Introduction to Data Assimilation

Data Assimilation Training Course
IIRS, ISRO, Dehra Dun
17-21 December 2012

Peter Jan van Leeuwen
Data Assimilation Research Center (DARC)
University of Reading
p.j.vanleeuwen@reading.ac.uk

How do we process new data?

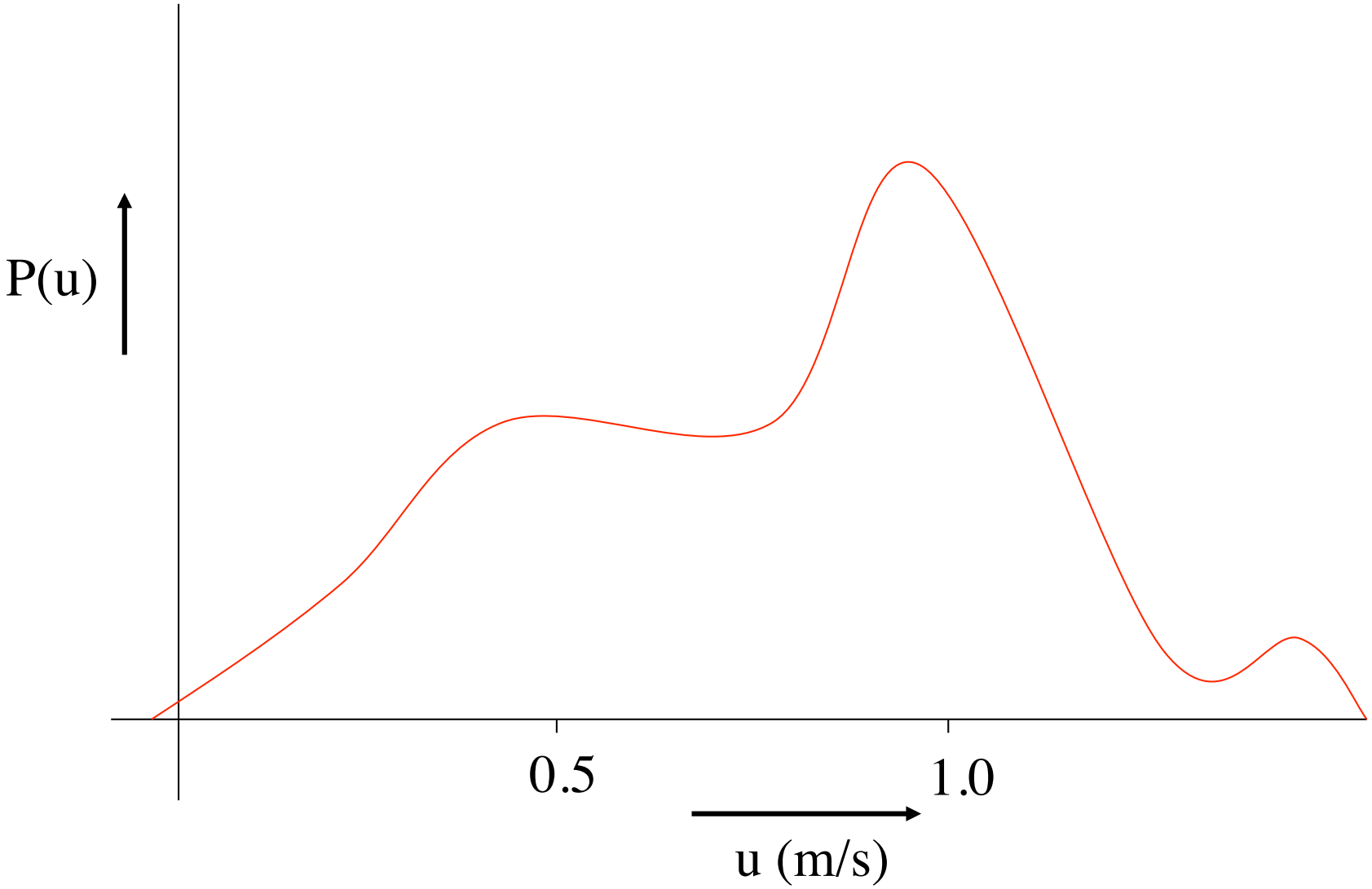
A process description

- Prior knowledge, from a model, a cat
- Observations, the dog
- Posterior knowledge, improvement of the model, the dog that has eaten the cat

What is missing?

Uncertainty !!!

