Variational methods in data assimilation

Summer school on data assimilation

IIRS, India, December 2012

Javier Amezcua
Ack: Ross Bannister, Nancy Nichols, Stefano Migliorini
Data Assimilation Research Center
University of Reading

Content

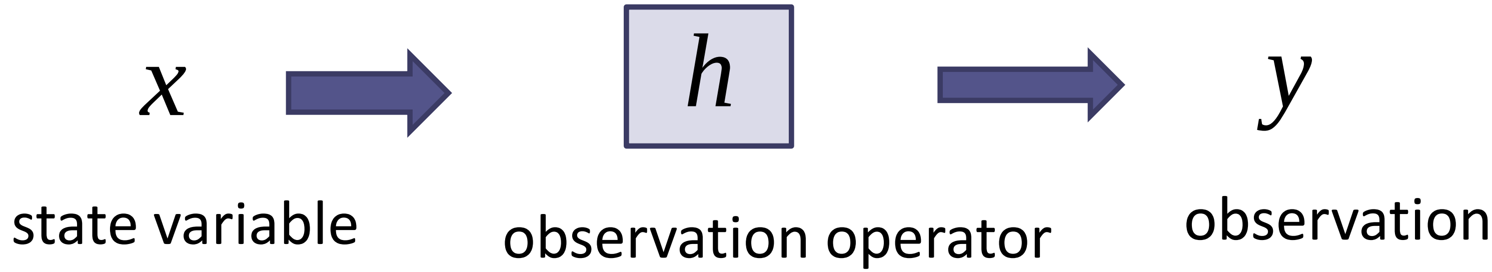
1. The inverse problem

2. Optimal interpolation and 3DVar

3. Advanced methods: 4DVar

The inverse problem

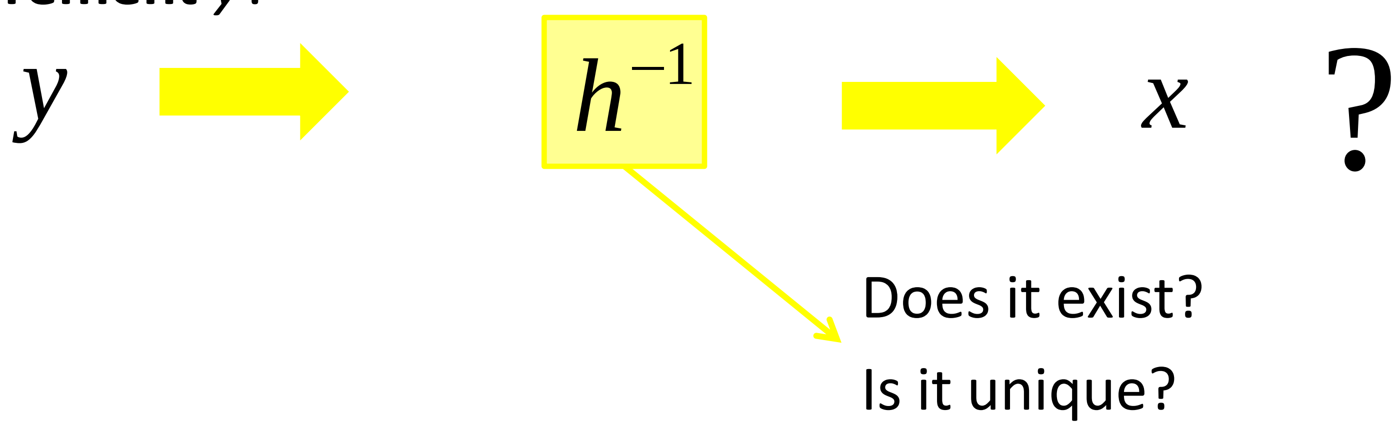
The inverse problem



$$y = h(x) + v$$

Observation error

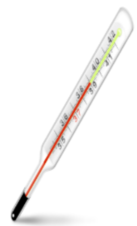
Can we infer the value of the **state variable** *x* with the information of the **measurement** *y*?

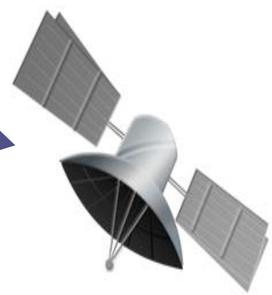


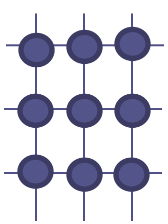
It depends... but in general it is an **'ill-posed' problem**.

The observation operator

state variable	observation operator	observation
Temperature in a location	$T_o = h(T) = T$ →	Temperature
T	$r_o = h(T) = \sigma T^4$ →	Radiance



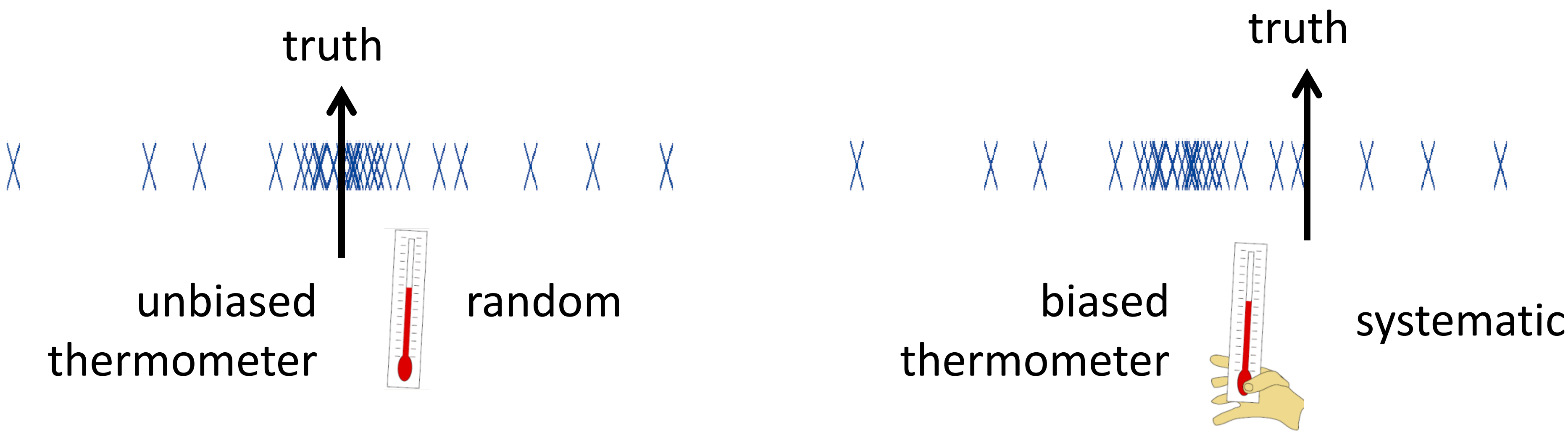




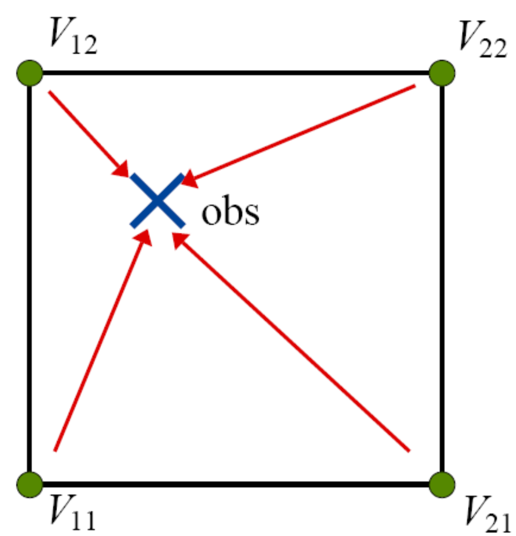
$$\mathbf{x} = \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_N \end{bmatrix} \longrightarrow H = \begin{bmatrix} h_{11} & h_{12} & \cdots & h_{1L} \\ h_{21} & h_{22} & \cdots & h_{2L} \\ \vdots & \vdots & \ddots & \vdots \\ h_{N1} & h_{N2} & \cdots & h_{NL} \end{bmatrix} \longrightarrow \mathbf{y} = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_L \end{bmatrix}$$

What are these observational errors?

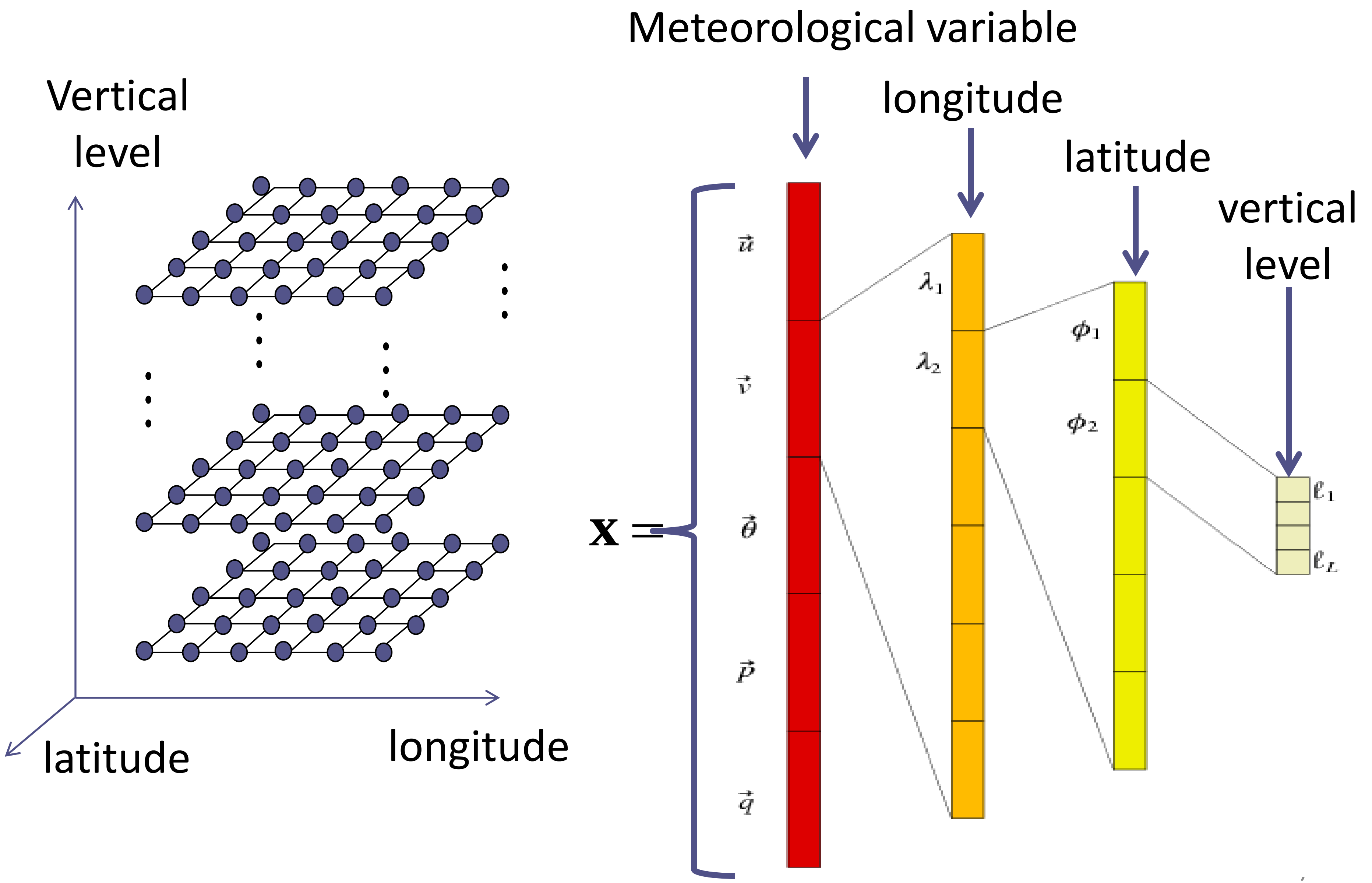
Example 1 – repeated observations of air temperature



