

Data Assimilation

Alan O'Neill

Data Assimilation Research Centre

University of Reading

Contents

- Motivation
- Univariate (scalar) data assimilation
- Multivariate (vector) data assimilation
 - Optimal Interpolation (*BLUE*)
 - 3d-Variational Method
 - Kalman Filter
 - 4d-Variational Method
- Applications of data assimilation in earth system science

Motivation

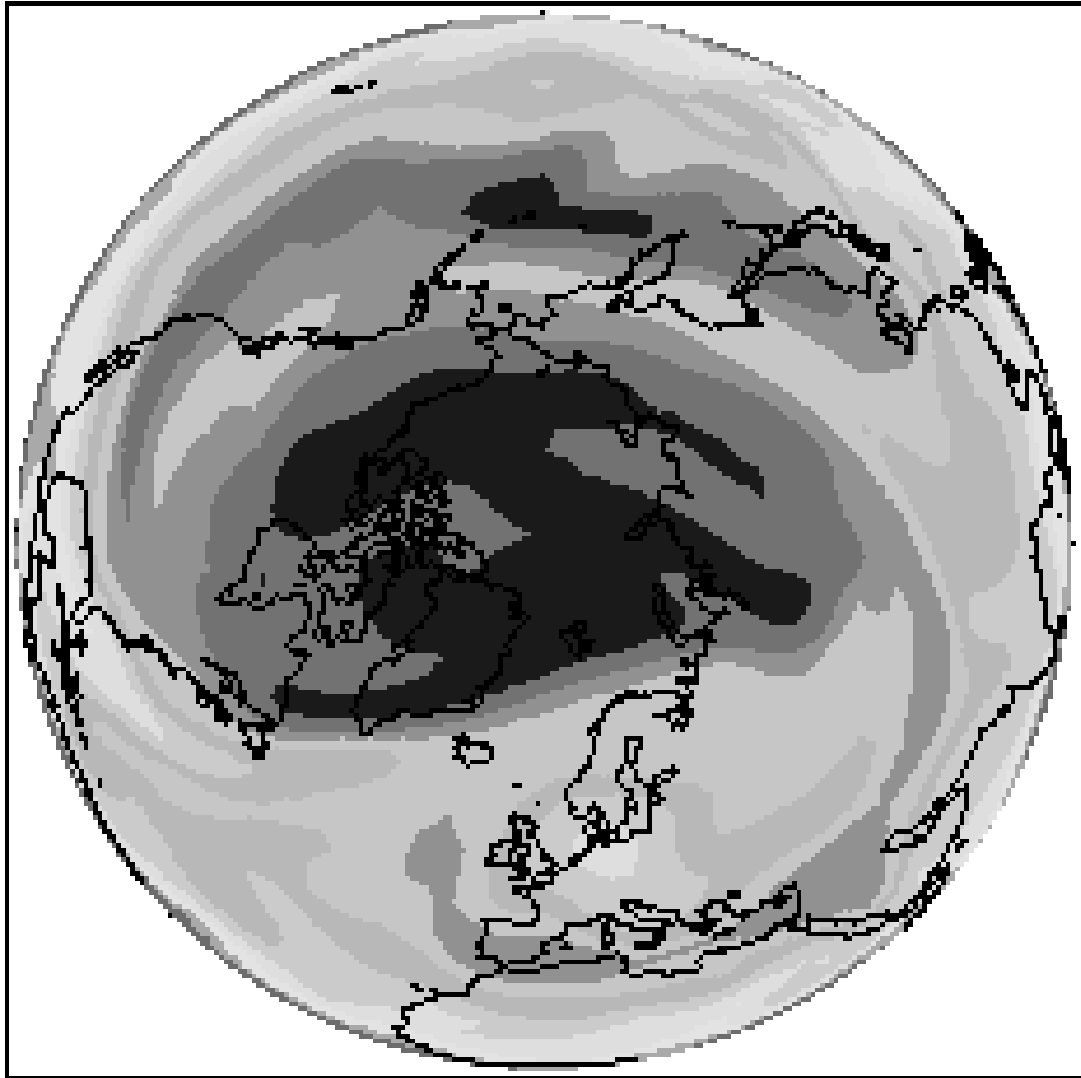
What is data assimilation?