A mathematical description: probability density functions



Intermezzo: conditional pdf

Conditional pdf: $$p(x, y) = p(x|y)p(y)$$

Similarly: $$p(x, y) = p(y|x)p(x)$$

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

Intermezzo: conditional pdf

We can use:

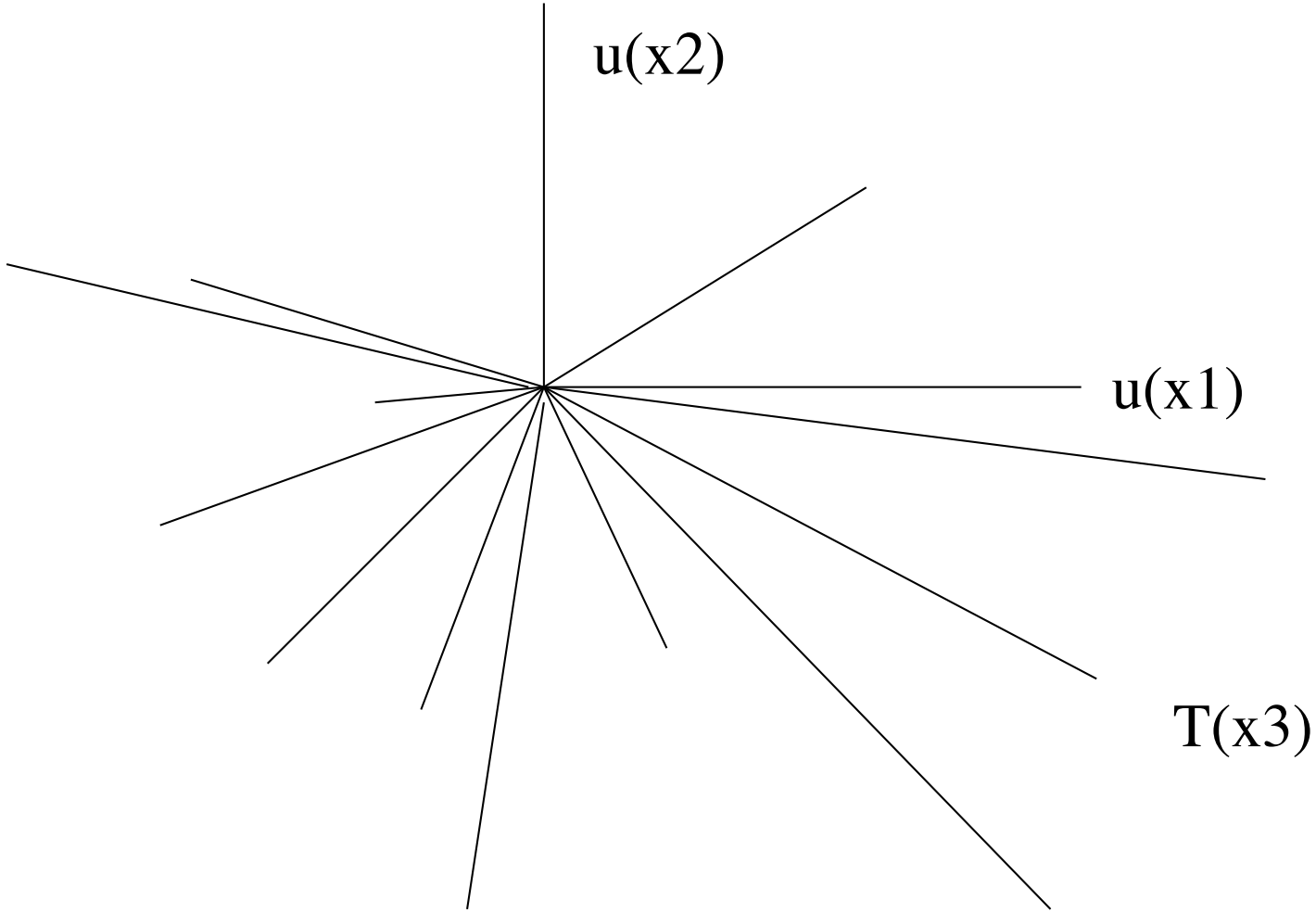
$$p(y) = \int p(x, y)\, dx = \int p(y|x)p(x)\, dx$$

Bayes Theorem

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

The model pdf *p(x)*

p (u(x1), u(x2), T(x3), …)



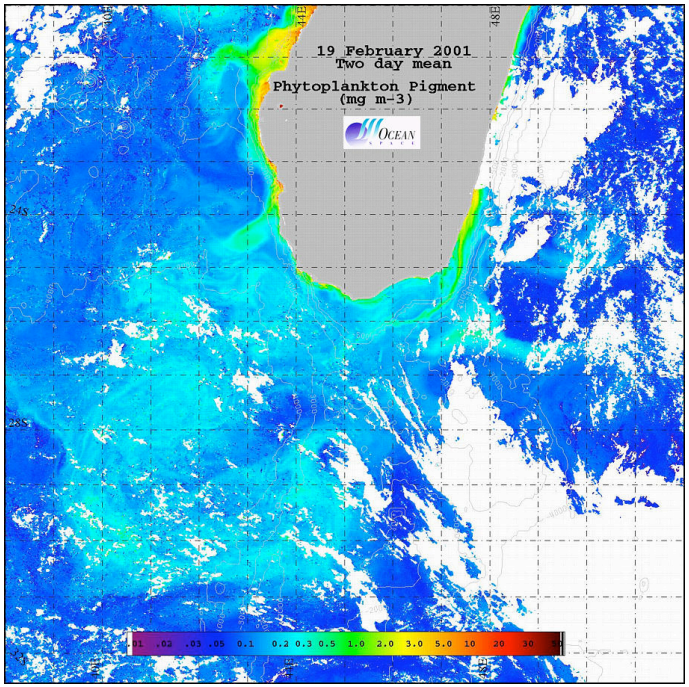
Observations
p(y|x)