Example 2 – representivity errors due to model grid



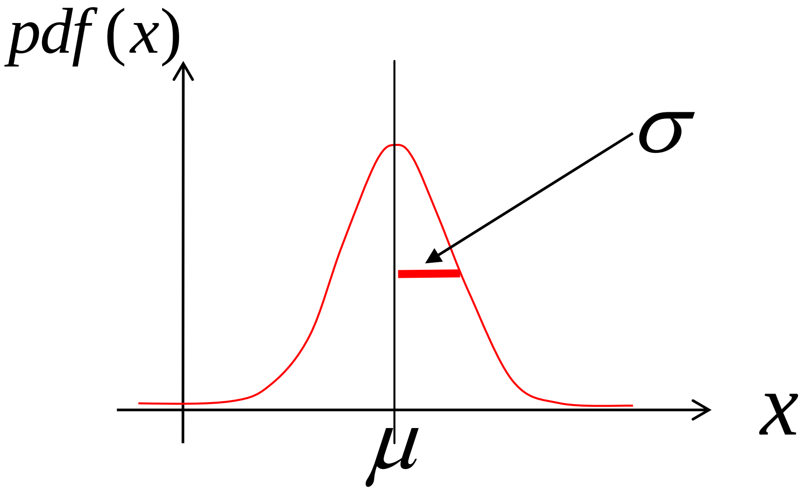
The state variables



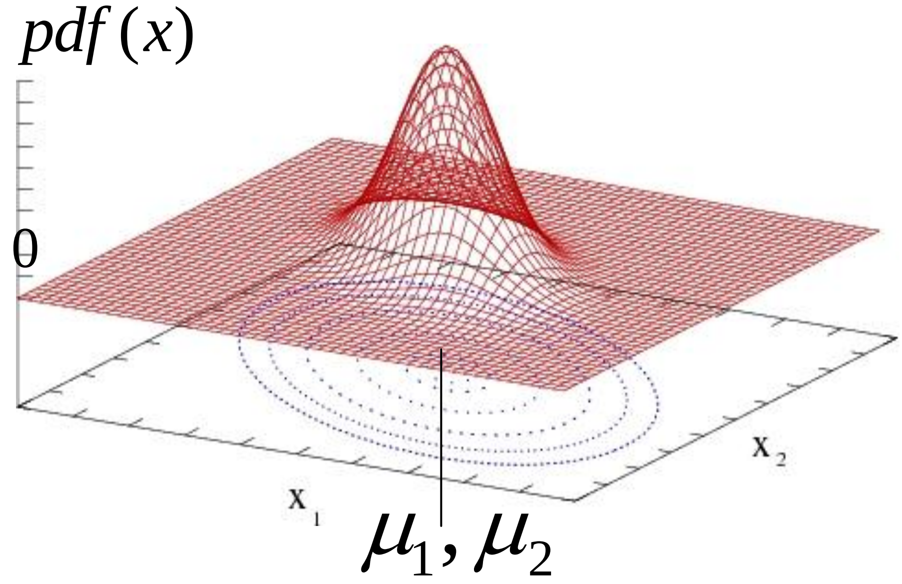
Statistical properties of errors

Errors (not only observational) are often considered to be Gaussian.

$$pdf(x) = \frac{1}{\sqrt{2\pi\sigma^2}} \mathrm{e}^{-\frac{(x-\mu)^2}{2\sigma^2}}$$



$$pdf(\mathbf{x}) = \frac{1}{\sqrt{(2\pi)^N |\mathbf{S}|}} \mathrm{e}^{-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})^{\mathrm{T}} \mathbf{S}^{-1} (\mathbf{x}-\boldsymbol{\mu})}$$



Solvability of the problem

- The problem can:

a) Have **exact solution**.

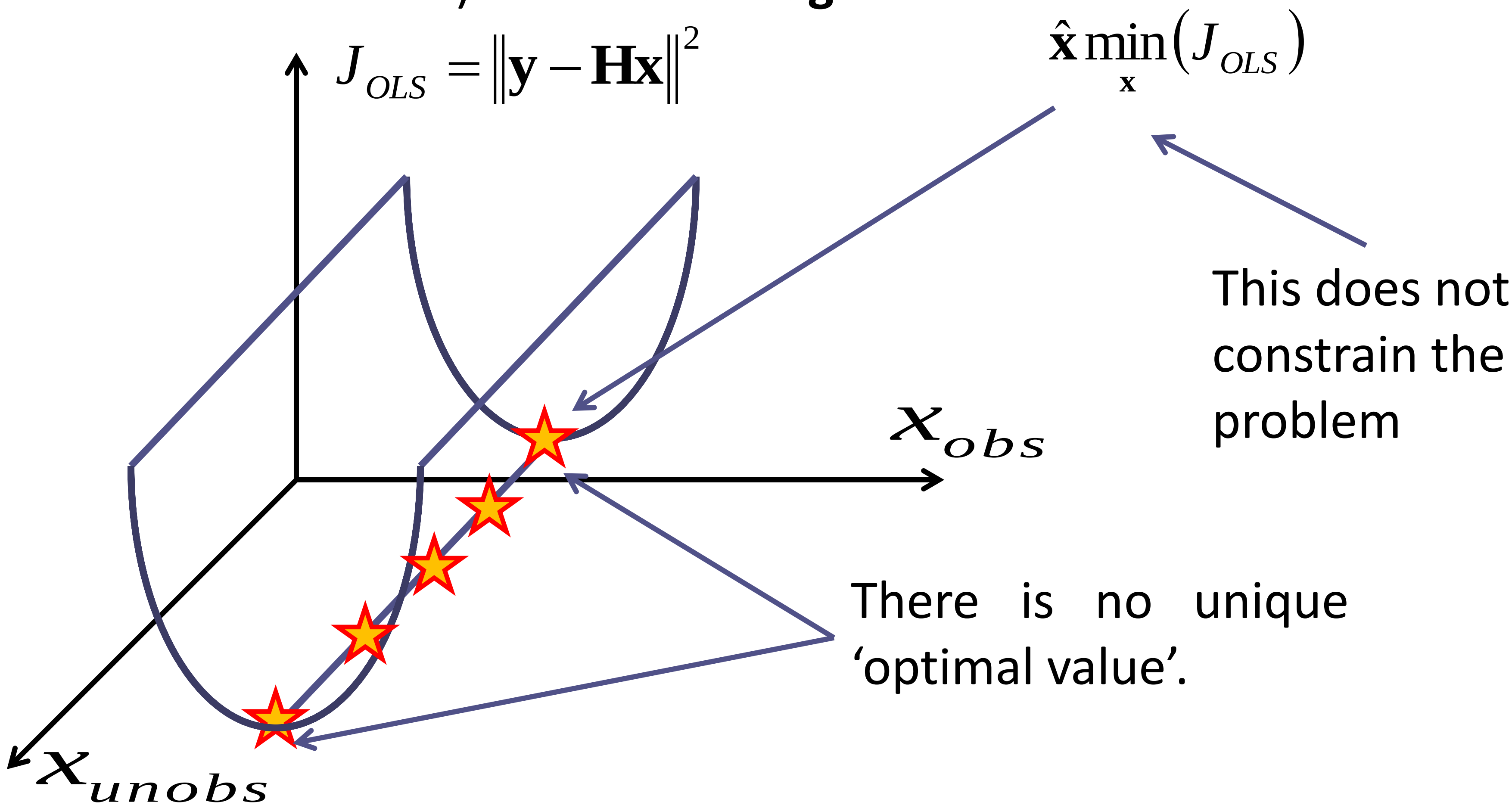
b) Be **overdetermined** (OLS approach)

c) Be **underdetermined** (regularization is required).

This involves minimizing a **residual**. (applies to b and c)

The underdetermined problem

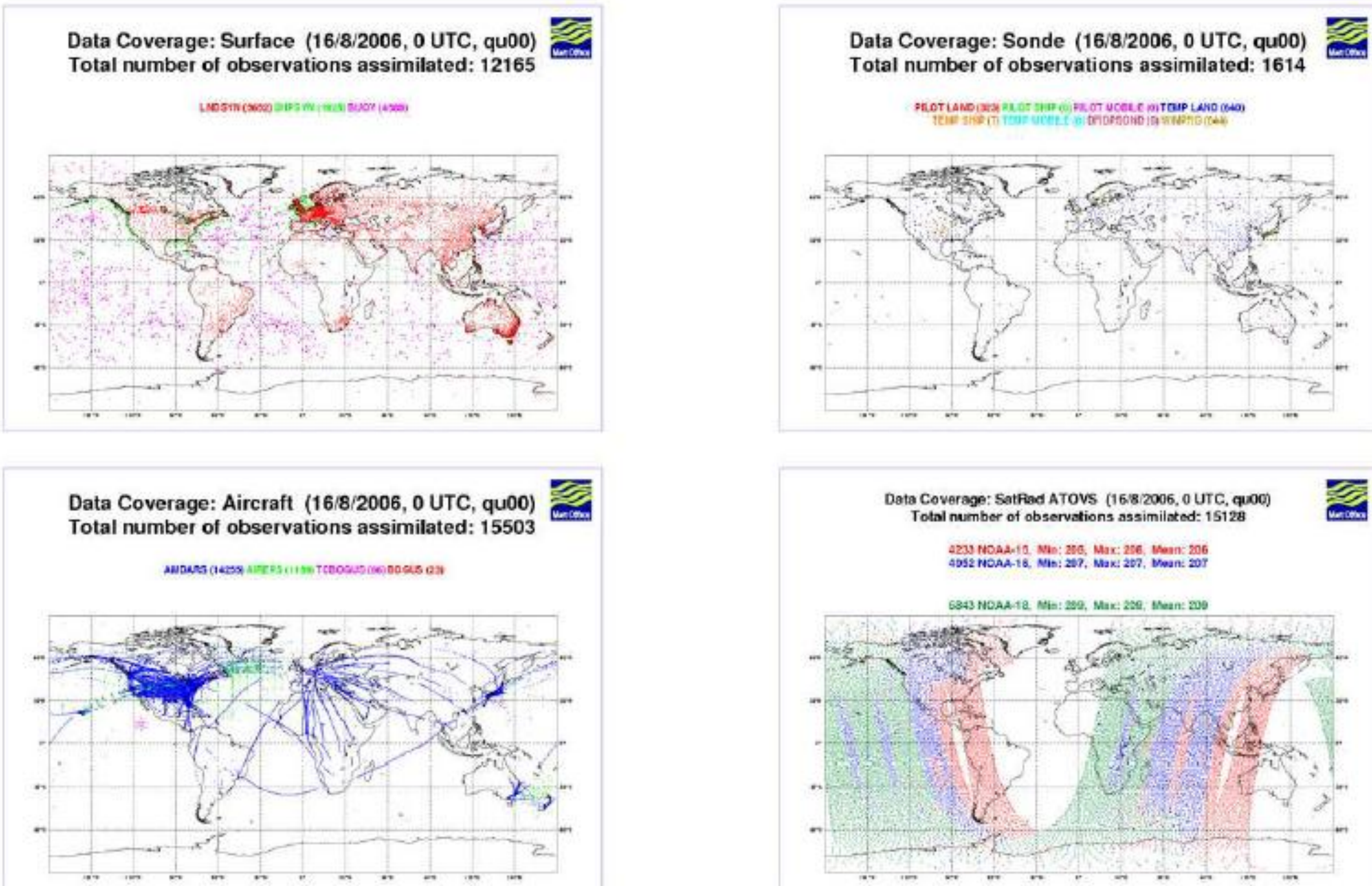
Does the traditional least squares solution help (as in the overdetermined case)? It is **not enough**!



In this case, the observations are not enough! We need **another source of information**.

The undetermined case

It happens a lot in real life!

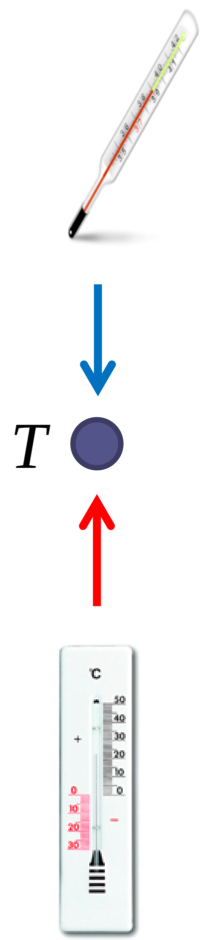


Example of observational coverage in the atmosphere. Met Office©.

Optimal Interpolation and 3DVar

Combining sources of information: a scalar example.

We want to determine the temperature at some location, and we use two thermometers to measure it.



$$T_1 = T + \varepsilon_1$$

(ε_1: error)

$$Var[\varepsilon_1] = \sigma_1^2$$

$$T_2 = T + \varepsilon_2$$

(ε_2: error)

$$Var[\varepsilon_2] = \sigma_2^2$$

Two important considerations:

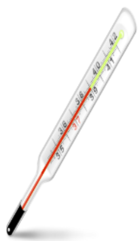
a) The thermometers are unbiased, i.e.

$$E[\varepsilon_1] = E[\varepsilon_2] = 0$$

b) The measurement errors are independent between the two thermometers.