Data assimilation is the technique whereby observational data are combined with output from a numerical model to produce an optimal estimate of the evolving state of the system.


Why We Need Data Assimilation



- **range of observations**
- **range of techniques**
- **different errors**
- **data gaps**
- **quantities not measured**
- **quantities linked**



Some Uses of Data Assimilation

- Operational weather and ocean forecasting
 - Seasonal weather forecasting
 - Land-surface process
 - Global climate datasets
 - Planning satellite measurements
 - Evaluation of models and observations
- 

Preliminary Concepts

What We Want To Know

$\mathbf{x}(t)$ **atmos. state vector**

$\mathbf{s}(t)$ **surface fluxes**

\mathbf{c} **model parameters**

$$\mathbf{X}(t) = (\mathbf{x}(t), \mathbf{s}(t), \mathbf{c})$$

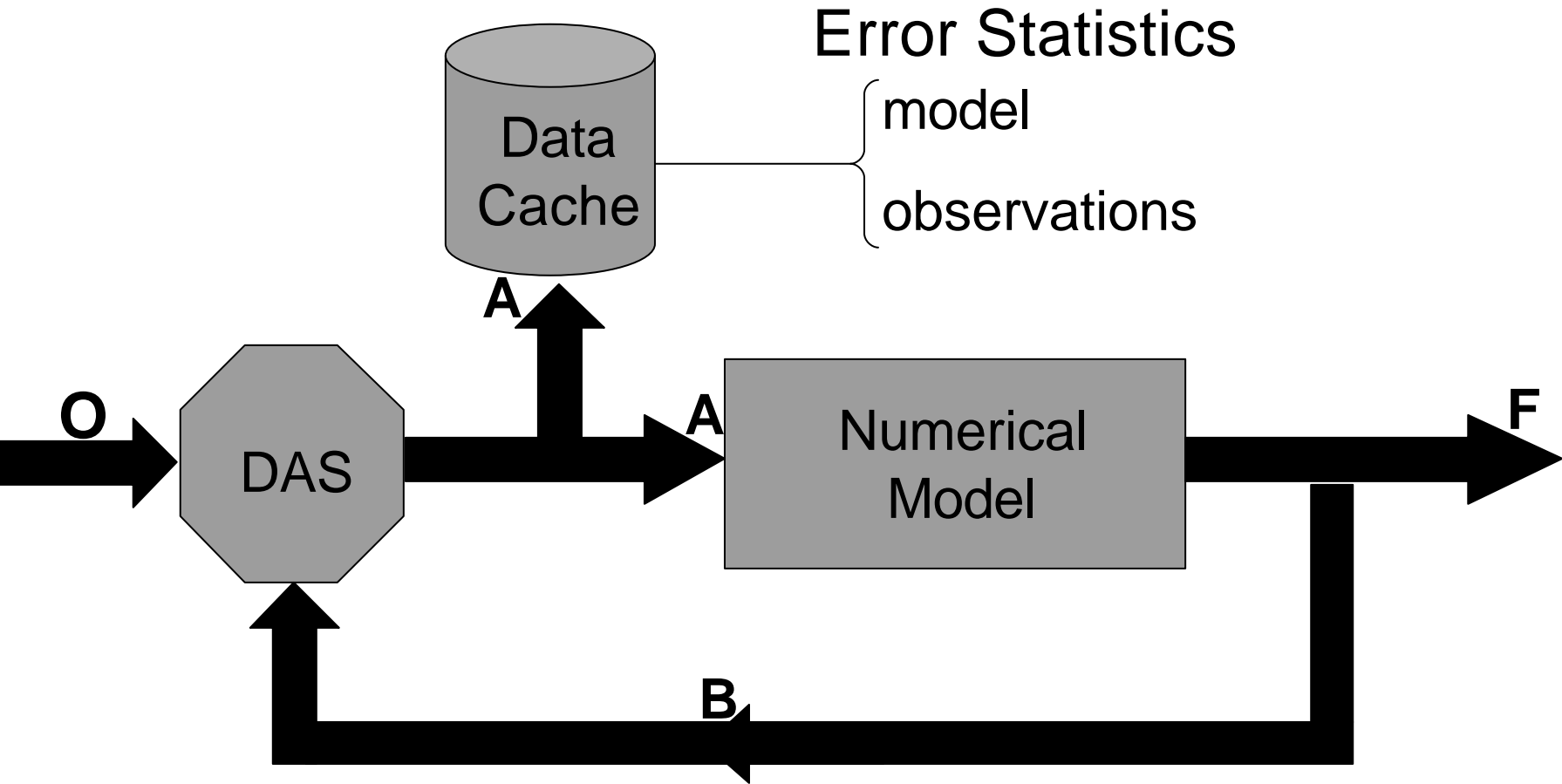
What We Also Want To Know

Errors in models

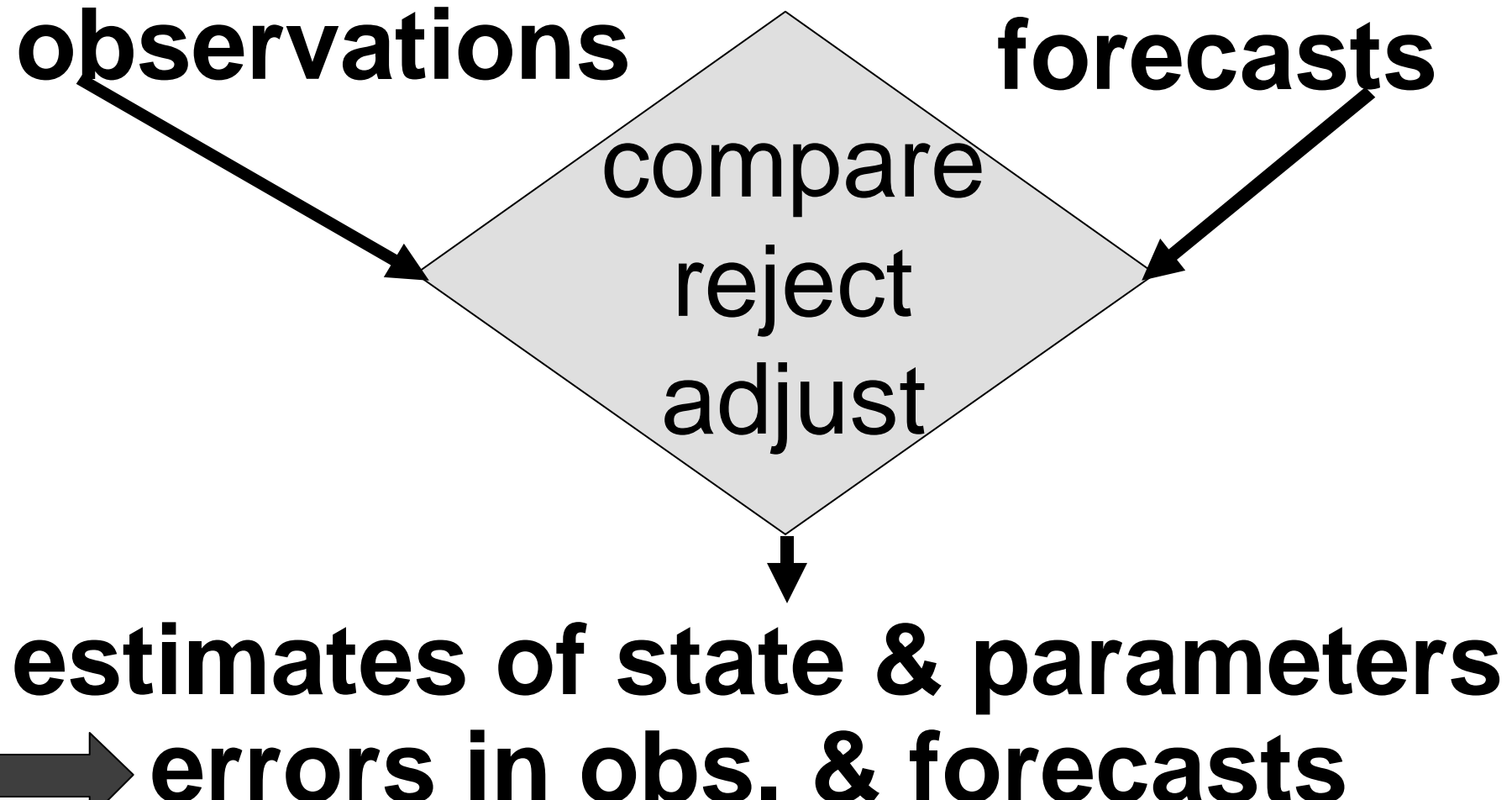
Errors in observations

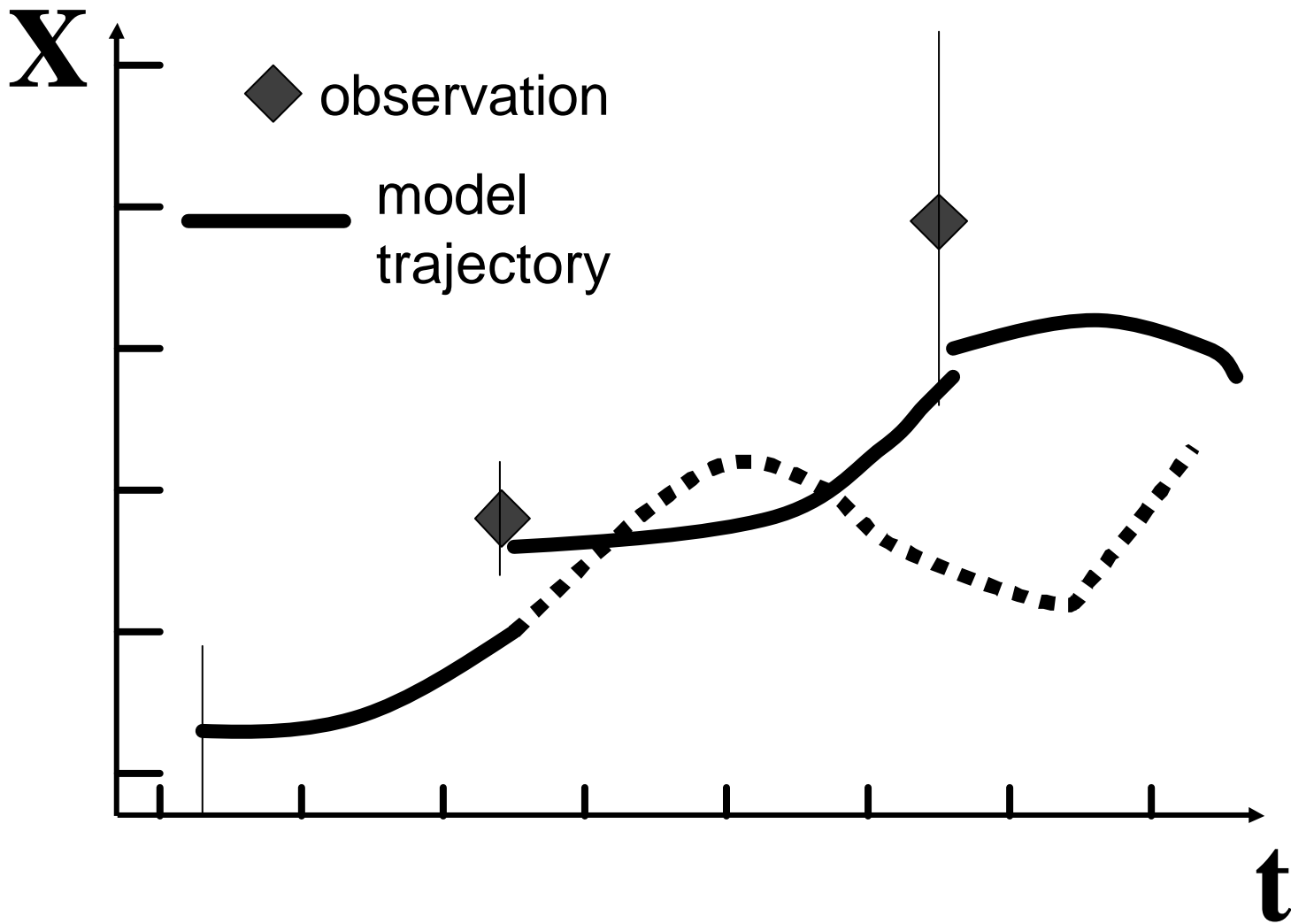
What observations to make

DATA ASSIMILATION SYSTEM



The Data Assimilation Process






Data Assimilation: an analogy

Driving with your eyes closed:
open eyes every 10 seconds and
correct trajectory

Basic Concept of Data Assimilation

- Information is accumulated in time into the model state and propagated to all variables.

What are the benefits of data assimilation?

- Quality control
 - Combination of data
 - Errors in data and in model
 - Filling in data poor regions
 - Designing observational systems
 - Maintaining consistency
 - Estimating unobserved quantities
- 