- In situ observations:

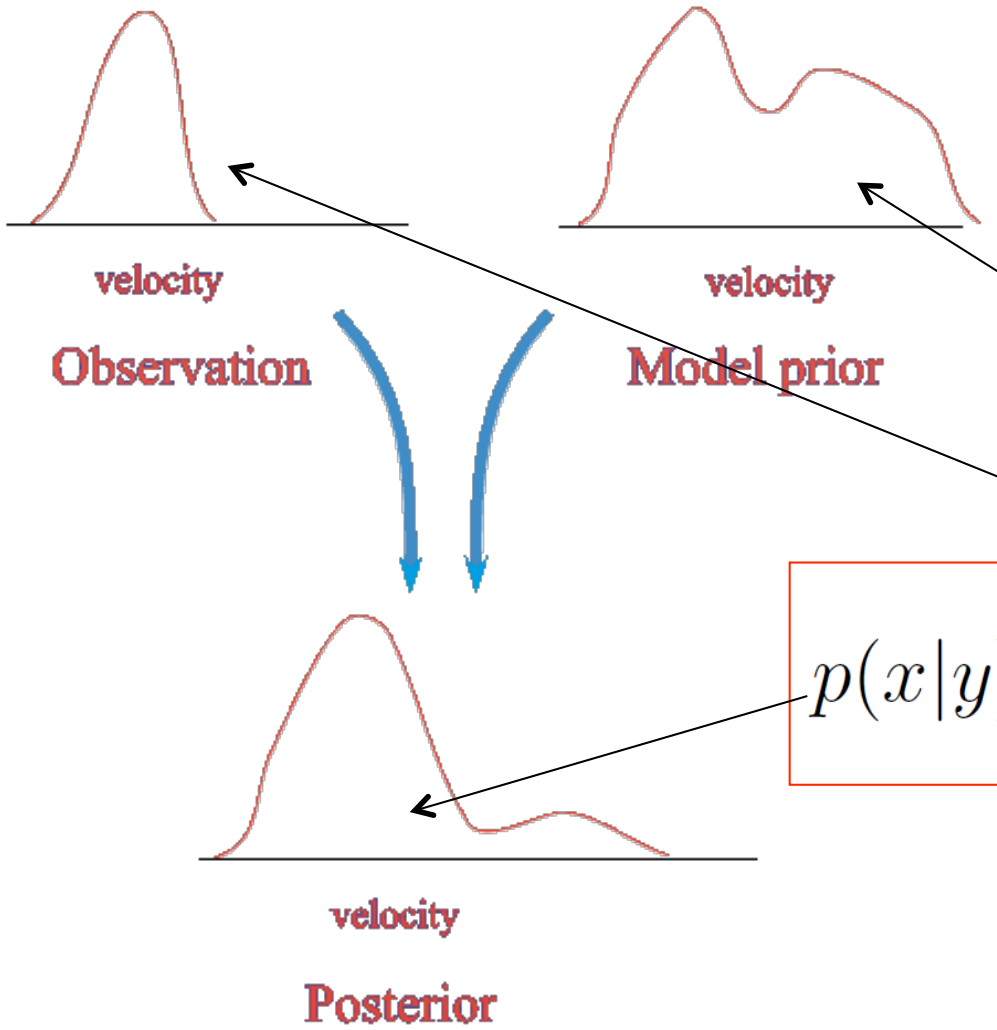
  irregular in space and time e.g. sparse hydrographic observations,

