearised) model
- Finds mean of posterior pdf, assuming linearity/Gaussianity
- Finds covariance of posterior pdf, assuming linearity/Gausianity
- Strong theoretical background

Ingredients:

- H observation operator
- R observation error covariance
- M linear (linearised) model operator
- (Q model evolution error covariance)

Potential problems:

- P too large to store for large-dimensional problems

Ensemble Kalman Filters

Assumptions:

- Prior is assumed Gaussian
- Observation errors are Gaussian
- H can be nonlinear
- Prior and posterior can be represented by small number of ensemble members

$$T = \left[1 + (X_b H)^T R^{-1} H X_b\right]^{-1/2}$$

$$K = X^f T T^T (H X^f)^T R^{-1}$$

$$\overline{x^a} = \overline{x^f} + K(y - H\overline{x^f})$$

$$X_a = X_b T$$

Ensemble Kalman Filters

Characteristics:

- Finds mean of the posterior pdf, assuming linearity/Gaussianity
- Finds ‘covariance’ of posterior pdf, assuming linearity/Gausianity
- Uses full nonlinear model through ensemble integrations
- Used extensively in real large-dimensional systems
- Rather weak theoretical background
- Extremely easy to make parallel

Ingredients:

- H observation operator
- R Observation error covariance
- (Q model evolution error covariance)
- Ensemble of model states

Potential problems:

- Needs inflation to avoid filter divergence, this needs tuning
- Needs localisation to counter rank deficiency and spurious correlations, localisation radius needs tuning

Hybrid 4DVar-EnKF

Assumptions:

- Prior is assumed Gaussian
- Observation errors are Gaussian
- H can be nonlinear (but needs linearisations)

Several different variants, the field is strongly in development

Hybrid 4DVar-EnKF

Characteristics:

- Flow-dependent B matrix
- Well-defined for linear problems
- Weak theoretical background for nonlinear problems
- Some variants can be made parallel and avoid adjoint

Ingredients:

- B model error covariance
- H observation oprator
- R observation error covariance
- (Q model evolution error covariance)
- (Tangent linear model and adjoint)
- Ensemble of model states

Potential problems:

- Needs inflation to avoid filter divergence. This needs tuning
- Needs localisation, localisation radius needs tuning
- Can end up in local minima

Particle Filters

Assumptions:

• Prior and Posterior pdf can be represented by small number of particles

$$p(x) = \sum_{i=1}^{N} \frac{1}{N} \delta(x - x_i)$$

$$p(x|y) = \sum_{i=1}^{N} w_i \delta(x - x_i)$$

$$w_i = \frac{p(y|x_i)}{\sum_j p(y|x_j)}$$

Particle Filters

Characteristics:

- Uses full nonlinear model through ensemble integrations
- Uses fully nonlinear update through Bayes theorem
- Needs to explore proposal density for efficiency
- Extremely parallel
- Strong theoretical background

Ingredients:

- H Observation operator
- R observation error covariance
- Q model evolution error covariance
- Ensemble of model states
- Efficient proposal density

Potential problems:

- Proposal density has tuning parameters
- No experience with real large-dimensional systems

Summary

Method	Description	Pros	Cons
A. Data insertion	Set grid points to observation values	1. Easy to do	1. No respect of uncertainty 2. What about observation voids? 3. Can't deal with indirect observations
B. Variational data assimilation	Minimize a cost function Many flavours: 3D, 4D, weak/ strong constraint	1. Respect of data uncertainty 2. Direct and indirect observations **3.** $\mathbf{P}_f$ gives smooth and balanced fields 4. Efficient 5. Can deal with (weakly) non-linear **h**	**1.** $\mathbf{P}_f$ is difficult to know, often static and suboptimal 2. High development costs **3.** **h**: need tangent linear, **H** and adjoint, $\mathbf{H}^T$ 4. Gaussian pdf
C. Kalman filtering	Evaluate KF equations	1. As B.1, B.2, B.3 **2.** $\mathbf{P}_f$ adapts with the state	1. As B.3, B.4 2. Difficult to use with non-linear **h** 3. Prohibitively expensive for large *n*
D. Ensemble Kalman filtering	Approximate KF equations with ensemble of *N* model runs Many flavours	1. As B.1,B.2, B.4, B.5, C.2 **2.** **h**: do not need **H** and $\mathbf{H}^T$ 3. Have measure of analysis spread	1. As B.4 2. Serious sampling issues when $N << n$ 3. Need ensemble inflation and localization schemes to overcome D.2
E. Hybrid	Cross between C/ D	1. As B.1, B.2, B.3, B.4, B.5, C.2	1. As D.2
F. Particle filter	Assign weights to ensemble members to represent any pdf	1. As. B.1, B.2 2. Can deal with non-linear **h** 3. Can deal with non-Gaussian pdf 4. Have measure of analysis spread	1. As D.2 2. Inefficient – members often become redundant 3. Need special techniques to overcome F.2

When to use what?

- When an adjoint is available use it!
- If not, it is hard to code up.
- Ensemble software code is available, relatively easy to add model
- If your system is not strongly nonlinear use 3/4DVar or EnKF
- If your system is strongly nonlinear use Particle Filter

Software support

- Explore TAF TAMC automatic adjoint compiler e.g. Ralph Giering
  Expensive, few 1000£ a year. Free compilers available, but not as fully featured. (Tapenade, …)
  http://www.fastopt.com/ for TAMC
  http://www-sop.inria.fr/tropics/tapenade.html for Tapenade
- Explore ensemble DA software packages like DART, PDAF and EMPIRE, typically no adjoint (but EMPIRE developing)
  Particle filters are now being implemented too in these packages.
  http://www.image.ucar.edu/DAReS/DART/ for DART
  http://pdaf.awi.de/trac/wiki for PDAF
  http://www.met.reading.ac.uk/~darc/empire/index.php for EMPIRE
  http://www.data-assimilation.net/ for DA tools in SANGOMA

Outlook

We will provide aftercare:

keep in touch, and ask for help if needed.

We hope you ENJOYED it!!!