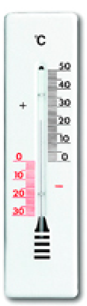
$$Cov[\varepsilon_1, \varepsilon_2] = 0$$

A simple scalar example



$$T_1 = T + \varepsilon_1$$

T



$$T_2 = T + \varepsilon_2$$

Let us linearly combine the two sources of information, and consider this our estimator.

$$\hat{T} = \underline{C_1 T_1} + \underline{C_2 T_2}$$

How do we choose these constants for the estimator to be **optimal**?

Optimal in what sense? It should:

a) Be unbiased:

$$E\left[\hat{T}\right] = T$$

b) Minimize the error of the estimation

$$\min_{\hat{T}} Var\left[\hat{T} - T\right] = \min_{\hat{T}} E\left[\left\|\hat{T} - T\right\|^2\right]$$

A simple scalar example

The conditions imply:

a) Be unbiased: $C_1 = (1 - C_2)$

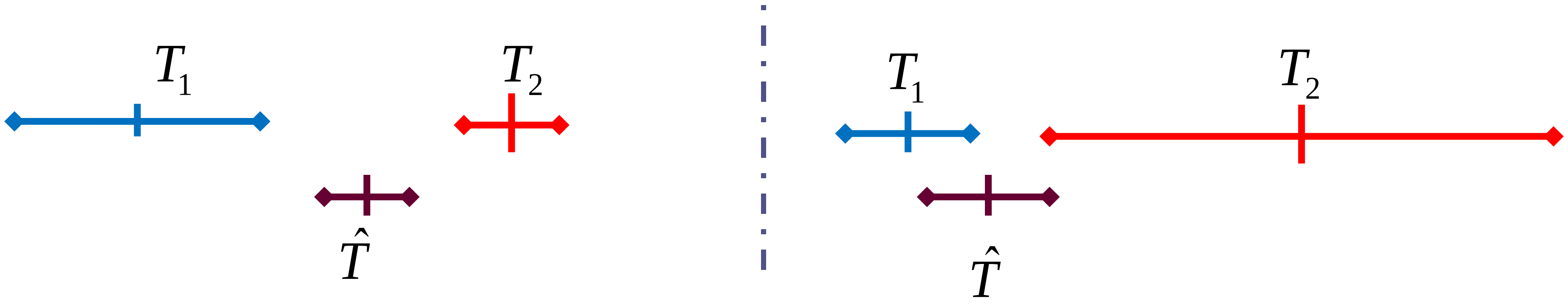
b) Minimize the residual $C_1 = \frac{\sigma_2^2}{\sigma_1^2 + \sigma_2^2}, C_2 = \frac{\sigma_1^2}{\sigma_1^2 + \sigma_2^2}$

$$\boxed{\hat{T} = \frac{\sigma_2^2}{\sigma_1^2 + \sigma_2^2} T_1 + \frac{\sigma_1^2}{\sigma_1^2 + \sigma_2^2} T_2}$$

This is the Best Linear Unbiased Estimator (BLUE)

A simple scalar example

$$\hat{T} = \frac{\sigma_2^2}{\sigma_1^2 + \sigma_2^2} T_1 + \frac{\sigma_1^2}{\sigma_1^2 + \sigma_2^2} T_2$$



Note that we minimized the cost function.

$$\hat{T} = \min_T J(T) = \frac{1}{2}\frac{(T_1 - T)^2}{\sigma_1^2} + \frac{1}{2}\frac{(T_2 - T)^2}{\sigma_2^2}$$

What are the sources of information?

A simple scalar example

What is the variance of this estimator?

$$\sigma_{\hat{T}}^2 = \mathrm{var}\left[\hat{T}\right] = \frac{\sigma_1^2 \sigma_2^2}{\sigma_1^2 + \sigma_2^2}$$

Note that: $\begin{cases} \sigma_{\hat{T}}^2 < \sigma_1^2 \\ \sigma_{\hat{T}}^2 < \sigma_2^2 \end{cases}$, i.e. the variance of the error is reduced with respect to any of the sources. Also

$$\boxed{\frac{1}{\sigma_{\hat{T}}^2}} = \boxed{\frac{1}{\sigma_1^2}} + \boxed{\frac{1}{\sigma_2^2}}$$

Precision of source 1

Precision of source 2

Precision of the optimal estimator

Sources of information

We have 2 sources of information for the **true** state of a system.

Dynamical **model**

$$\mathbf{x}_t = f(\mathbf{x}_{t-1}) + \mathbf{w}_t \quad \mathbf{x} \in \Re^N$$

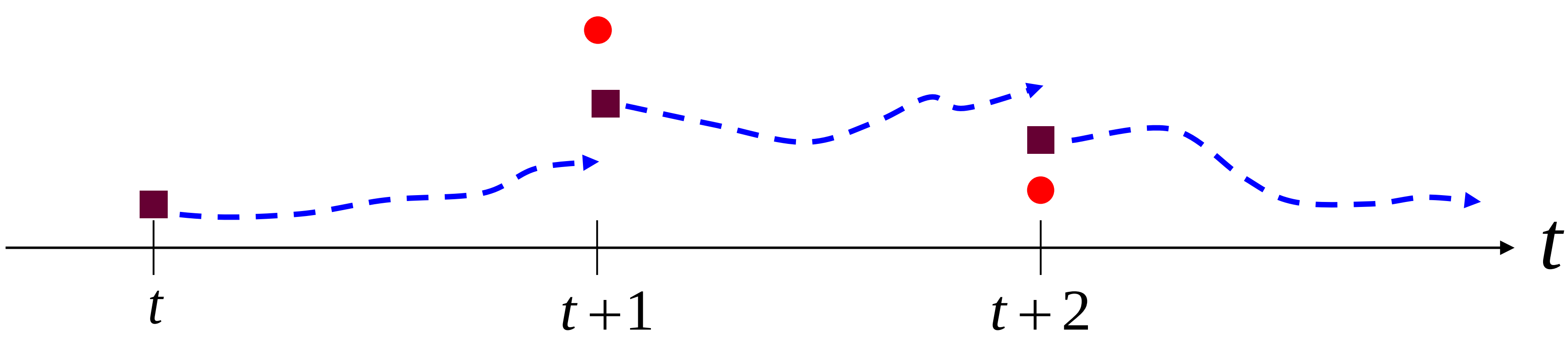
Model error: $E(\mathbf{w}_t) = \mathbf{0}$, $Cov(\mathbf{w}_t) = \mathbf{Q}$

Observations

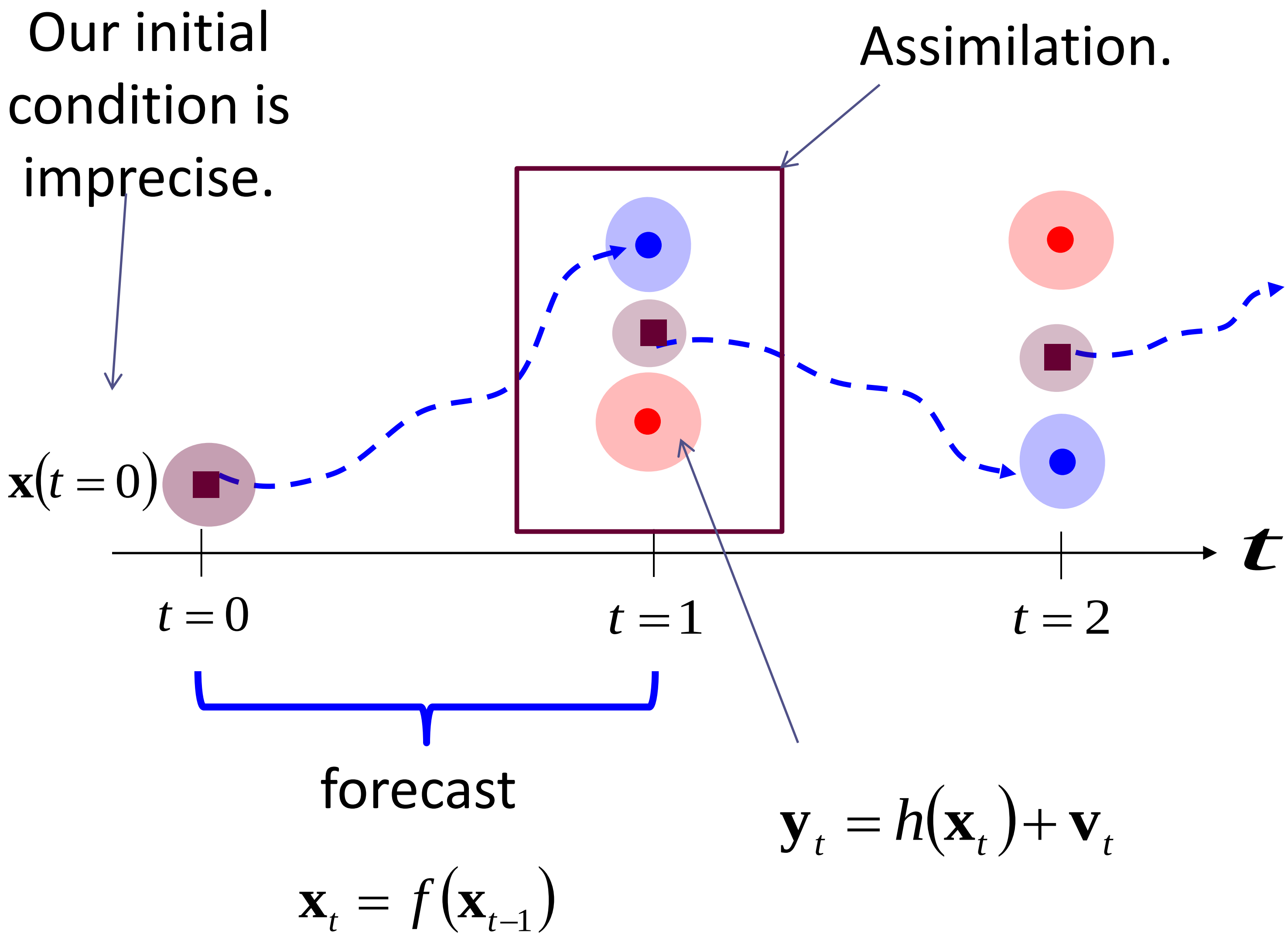
$$\mathbf{y}_t = h(\mathbf{x}_t) + \mathbf{v}_t \quad \mathbf{y} \in \Re^L$$

Observation error: $E(\mathbf{v}_t) = \mathbf{0}$, $Cov(\mathbf{v}_t) = \mathbf{R}$

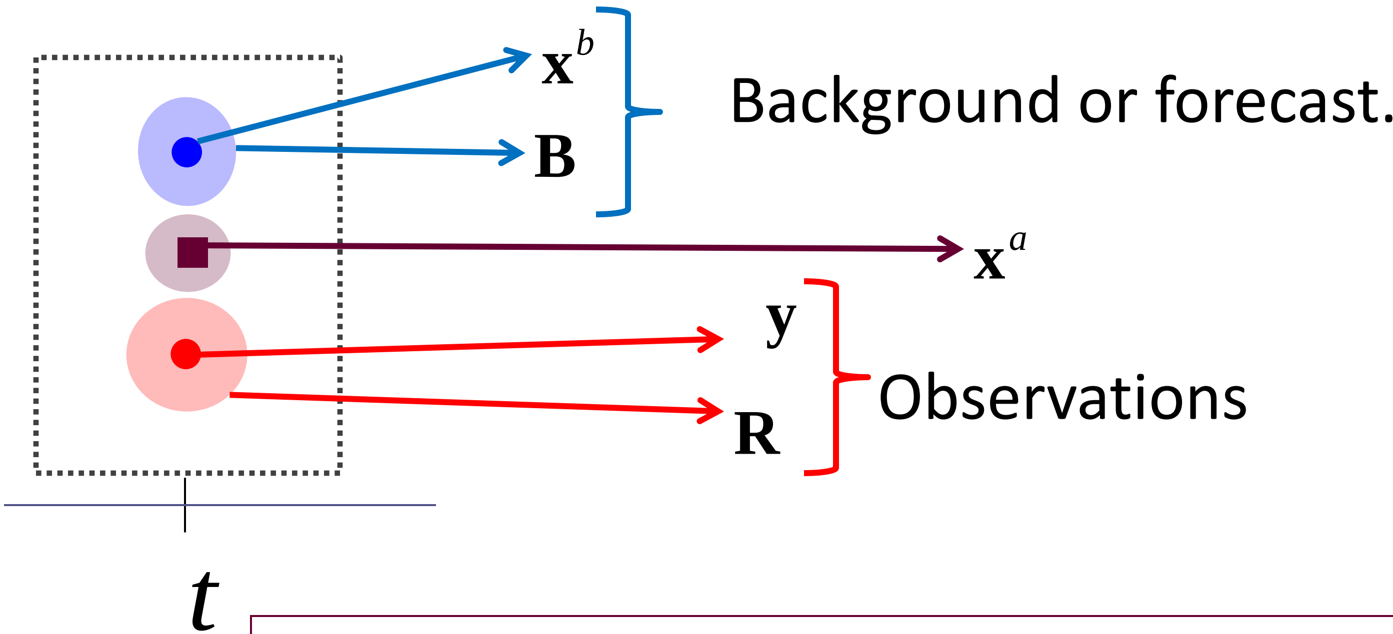
Analysis (‘best’ estimate of the **truth**)



Perfect model (for the moment)



3DVar



$$\mathbf{x}^a = \min_{\mathbf{x}} J(\mathbf{x})$$

$$J(\mathbf{x}) = \frac{1}{2}\left(\mathbf{x}-\mathbf{x}^b\right)^T \mathbf{B}^{-1}\left(\mathbf{x}-\mathbf{x}^b\right) + \frac{1}{2}\left(\mathbf{y}-h(\mathbf{x})\right)^T \mathbf{R}^{-1}\left(\mathbf{y}-h(\mathbf{x})\right)$$

J_b

J_o

This is a multidimensional minimization problem!