- Satellite observations: indirect

  e.g. of the sea-surface



Data assimilation: general formulation

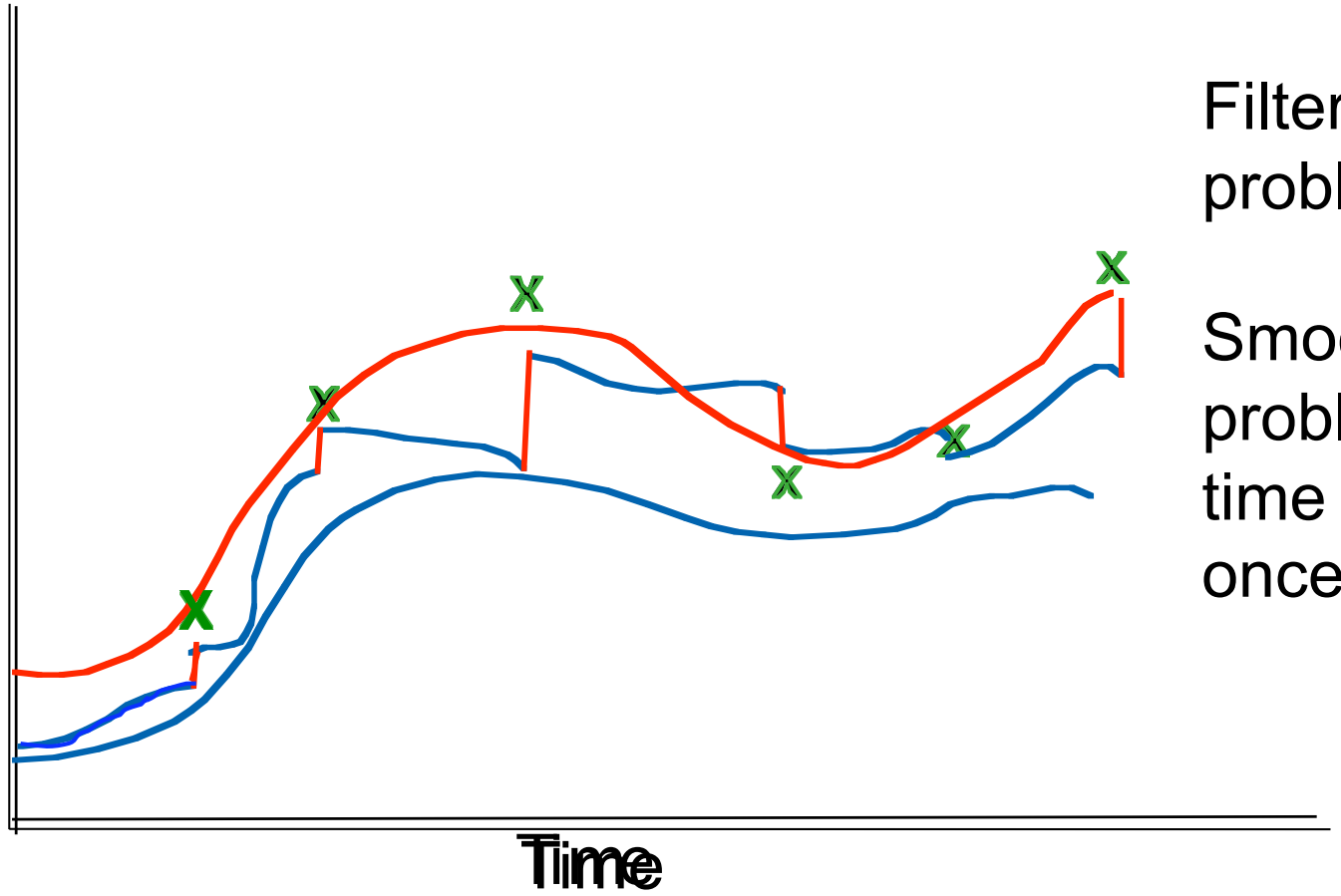


Bayes theorem:

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

The solution is a pdf!

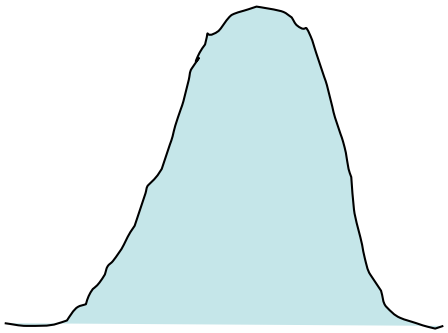
Filters and smoothers



Filter: solve 3D problem sequentially

Smoother: solve 4D problem in specific time window all at once

The Gaussian assumption



$$p(T) = \frac{1}{\sqrt{2\pi\sigma}} \exp\left(-\frac{\left(T-\overline{T}\right)^2}{2\sigma^2}\right)$$

Prior pdf: multivariate Gaussian:

$$p(x) \propto \exp\left[-\frac{1}{2}(x-x_b)^T B^{-1}(x-x_b)\right]$$

Likelihood: multivariate Gaussian

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

(Ensemble) Kalman Filter I

Use Gaussianity in Bayes at a specific time:

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

Multiplication:

$$p(x|y) \propto \exp\left[-\frac{1}{2}(x - x_b)^T B^{-1}(x - x_b) - \frac{1}{2}(y - H(x))^T R^{-1}(y - H(x))\right]$$

Complete the squares to find again a Gaussian (only for linear H !!!):

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

(Ensemble) Kalman Filter III

Two possibilities to find the expressions for the mean and covariance:

1) Completing the squares
2) Assume solution is linear combination of model and observations.

Both lead to the Kalman filter equations, which are just the least squares solutions (best linear unbiased estimator, BLUE):

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

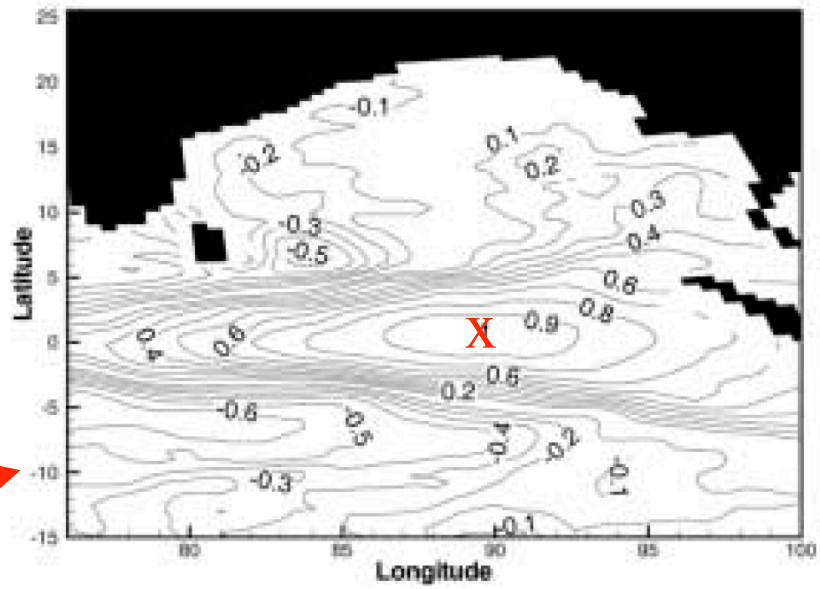
influence region — K the Kalman Gain — weighting — innovation