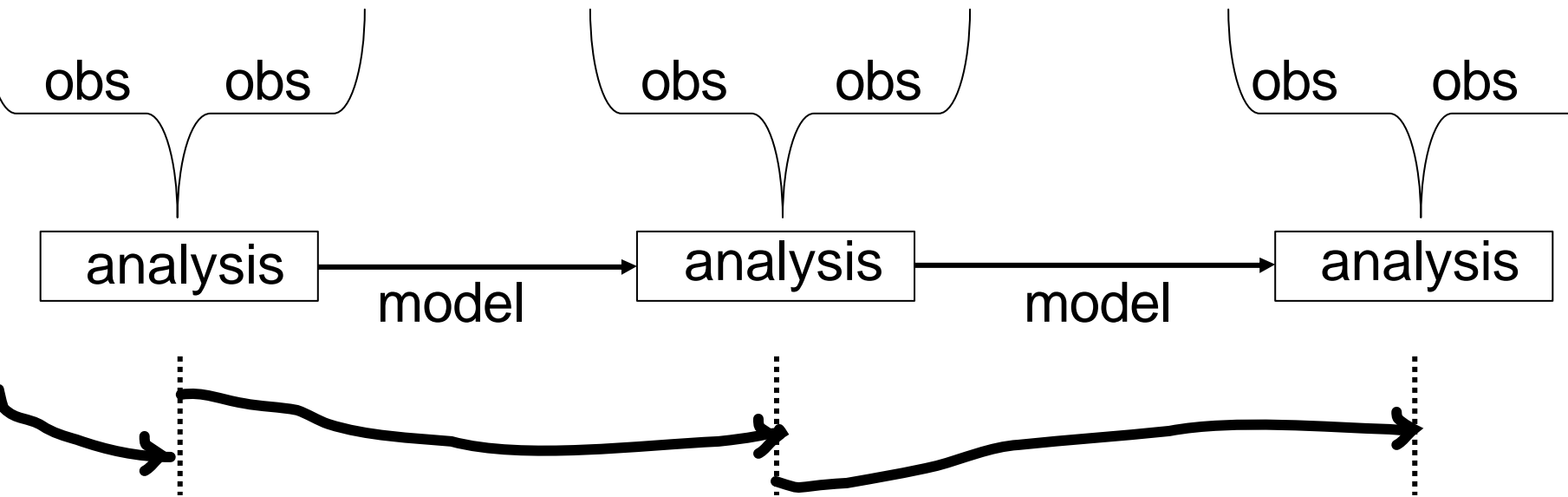
Methods of Data Assimilation

- Optimal interpolation (or approx. to it)
- 3D variational method (3DVar)
- 4D variational method (4DVar)
- Kalman filter (with approximations)