Where do the covariances come from?...

Where do the covariances come from?

R is considered known (from the specifications of the instrument). Moreover, it is usually considered diagonal.

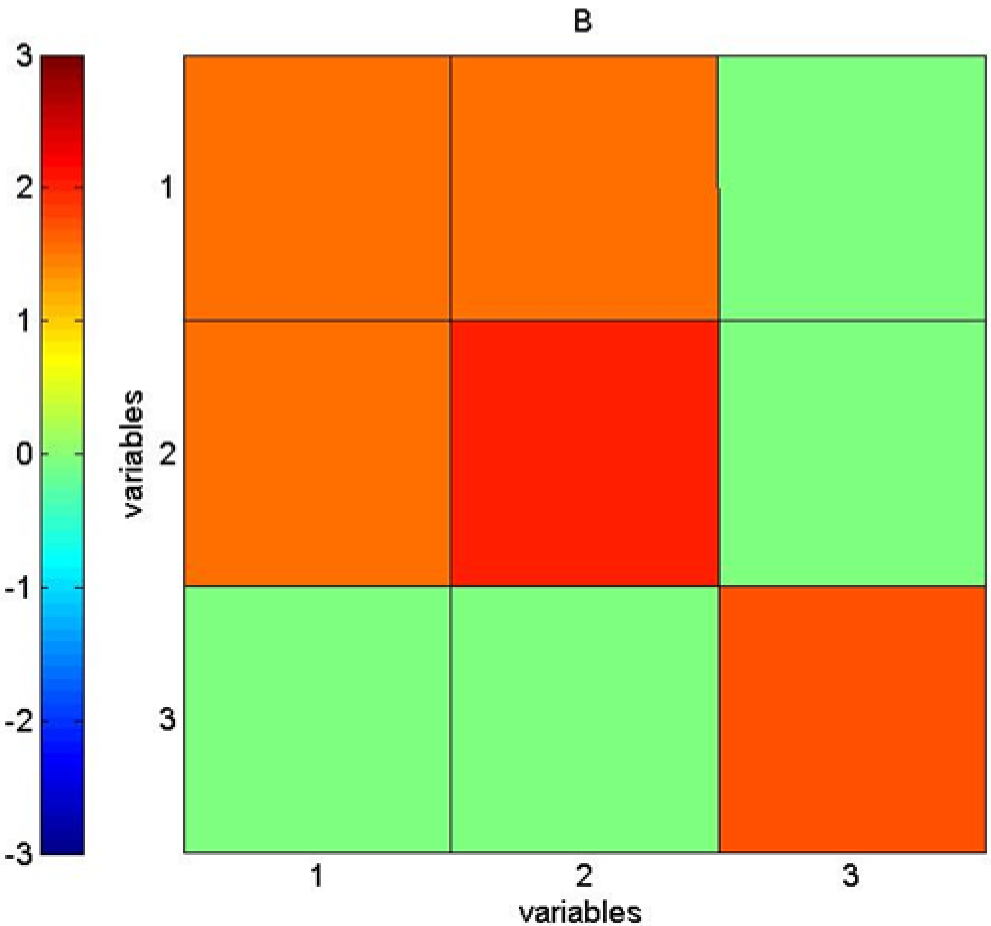
$$\mathbf{R} = \begin{bmatrix} \sigma_1^2 & 0 & 0 & 0 \\ 0 & \sigma_2^2 & 0 & 0 \\ 0 & 0 & \ddots & 0 \\ 0 & 0 & 0 & \sigma_L^2 \end{bmatrix}$$

← This is not particularly true for some observations coming from the same instrument, like in satellites…

B is trickier… It expresses the relationship among variables in the model. **B** has a very important job. It transmits information in the changes of observed variables to unobserved variables! We consider it to be **static**.

The background error covariance

Example for the Lorenz 1963 model. In this case **B** was obtained from comparing the state of the model at different instants for a very long run (and 'tunning' the magnitudes…).



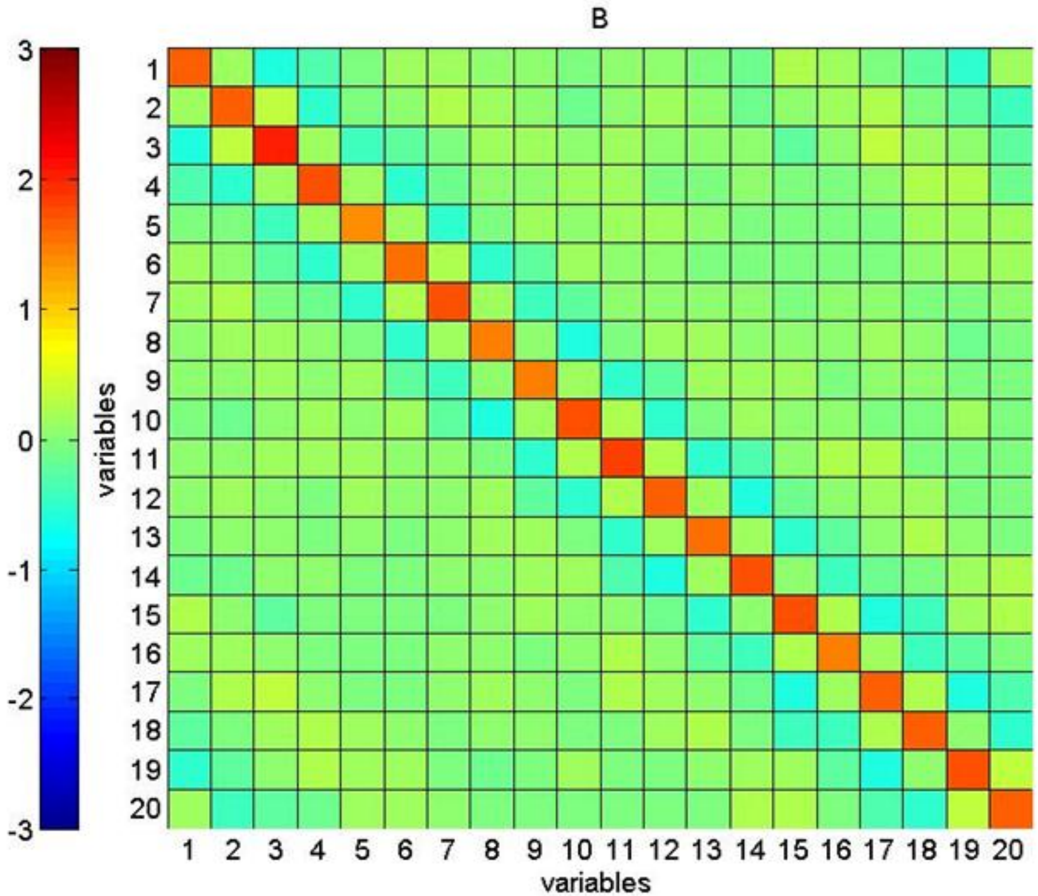
What is this saying?

$x^{(1)}$ and $x^{(2)}$ are strongly correlated: **knowing the state of one gives information about the other**.

On the contrary, $x^{(3)}$ is not correlated to the other 2. **Knowing its value does not yield information on the others.**

Covariances

B for the Lorenz 1996 model with 20 variables. Again, it was obtained by comparing different instants of a long run.

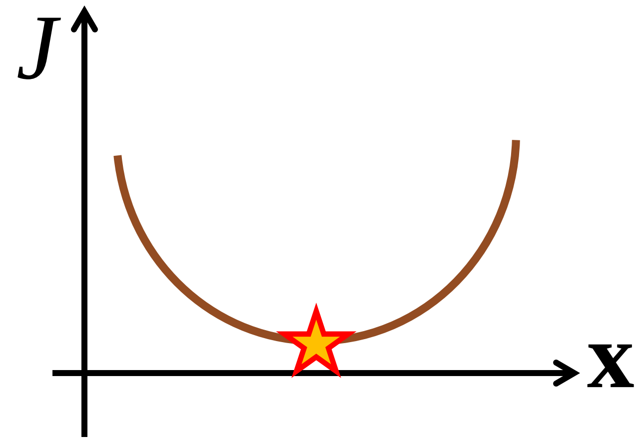


So far, we have not included any error in the forecast. If there were, it would affect **B**.

3DVar: a minimization problem

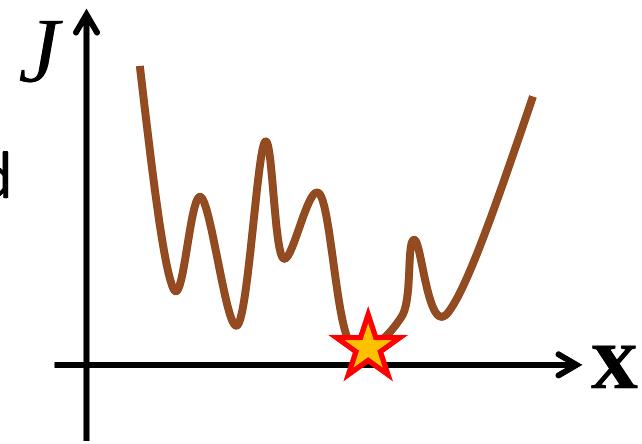
$$J(\mathbf{x}) = \frac{1}{2}\left(\mathbf{x}-\mathbf{x}^b\right)^T \mathbf{B}^{-1}\left(\mathbf{x}-\mathbf{x}^b\right) + \frac{1}{2}\left(\mathbf{y}-\mathbf{H}\mathbf{x}\right)^T \mathbf{R}^{-1}\left(\mathbf{y}-\mathbf{H}\mathbf{x}\right)$$

If the observation operator is linear, there exists a global minimum and no local minima.



$$J(\mathbf{x}) = \frac{1}{2}\left(\mathbf{x}-\mathbf{x}^b\right)^T \mathbf{B}^{-1}\left(\mathbf{x}-\mathbf{x}^b\right) + \frac{1}{2}\left(\mathbf{y}-h(\mathbf{x})\right)^T \mathbf{R}^{-1}\left(\mathbf{y}-h(\mathbf{x})\right)$$

The minimization becomes complicated when the observation operator is nonlinear.