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

The error covariance:

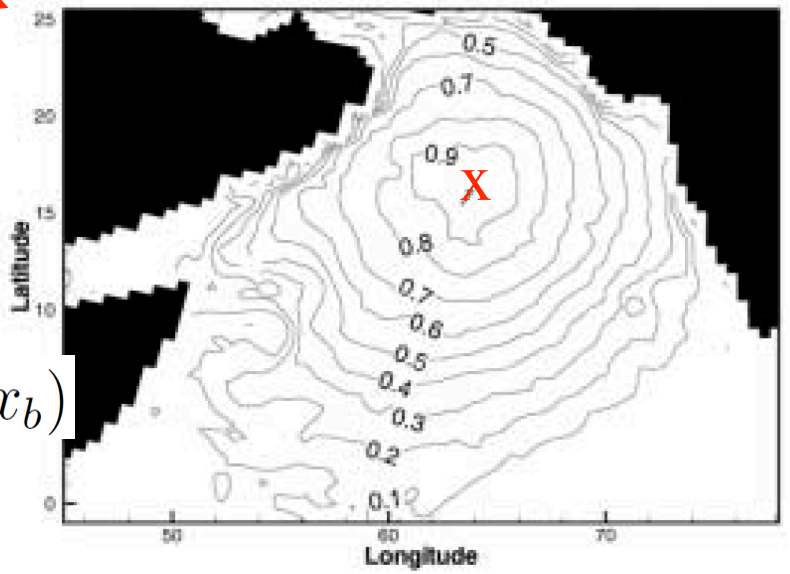
Tells us how model variables co-vary.

Spatial correlation of SSH and SST in the Indian Ocean



In the Kalman filter this comes in via the BH^T term:

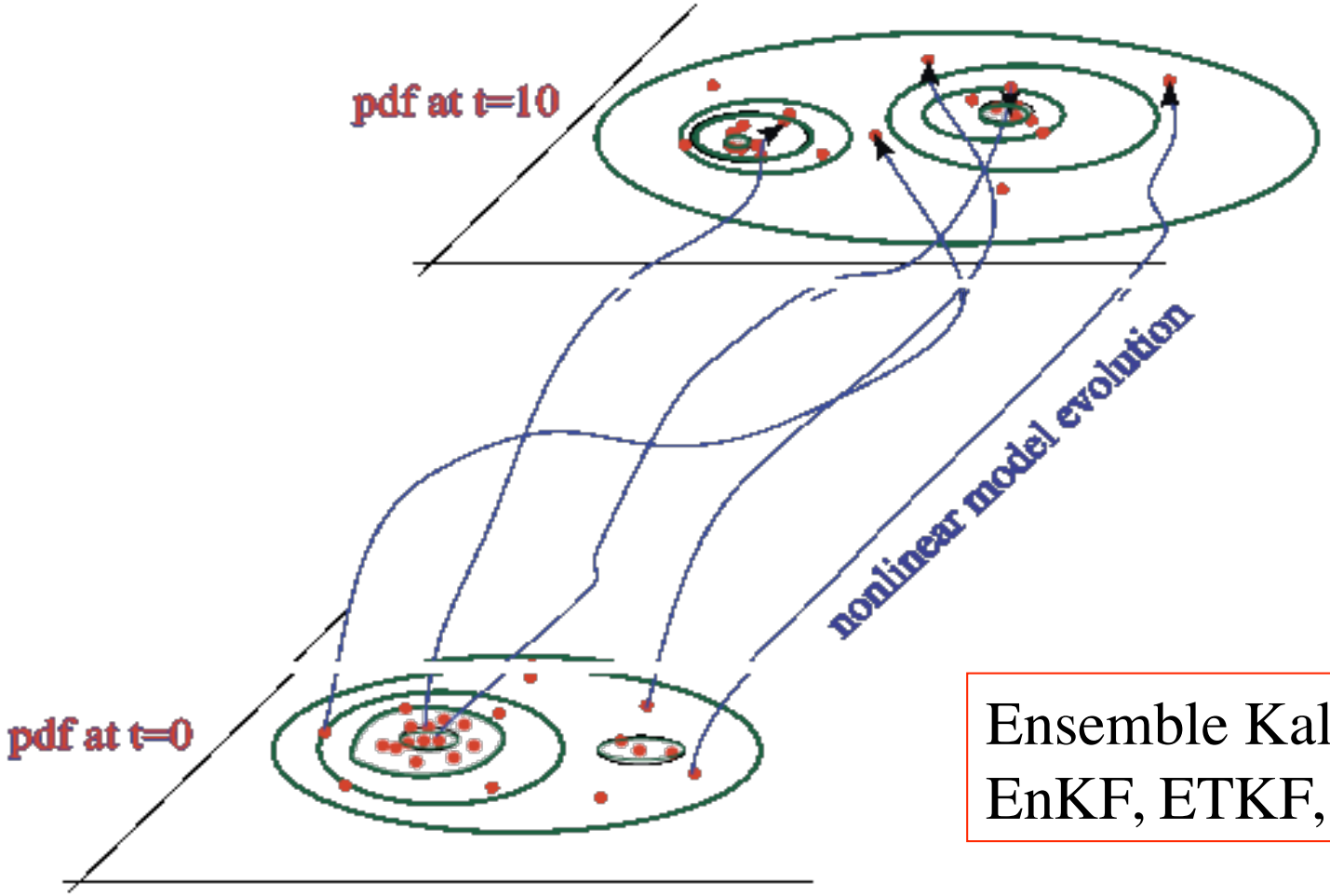
$$x_a = x_b + BH^T(HBH^T + R)^{-1}(y - Hx_b)$$