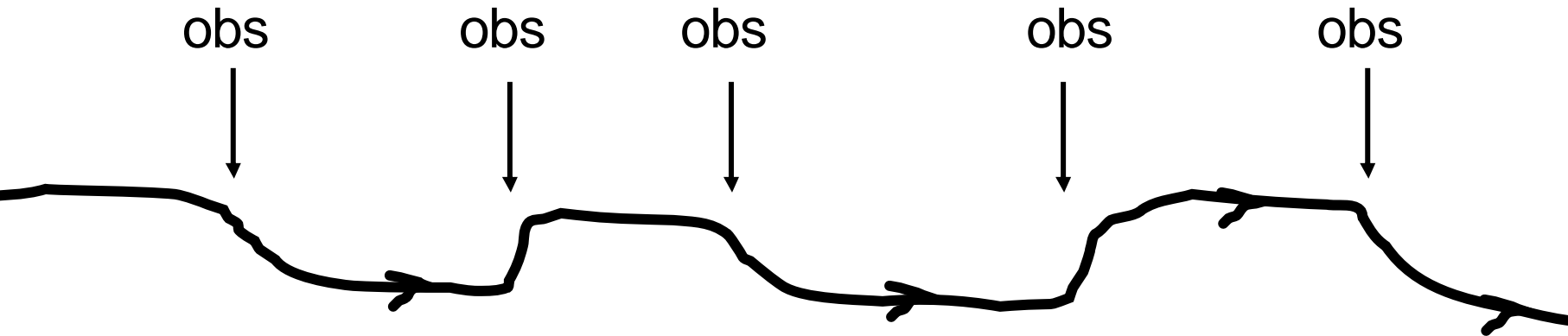
Types of Data Assimilation

- Sequential
- Non-sequential
- Intermittent
- Continuous

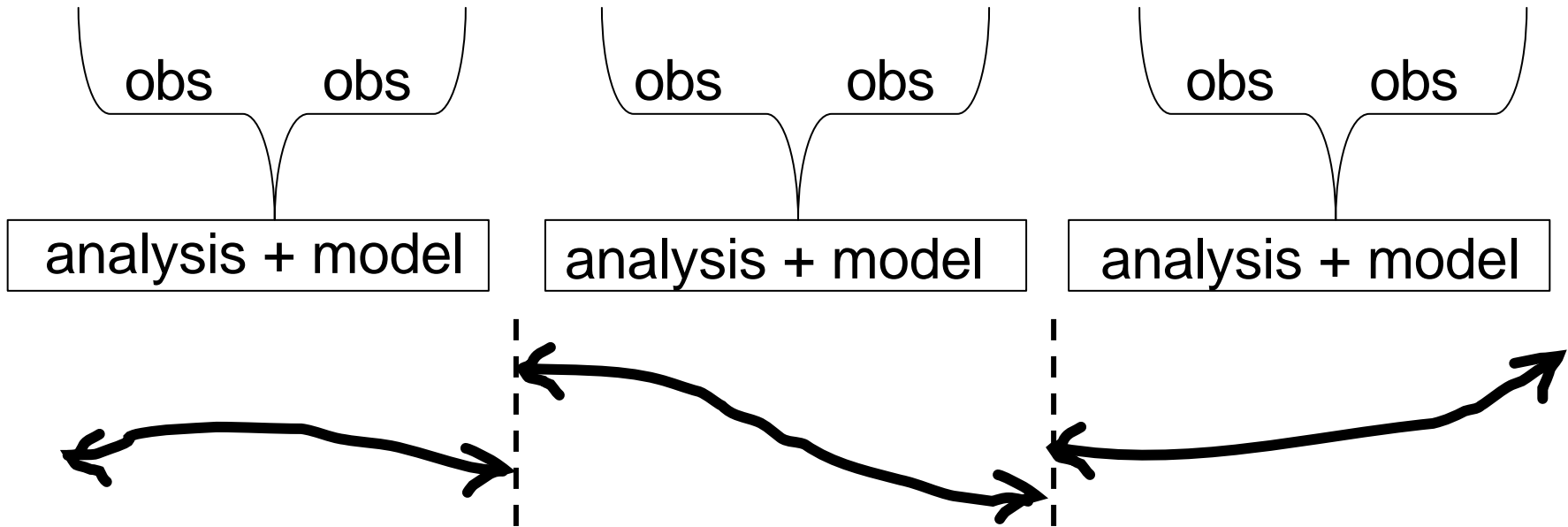
Sequential Intermittent Assimilation



Sequential Continuous Assimilation



Non-sequential Intermittent Assimilation



Non-sequential Continuous Assimilation

obs

obs

obs

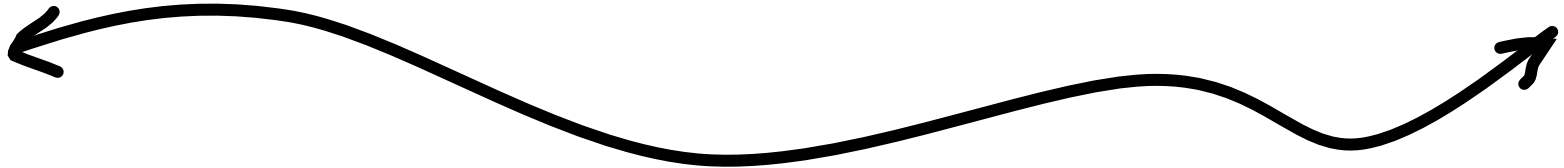
obs

obs

obs



analysis + model



Statistical Approach to Data Assimilation



Data Assimilation Made Simple (scalar case)

Least Squares Method (Minimum Variance)

$$T_1 = T_t + \mathbf{e}_1$$

$$T_2 = T_t + \mathbf{e}_2$$

$$\langle \mathbf{e}_1 \rangle = \langle \mathbf{e}_2 \rangle = 0$$

$$\langle (\mathbf{e}_1)^2 \rangle = \mathbf{s}_1^2$$

$$\langle (\mathbf{e}_2)^2 \rangle = \mathbf{s}_2^2$$

$\langle \mathbf{e}_1 \mathbf{e}_2 \rangle = 0$, the two measurements are

uncorrelated

Estimate T_t as a linear combination
of the observations :

$$T_a = a_1 T_1 + a_2 T_2$$

The analysis should be unbiased : $\langle T_a \rangle = T_t$

$$\Rightarrow a_1 + a_2 = 1$$

Least Squares Method Continued

Estimate T_a by minimizing its mean squared error :

$$\begin{aligned}\mathbf{s}_a^2 &= \langle (T_a - T_t)^2 \rangle = \langle (a_1(T_1 - T_t) + a_2(T_2 - T_t))^2 \rangle \\ &= \langle (a_1\mathbf{e}_1 + a_2\mathbf{e}_2)^2 \rangle = a_1^2\mathbf{s}_1^2 + a_2^2\mathbf{s}_2^2\end{aligned}$$