3DVar: a minimization problem

Conditions for the solution:

$$J(\mathbf{x}) = \frac{1}{2}\left(\mathbf{x}-\mathbf{x}^b\right)^T \mathbf{B}^{-1}\left(\mathbf{x}-\mathbf{x}^b\right) + \frac{1}{2}\left(\mathbf{y}-h(\mathbf{x})\right)^T \mathbf{R}^{-1}\left(\mathbf{y}-h(\mathbf{x})\right)$$

$$\nabla J(\mathbf{x}) = 0$$

$$\mathbf{B}^{-1}\left(\mathbf{x}-\mathbf{x}^b\right) - \underline{\mathbf{H}'}^T \mathbf{R}^{-1}\left(\mathbf{y}-h(\mathbf{x})\right) = 0$$

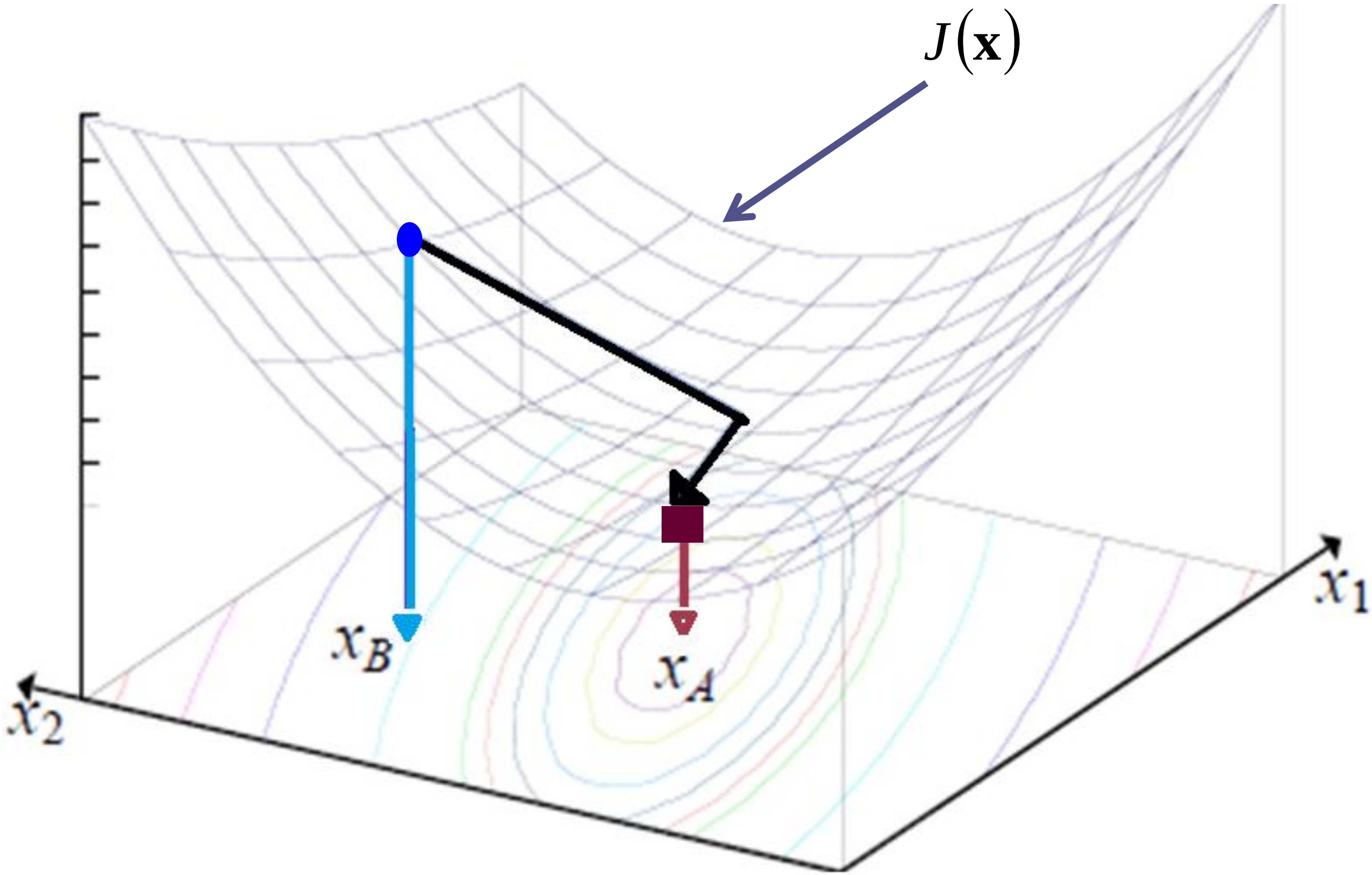
This is the Jacobian of the observation operator

$$\mathbf{H}' = \left.\frac{h(\mathbf{x})}{\partial \mathbf{x}}\right|_{\mathbf{x}_{ref}}$$

$$\nabla J(\mathbf{x}) > 0$$

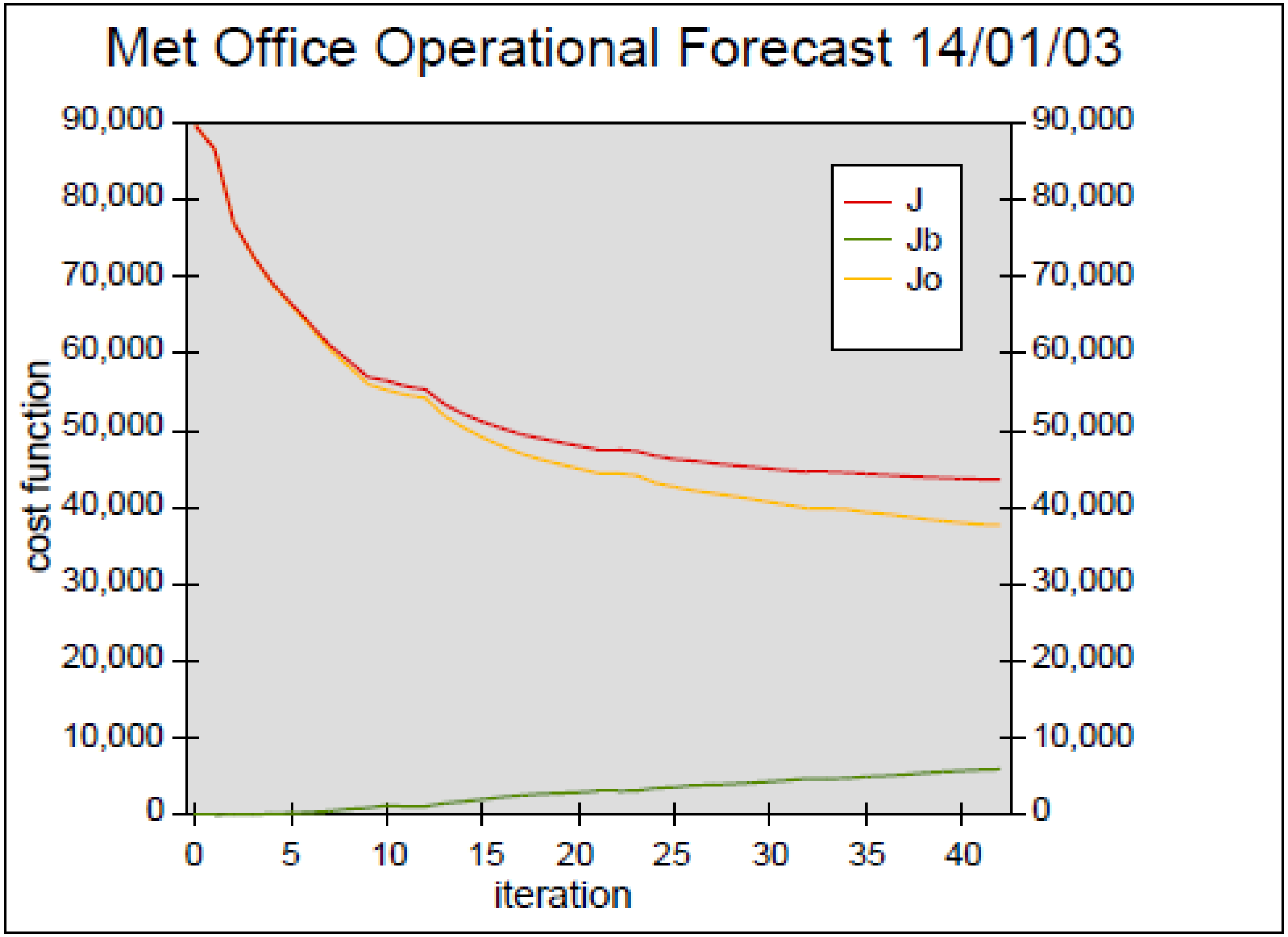
$$\mathbf{B}^{-1} + \mathbf{H}'^T \mathbf{R}^{-1}\mathbf{H}' > 0$$

3DVar: a minimization problem



There are many methods to do this minimization. One must choose efficiency and feasibility.

Minimizing the cost function



Value of the cost function and its components as a function of **iteration** for Met Office © 3DVar.

Preconditioning

Let us take a look on the cost function again.

$$J(\mathbf{x}) = \frac{1}{2}\left(\mathbf{x}-\mathbf{x}^b\right)^T \mathbf{B}^{-1}\left(\mathbf{x}-\mathbf{x}^b\right) + \frac{1}{2}\left(\mathbf{y}-h(\mathbf{x})\right)^T \mathbf{R}^{-1}\left(\mathbf{y}-h(\mathbf{x})\right)$$

B can be badly conditioned. The following change of variable may useful:

$$\mathbf{v} = \mathbf{B}^{-1/2}\left(\mathbf{x}-\mathbf{x}^b\right)$$

Then, the cost function looks like:

$$J(\mathbf{v}) = \frac{1}{2}\mathbf{v}^T\mathbf{v} + \frac{1}{2}\left(\mathbf{y}-h\left(\mathbf{B}^{1/2}\mathbf{v}+\mathbf{x}^b\right)\right)^T \mathbf{R}^{-1}\left(\mathbf{y}-h\left(\mathbf{B}^{1/2}\mathbf{v}+\mathbf{x}^b\right)\right)$$

This is better conditioned.

Preconditioning

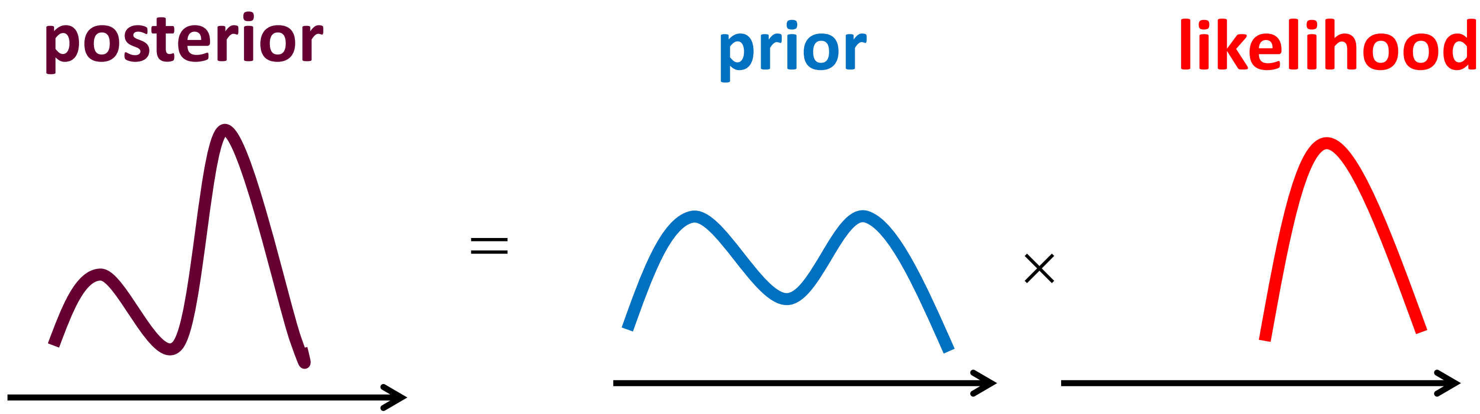
Knowledge of the Hessian of the cost function can be used to preconditioned.

For example, ECMWF uses a Lanczos algorithm to minimize the cost function and save the leading eigenvectors to precondition the next assimilation.