Haugen and Evensen, 2002

Kalman filters in practice: Ensembles

How to propagate (or even store) the covariance matrix?



Ensemble Kalman Filter
EnKF, ETKF, EAKF, …

Ensemble Kalman Filter: the update

Ensemble perturbation matrix

$$X_b = (x_b^1 - x_b, \cdots, x_b^N - x_b)^T$$

to represent prior covariance as:

$$B = \frac{1}{N-1} X_b X_b^T$$

Write posterior ensemble perturbations as:

$$X_a = X_b T \qquad \text{with} \qquad P = \frac{1}{N-1} X_a X_a^T$$

Use $P = (1 - KH)B$ to find

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

Variational methods

A variational method looks for the most probable state, which is the maximum of this posterior pdf also called the mode.

Instead of looking for the maximum one solves for the minimum of a so-called costfunction.

The pdf can be rewritten as $$p(x|y) \propto \exp\left[-\frac{1}{2}J\right]$$

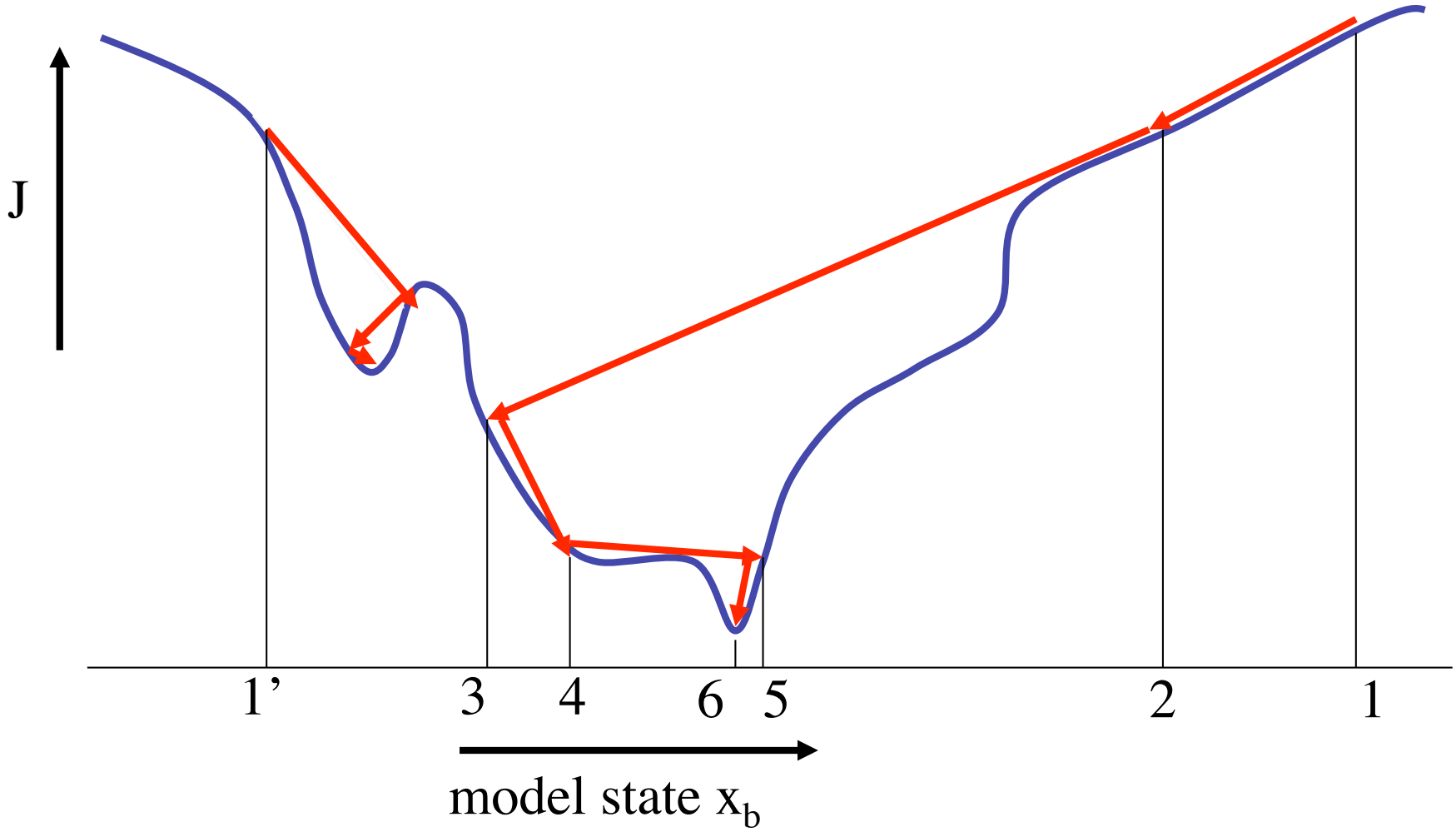
in which

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

Find min J from variational derivative:
J is costfunction or penalty function

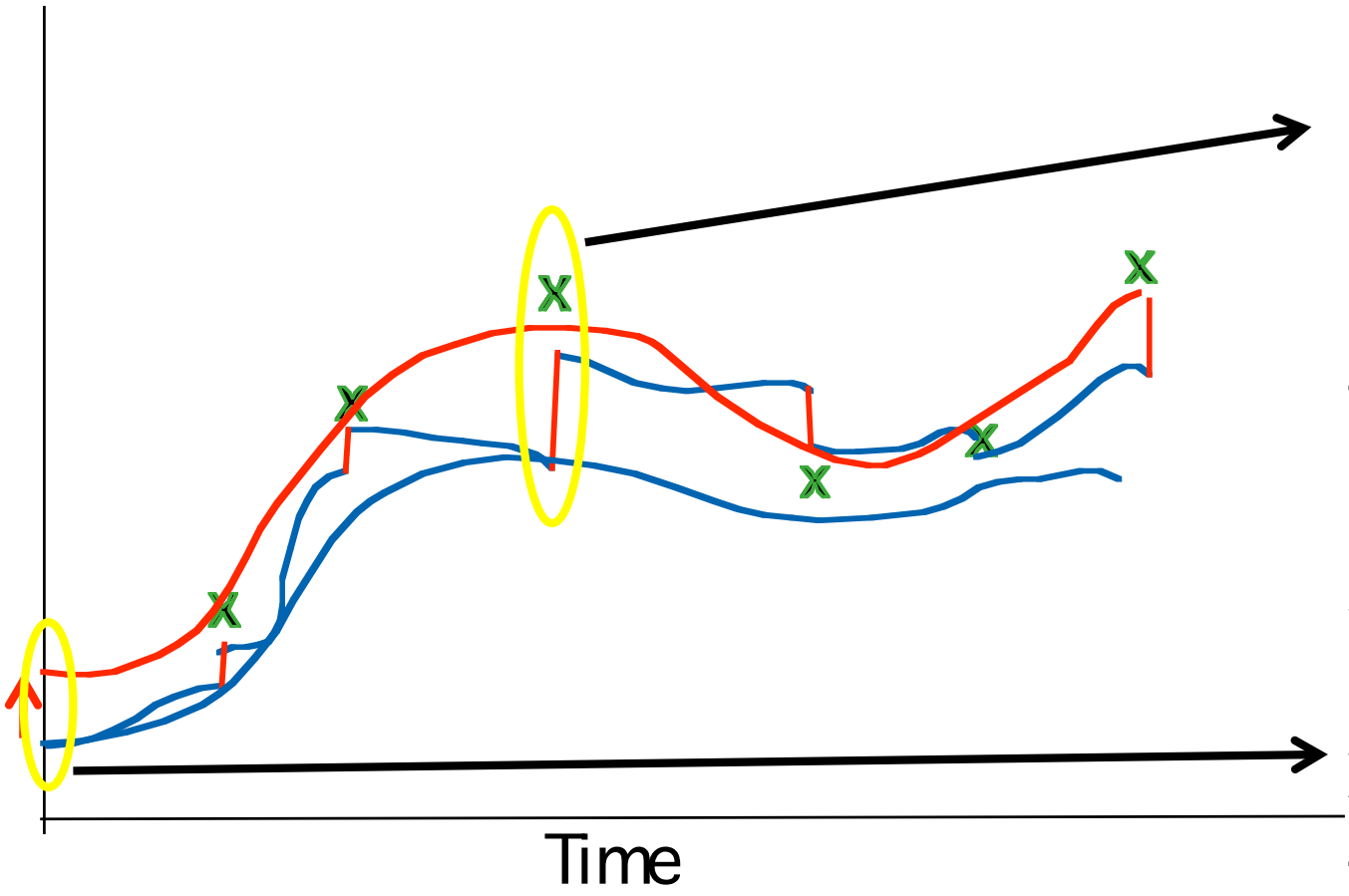
$$\frac{\delta J}{\delta x_b} = 0$$

Gradient descent methods: Gauss-Newton iterations



4DVar

There is an interesting extension to this formulation to a smoother.



Filter:
Solve a 3D problem at each observation time.