A Bayesian interpretation

We can **combine** the two sources using **Bayes theorem**:

$$p(\mathbf{x} \mid \mathbf{y}) = \frac{1}{A} p(\mathbf{x}) p(\mathbf{y} \mid \mathbf{x})$$



Can be viewed as a **normalization** constant

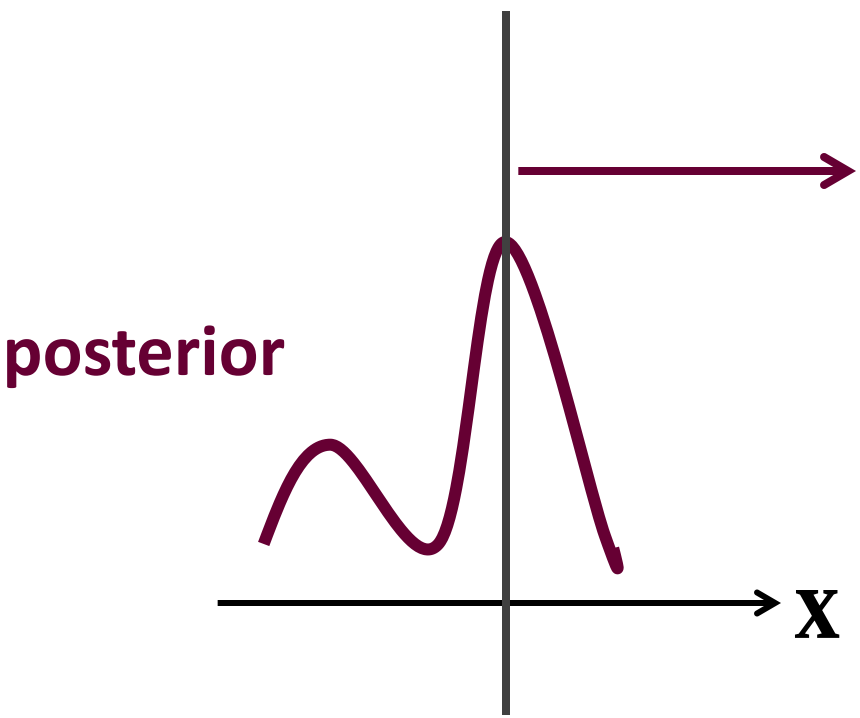
$$A = p(\mathbf{y}^n) = \int p(\mathbf{y}^n \mid \mathbf{x}^n) p(\mathbf{x}^n) d\mathbf{x}^n$$

A Bayesian interpretation

The 3Var solution is a Maximum A Posteriori (MAP) estimator.

$$p(\mathbf{x}\,|\,\mathbf{y}) = \frac{1}{A}\,p(\mathbf{x})\,p(\mathbf{y}\,|\,\mathbf{x})$$

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



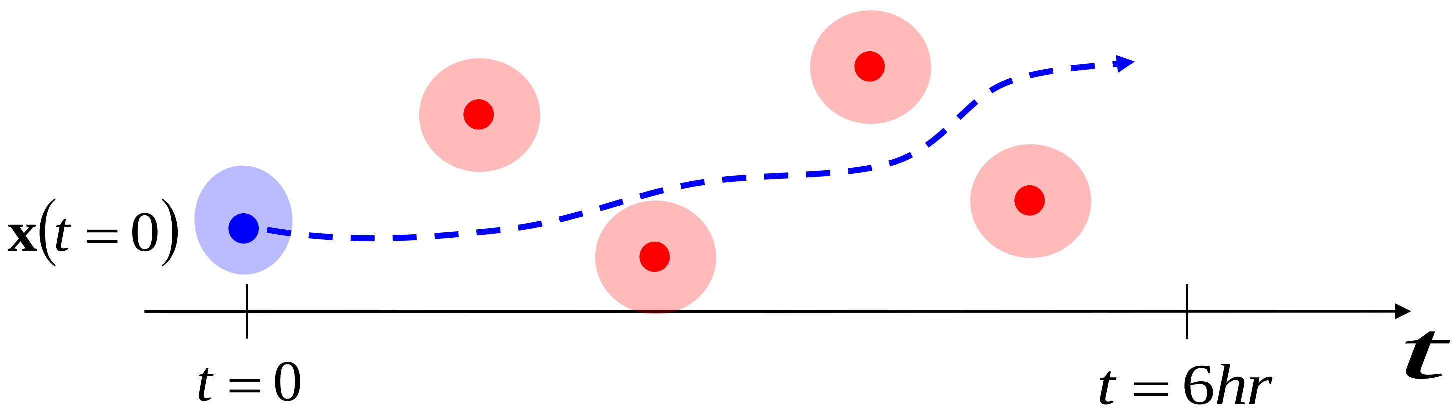
The value that minimizes the cost function corresponds to the mode of the posterior distribution.

Advanced methods:
4DVar

One extra dimension: time

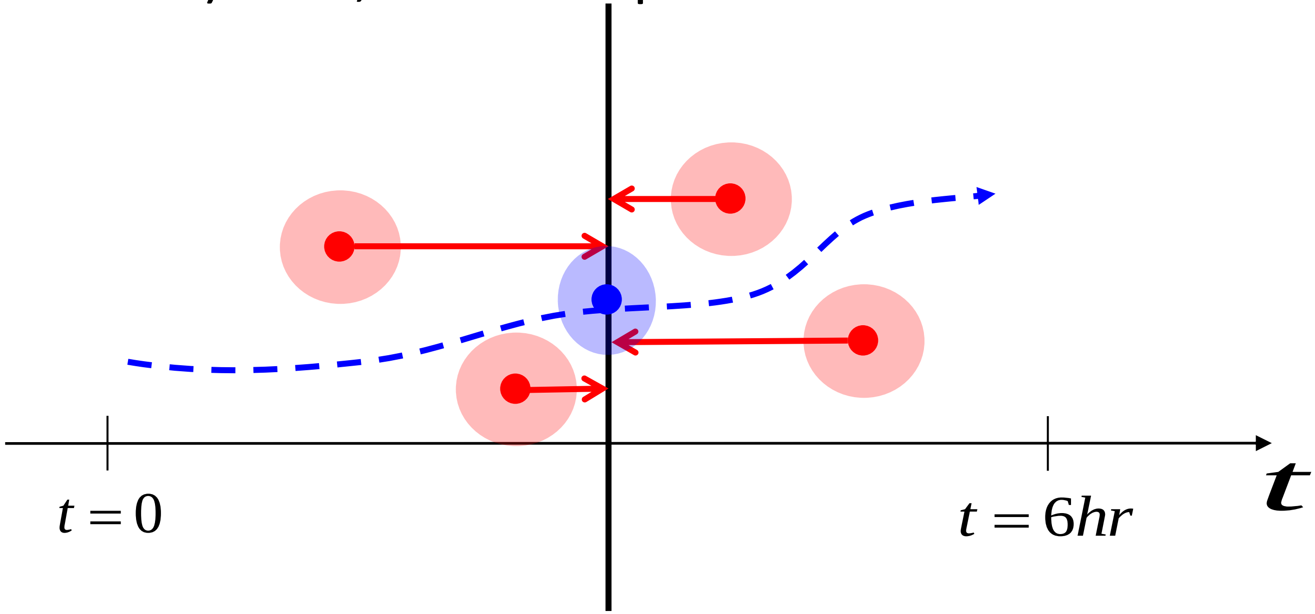
So far, **we have considered the observation and assimilation to occur at the same time**. This is rarely the case, one cannot assimilate every time a new observation arrives.

How do we assimilate **observations distributed in a** forecast **window**?



One extra dimension: time

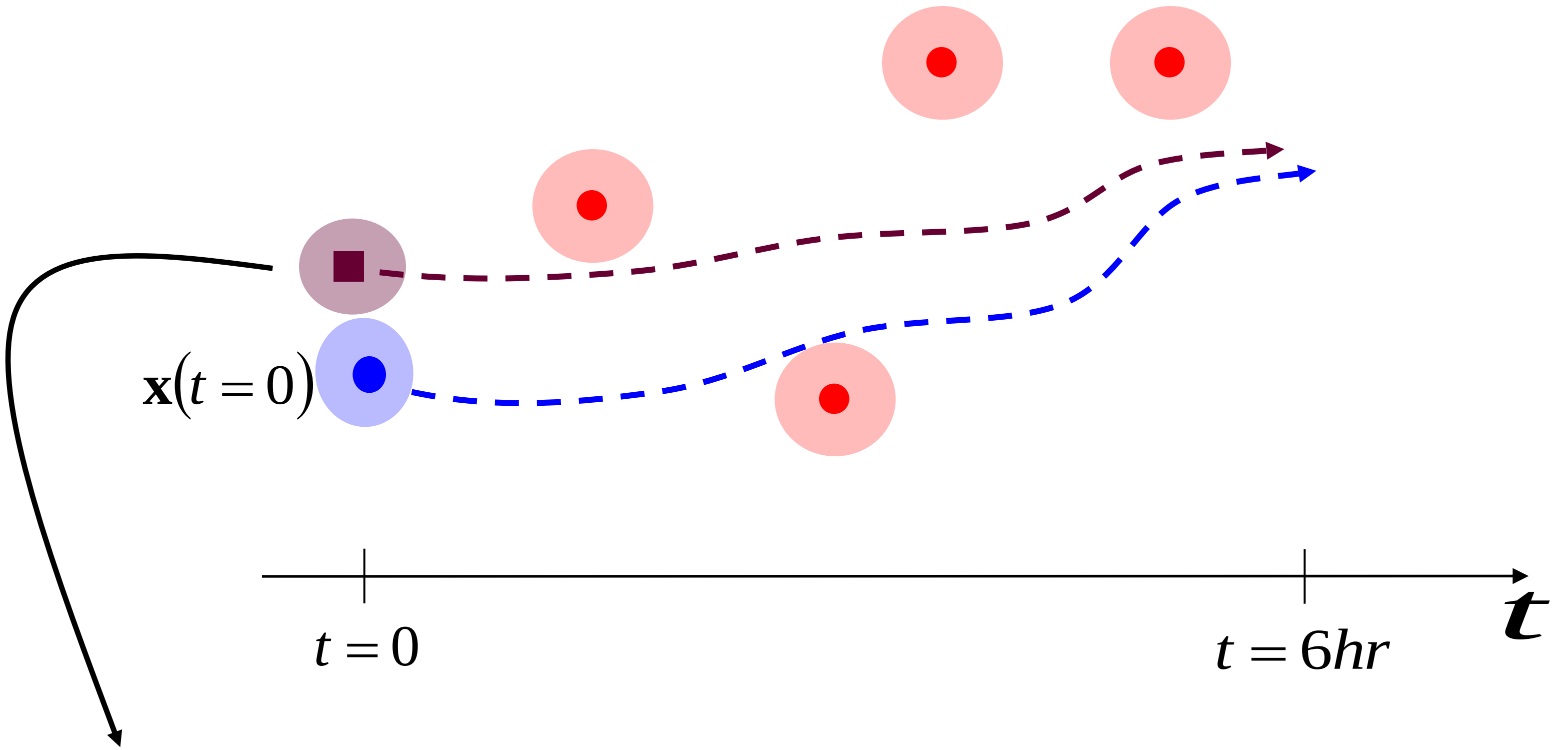
a) Consider the observations to have occurred at the same (intermediate) time, and then perform 3DVar.



b) Find a **trajectory** that fits the observations **throughout the** whole **forecast window**. This is 4DVar: we have added an extra dimension to the minimization. Two flavours: strong constraint and weak constraint.

Strong constraint 4D-Var

In the absence of model error, this reduces to finding the **initial condition** that produces the **trajectory** that **fits observations best** throughout the **whole window**.



Since the model is perfect, the **initial condition determines the whole trajectory**.

Strong constraint 4DVar

Again, the solution is the minimizer of a (more complicated) cost function.

$$\mathbf{x}^a = \min_{\mathbf{x}^0} J\left(\mathbf{x}^0\right)$$

$$J\left(\mathbf{x}^0\right) = \frac{1}{2}\left(\mathbf{x}^0 - \mathbf{x}^b\right)^T \mathbf{B}^{-1}\left(\mathbf{x}^0 - \mathbf{x}^b\right) + \frac{1}{2}\sum_{t_{obs}=1}^{T_{obs}}\left(\mathbf{y}^{t_{obs}} - h^{t_{obs}}\left(\mathbf{x}^0\right)\right)^T \mathbf{R}^{-1}\left(\mathbf{y}^{t_{obs}} - h^{t_{obs}}\left(\mathbf{x}^0\right)\right)$$

J_b

J_o contains the model.

The ‘modified’ observation operator includes the time evolution of the initial condition.

$$h^{t_{obs}}\left(\mathbf{x}^0\right) = h\left(\mathbf{x}\left(t = t_{obs}\right)\right) = h\left(f^{\,t \to t_{obs}}\left(\mathbf{x}^0\right)\right)$$

Strong constraint 4DVar

The conditions for the solution:

$$\nabla J\left(\mathbf{x}^0\right) = 0$$

$$\mathbf{B}^{-1}\left(\mathbf{x}^0 - \mathbf{x}^b\right) + \sum_{t_{obs}=1}^{T_{obs}} \mathbf{F}'^{t_{obs}\,T}\, \mathbf{H}'^T\, \mathbf{R}^{-1}\left(\mathbf{y}^{t_{obs}} - h^{t_{obs}}\left(\mathbf{x}^0\right)\right) = 0$$

$$\mathbf{F}'^{t_{obs}} = \left.\frac{\partial f\left(\mathbf{x}^0\right)}{\partial \mathbf{x}^0}\right|_{\mathbf{x}^{0,c}}$$

$\mathbf{F}'$ is known as the tangent linear model (TLM) and $\mathbf{F}'^T$ is known as the adjoint.

The TLM and the adjoint

The **TLM** and the **adjoint** have very important tasks.

The **TLM** (linearly) evolves perturbations around a control trajectory.

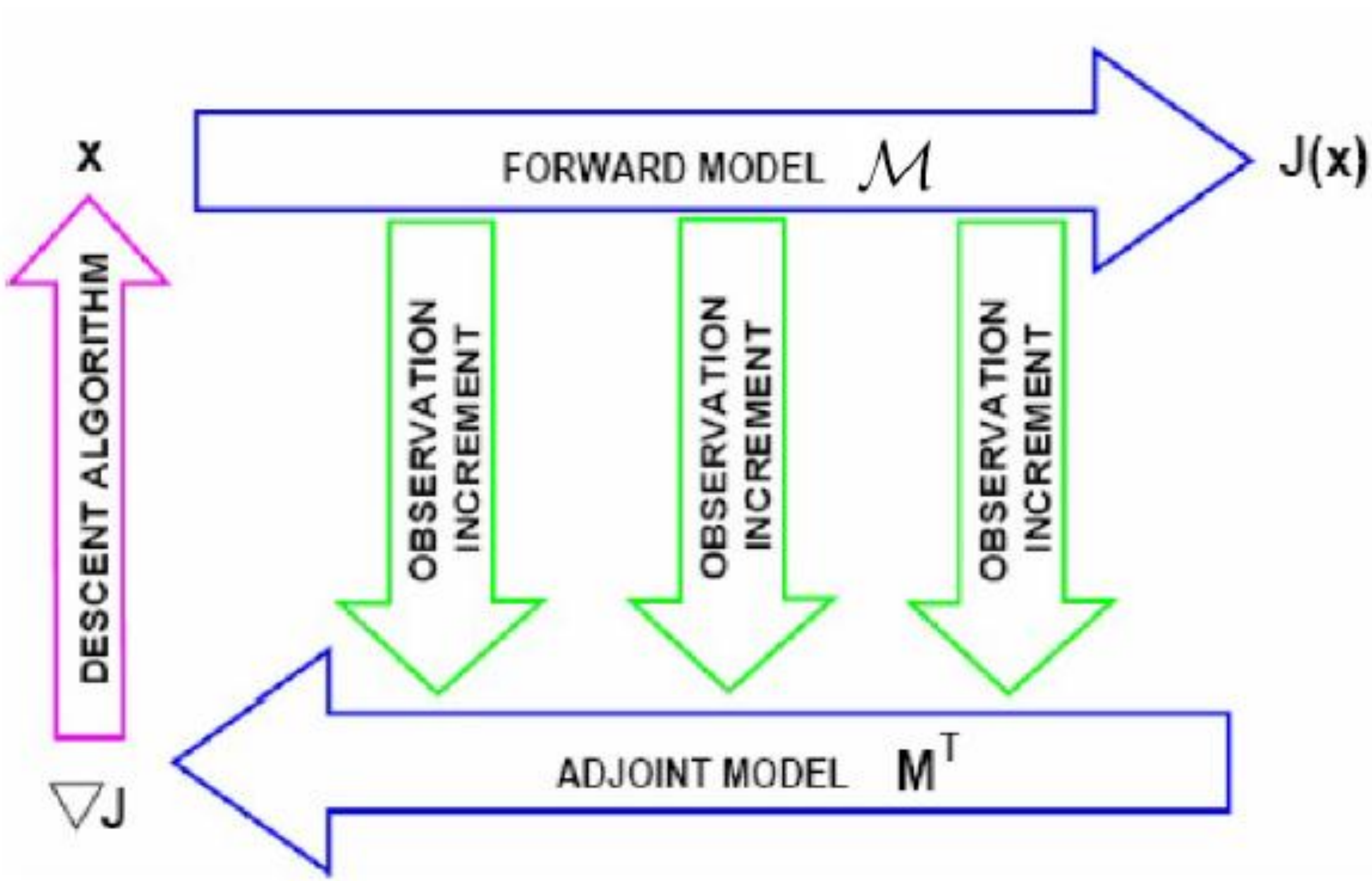
The **adjoint** (linearly) projects the impact of observations to the start of the assimilation window.

$$\sum_{t_{obs}=1}^{T_{obs}} \mathbf{F'}^{t_{obs}T} \, \mathbf{H'}^{T} \, \mathbf{R}^{-1} \left(\mathbf{y}^{t_{obs}} - h^{t_{obs}} \left(\mathbf{x}^{0} \right) \right)$$

They are **approximations**… the stronger the nonlinearity and the longer the assimilation window the less accurate they will become.

The 4DVar cycle

The minimization through time is designed in the following recursive fashion:



Incremental formulation

Let:

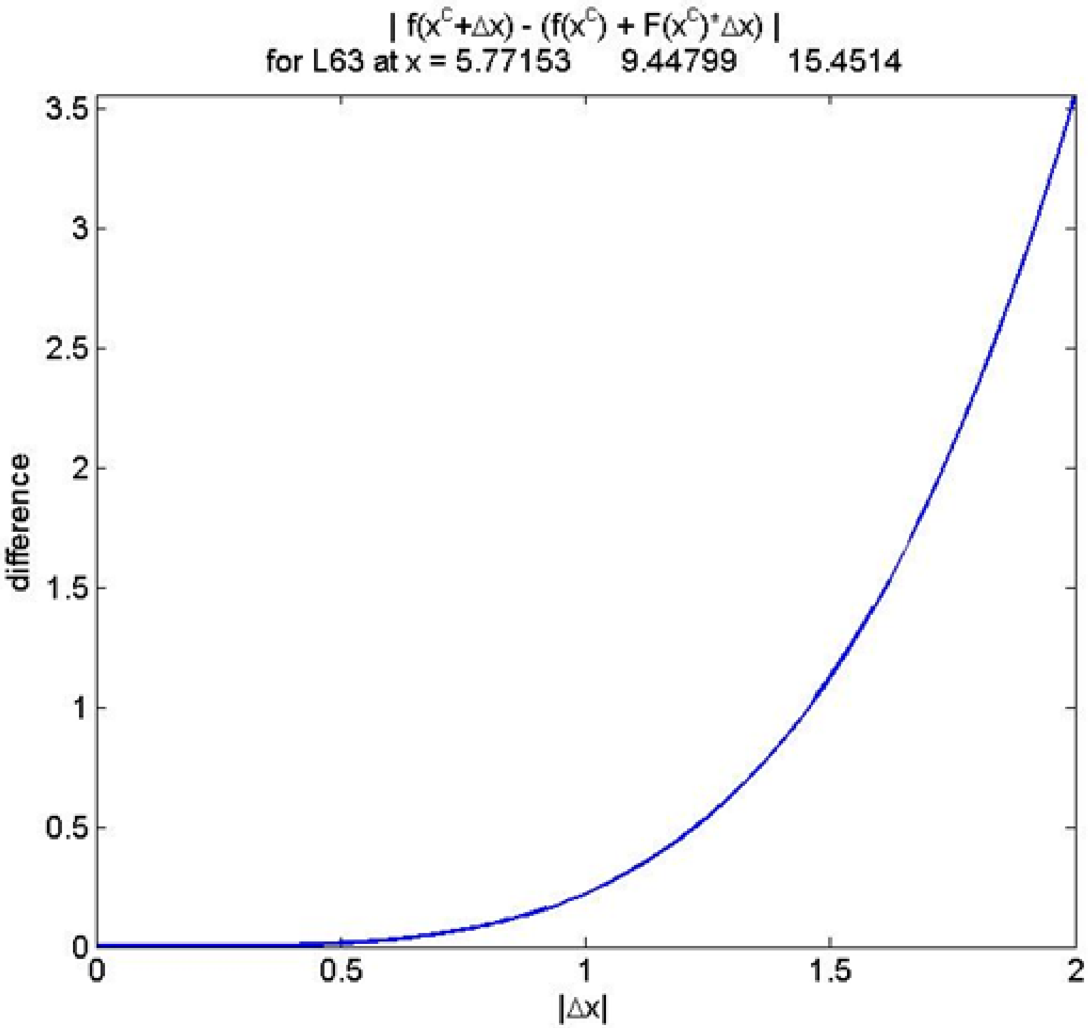
$$\delta \mathbf{x}^0 = \mathbf{x}^0 - \mathbf{x}^b, \qquad \delta \mathbf{y}^{t_{obs}} = \mathbf{y}^{t_{obs}} - h^{t_{obs}}\left(\mathbf{x}^{0,b}\right)$$

$$J = \frac{1}{2}\left(\delta \mathbf{x}^0\right)^T \mathbf{B}^{-1}\left(\delta \mathbf{x}^0\right) + \frac{1}{2}\sum_{t_{obs}=1}^{T_{obs}}\left(\delta \mathbf{y}^{t_{obs}} - \mathbf{HF}^{0\to tobs}\delta \mathbf{x}^0\right)^T \mathbf{R}^{-1}\left(\delta \mathbf{y}^{t_{obs}} - \mathbf{HF}^{0\to tobs}\delta \mathbf{x}^0\right)$$

This works when the nonlinearity is not terribly strong.

Outer loops are iterations of the nonlinear formulation. Inner loops are iterations of the **incremental formulatio**n.

Testing the TLM (static)



Using the Lorenz 63 model, we study how the 1st order truncation of the Taylor loses accuracy.

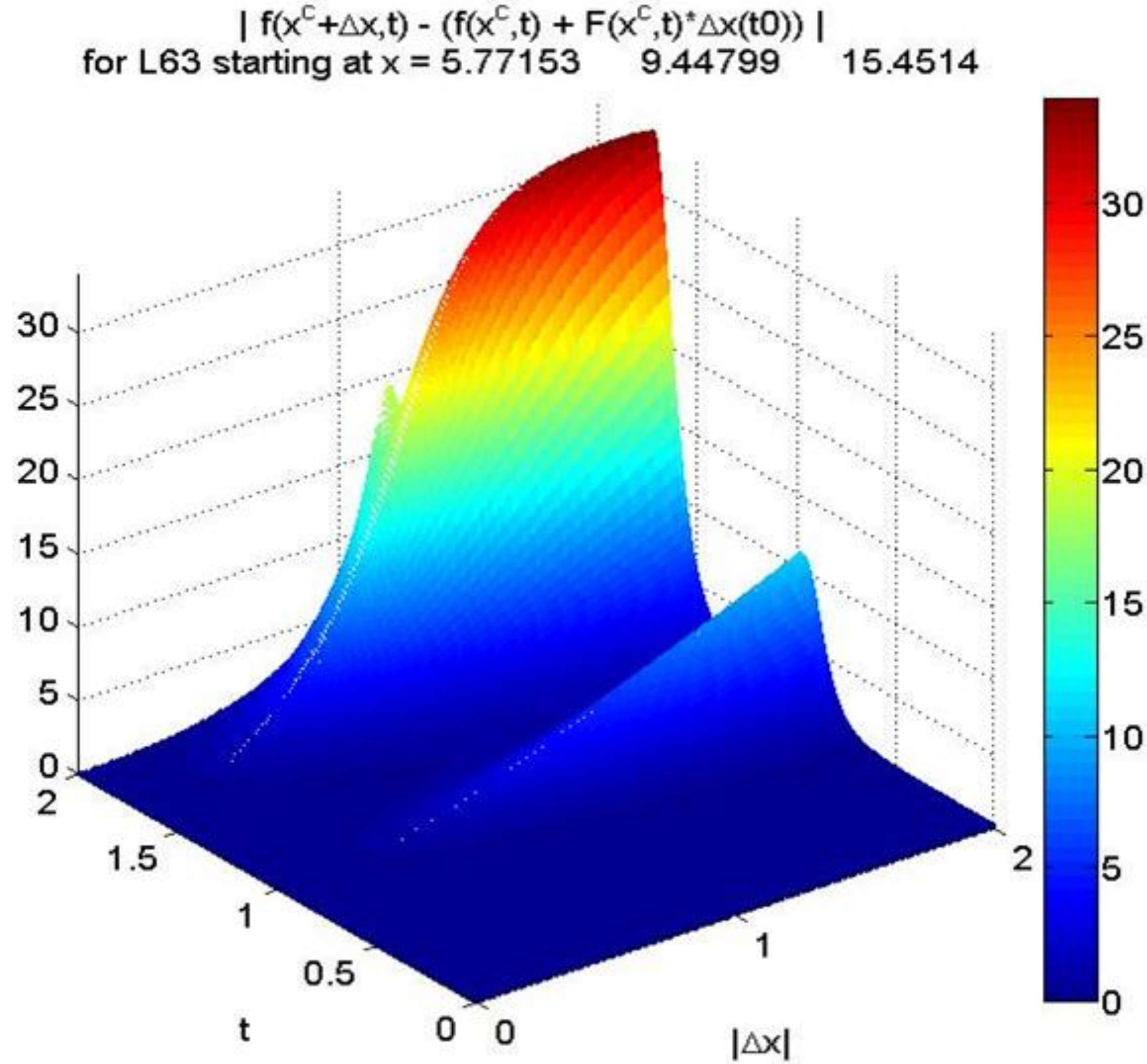
$$f\left(\mathbf{x}^c + \delta\mathbf{x}\right) \approx f\left(\mathbf{x}^c\right) + \mathbf{F}'\big|_{\mathbf{x}^c} \delta\mathbf{x}$$

Testing the TLM (dynamic)

$$f^{0\to t}\left(\mathbf{x}^{0,c}+\delta\mathbf{x}^{0}\right)\approx f^{0\to t}\left(\mathbf{x}^{c}\right)+\mathbf{F'}^{0\to t}\left(\mathbf{x}^{0,c}\right)\delta\mathbf{x}^{0}$$

Evaluating the difference between the exact time evolution of a trajectory, with respect to the nonlinear evolution of the control and the linear evolution of the perturbations using the TLM.