4DVar:
Solve a 3D problem at the beginning of the time window using all observations.
Note we get a new full 4D solution!

4DVar: the dynamical model

The dynamical model is denoted by *M*:

$$x_1 = M(x_0)$$

Using the model operator twice brings us to the next time step:

$$x_2 = M(x_1) = M(M(x_0))$$

And some short-hand notation:

$$x_2 = M_{0\rightarrow 2}(x_0)$$

4DVar: the costfunction

The total costfunction that we have to minimize now becomes:

$$J = (x - x_b)^T B^{-1}(x - x_b) + \sum_{t_{obs}=1}^{M} (y_i - H_i(x))^T R^{-1}(y_i - H_i(x))$$

in which the measurement operator H_i contains the forward model:

$$H_i(x) = H(M_{0\rightarrow i}(x_0))$$

This nonlinear costfunction is minimised iteratively.

4DVar: the adjoint

The solution to the linear iterates can be written as:

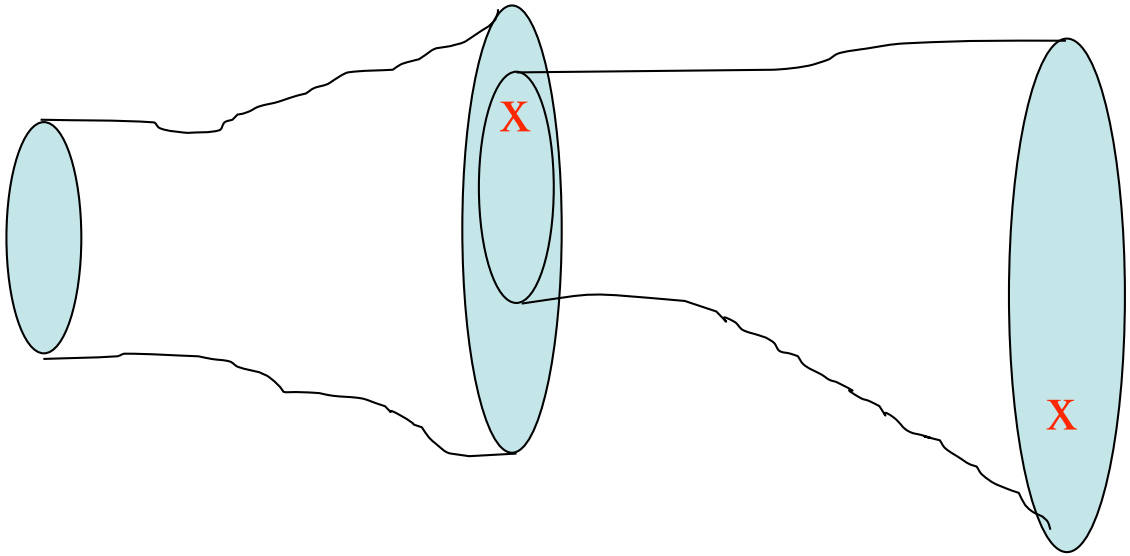
$$x_a = x_b + BH^T(HBH^T + R)^{-1}(y - Hx_b)$$

in which H now contains the model equations.

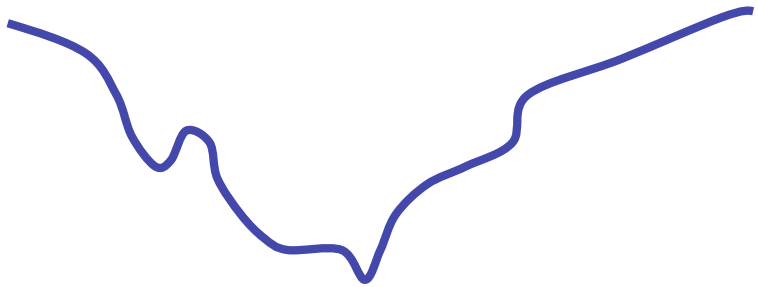
Note that H^T contains the adjoint model equations, running from end of the time window to the initial time.

Present-day data-assimilation methods for NWP:

- EnKF:



- 4DVar:



- Hybrid methods: Combine the best of both.
- Nonlinear data-assimilation methods…

Nonlinear filtering: Particle filter

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

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

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

with $w_i = \frac{p(y|x_i)}{\sum_j p(y|x_j)}$ the weights.

What are these weights?

- The weight w_i is the normalised value of the pdf of the observations given model state x_i.
- For Gaussian distributed variables is is given by:

$$\begin{aligned} w_i &\propto p(y|x_i) \\ &\propto \exp\left[-\frac{1}{2}\left(y - H(x_i)\right) R^{-1} \left(y - H(x_i)\right)\right] \end{aligned}$$

- One can just calculate this value
- That is all !!!
- Or is it? More needed for high-dimensional problems…

Summary and outlook

- We know how to formulate the data assimilation problem using Bayes Theorem.
- We have derived the Kalman Filter and shown that it is the best linear unbiased estimator (BLUE).
- We derived 3D and 4DVar and discussed some of their properties.
- We looked at a fully nonlinear method, the particle filter.
- This forms the basis for what is to come the rest of the week!

ENJOY