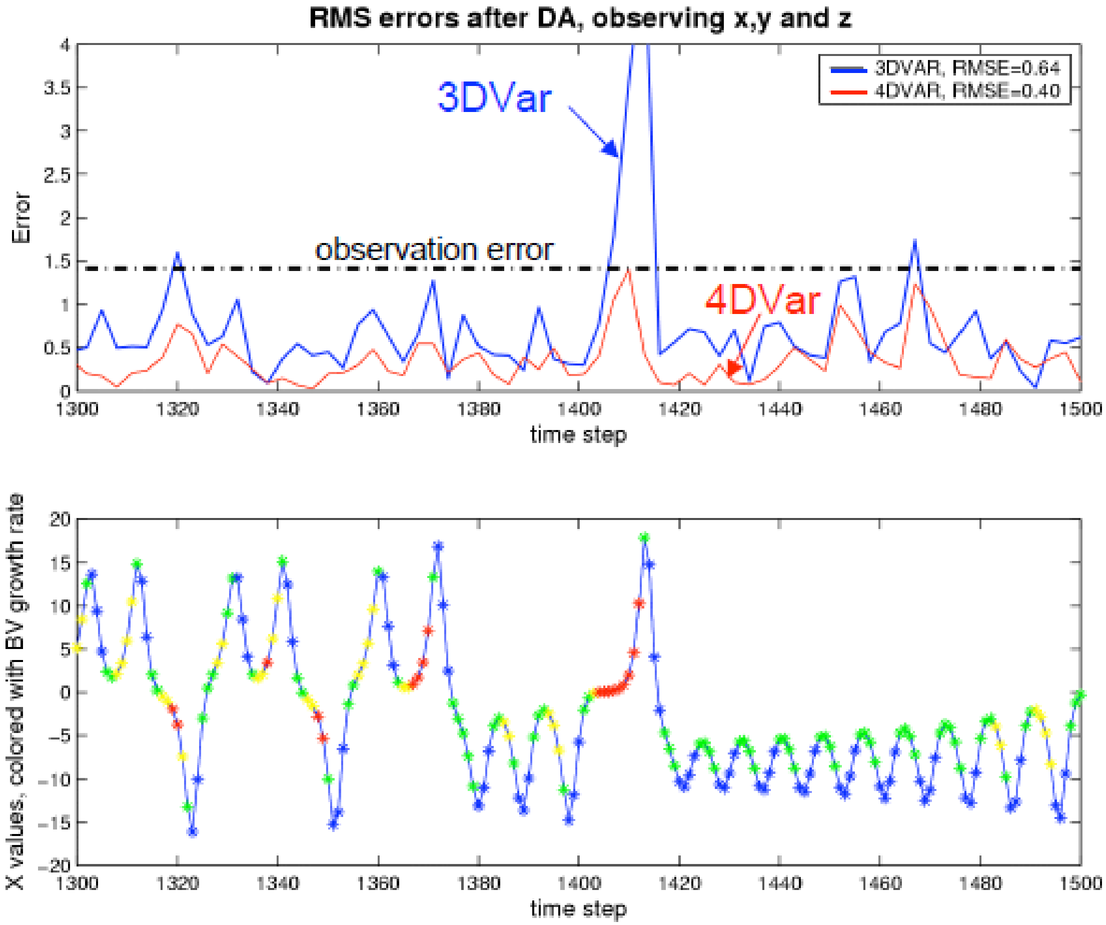
Again, Lorenz 1963 is used.



3D vs 4DVar

4DVar has important information from the future (after all, it is a smoother), 3DVar does not.

The figure shows a comparison of the performance of the two methods. Taken from Evans et al, 2005.



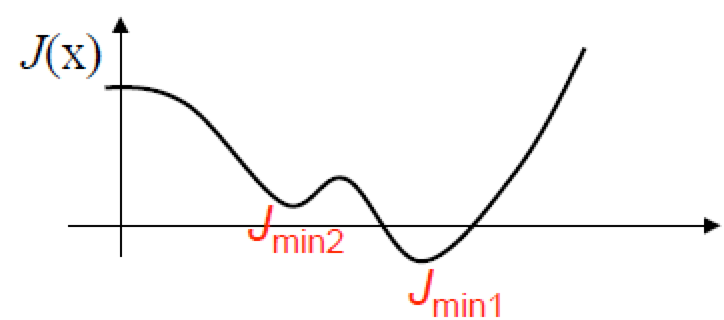
DA cycle and observations: $8\Delta t$, $\mathbf{R}=2^*\mathbf{I}$
4D-Var assimilation window: $24\Delta t$

How long should the assimilation window be?

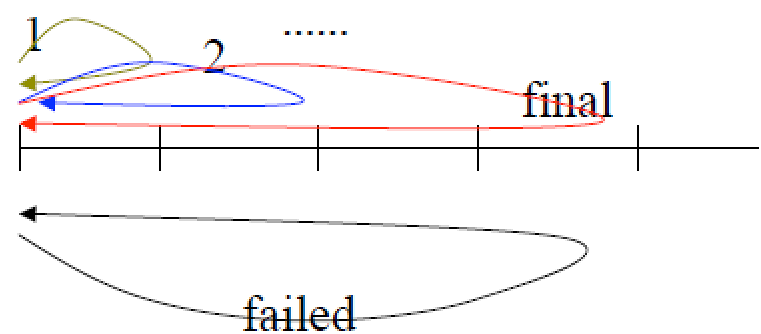
The longer the 4D assimilation window the more observations we'll have… but also the more nonlinear the forecast will be. The best should be somewhere in the middle.

	Win=8	16	24	32	40	**48**	56	64	72
Fixed window	0.59	0.59	0.47	0.43	0.62	**0.95**	0.96	0.91	0.98
Start with short window	0.59	0.51	0.47	0.43	0.42	**0.39**	0.44	0.38	0.43

Performance of 4DVar using the Lorenz 1963 and different lengths of assimilation window (Kalnay *et al.*, 2007).



It is recommendable to do the minimization progressively while increasing the assimilation window (Pires *et al.*, 1996).



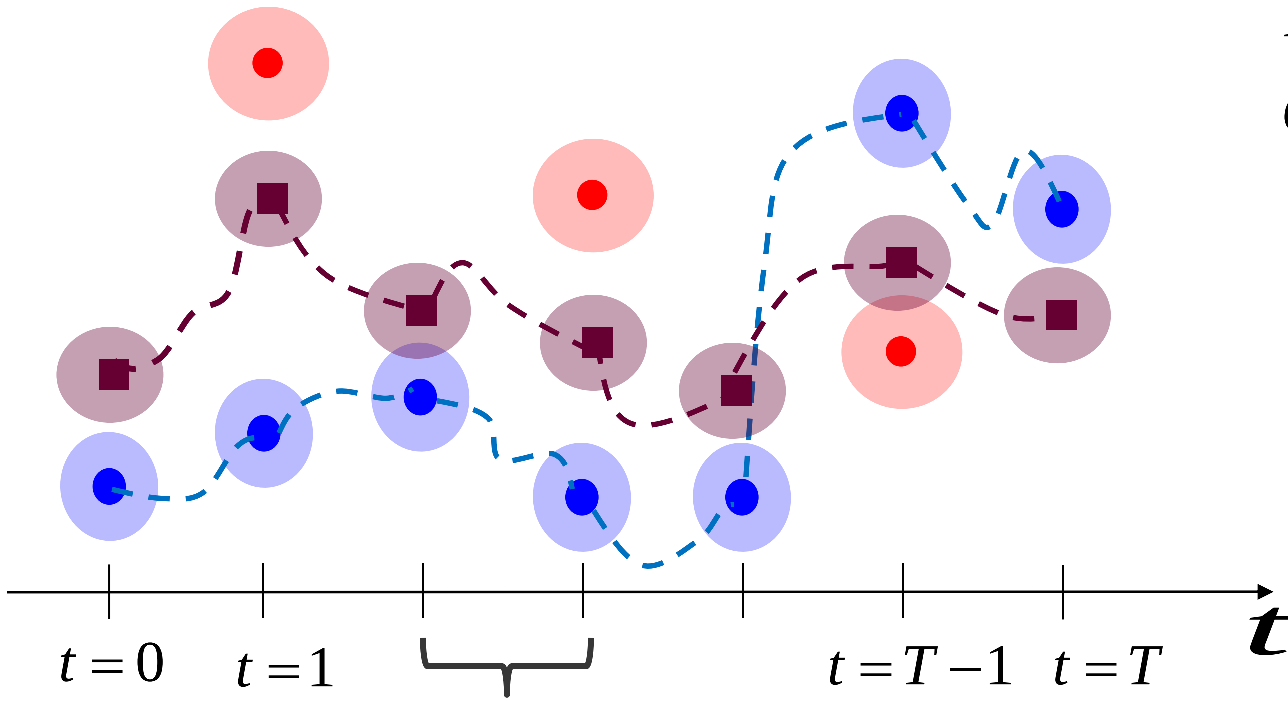
Weak constraint 4DVar

What if the model is not perfect?

$$\mathbf{x}_t = f(\mathbf{x}_{t-1}) + \mathbf{w}_t$$

$$E[\mathbf{w}_t] = \mathbf{0}$$

$$Cov[\mathbf{w}_t] = \mathbf{Q}$$



The forecast contains a random component.

In this case it is not enough to find an optimal **initial condition**, since it does not uniquely determine the trajectory. We have **more control variables** (the **random jumps** from a time step to the next).

Weak constraint 4DVar

The minimization has an extra term.

$$\mathbf{x}_t = f\left(\mathbf{x}_{t-1}\right) + \mathbf{w}_t$$

$$\mathbf{x}^{a,0:T} = \min_{\mathbf{x}^{0:T}} J\left(\mathbf{x}^{0:T}\right)$$

$$\left\{\mathbf{x}^{a,0}, \mathbf{w}^{a,1:T}\right\} = \min_{\mathbf{x}^{0}, \mathbf{w}^{1:T}} J\left(\mathbf{x}^{0}, \mathbf{w}^{1:T}\right)$$

$$J\left(\mathbf{x}^{0:T}\right) = \frac{1}{2}\left(\mathbf{x}^0 - \mathbf{x}^b\right)^T \mathbf{B}^{-1}\left(\mathbf{x}^0 - \mathbf{x}^b\right) \longrightarrow J_b$$

$$+ \frac{1}{2}\sum_{t_{obs}=1}^{T_{obs}}\left(\mathbf{y} - h\left(\mathbf{x}^{t_{obs}}\right)\right)^T \mathbf{R}^{-1}\left(\mathbf{y} - h\left(\mathbf{x}^{t_{obs}}\right)\right) \longrightarrow J_o$$

$$+ \frac{1}{2}\sum_{t=1}^{T}\left(\mathbf{x}^t - f^{\,t-1\to t}\left(\mathbf{x}^{t-1}\right)\right)^T \mathbf{Q}^{-1}\left(\mathbf{x}^t - f^{\,t-1\to t}\left(\mathbf{x}^{t-1}\right)\right) \longrightarrow J_m$$

Some final comments

- What is the best way to compute **B**? Also, how much are we losing by not evolving this background covariance? Especially important in 3DVar.

- Preconditioning and transformations to the formulation in order to make the minimization problem more feasible.

- One of the main difficulties in 4DVar is to compute the TLM/Adjoint for different models. It is hard work…