subject to the constraint $a_1 + a_2 = 1$

Least Squares Method

Continued

$$a_1 = \frac{\frac{1}{s_1^2}}{\frac{1}{s_1^2} + \frac{1}{s_2^2}}$$

$$a_2 = \frac{\frac{1}{s_2^2}}{\frac{1}{s_1^2} + \frac{1}{s_2^2}}$$

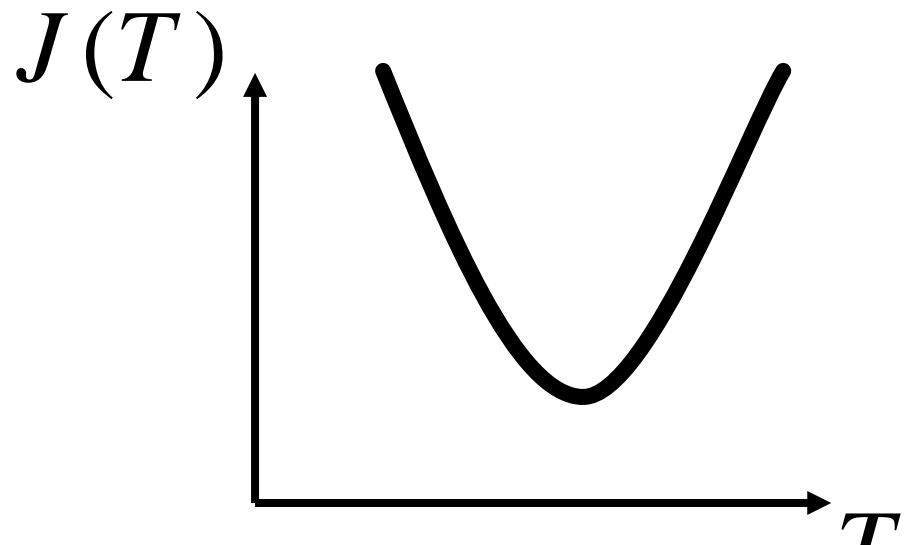
$$\Rightarrow \frac{1}{s_a^2} = \frac{1}{s_1^2} + \frac{1}{s_2^2}$$

The precision of the analysis is the sum of the precisions of the measurements. The analysis therefore has higher precision than any single measurement (if the statistics are correct).

Variational Approach

$$J(T) = \frac{1}{2} \left[\frac{(T - T_1)^2}{\mathbf{s}_1^2} + \frac{(T - T_2)^2}{\mathbf{s}_2^2} \right]$$

T_a is the value of T for which $\frac{\partial J}{\partial T} = 0$



Maximum Likelihood Estimate

- Obtain or assume probability distributions for the errors
- The best estimate of the state is chosen to have the greatest probability, or maximum likelihood
- If errors normally distributed, unbiased and uncorrelated, then states estimated by minimum variance and maximum likelihood are the same

Maximum Likelihood Approach (Bayesian Derivation)

Assume we have already made an observation T_1 ,
(the background forecast in data assimilation).

Then the probability distribution of the truth T
for Gaussian error statistics is :

$$p(T) \propto \exp\left(-\frac{(T - T_1)^2}{2\mathbf{s}_1^2}\right) \quad \text{the prior pdf.}$$

Maximum Likelihood Continued

Bayes' s formula for the posterior pdf given the observation T_2 is :

$$p(T | T_2) = \frac{p(T_2 | T) p(T)}{p(T_2)} \propto \exp\left(-\frac{(T_2 - T)^2}{2\mathbf{s}_2^2}\right) \exp\left(-\frac{(T - T_1)^2}{2\mathbf{s}_1^2}\right)$$

since $p(T_2)$ is normalising factor independent of T.

Maximize the posterior pdf (or ln) to estimate the truth.

We get the same answer as minimizing the cost function.

Equivalence holds for multi - dimensional case (for Gaussian statistics)

Simple Sequential Assimilation

Let $T_1 = T_b$ $T_2 = T_o$

$T_a = T_b + W(T_o - T_b)$ where $(T_o - T_b)$ is the "innovation".

The optimal weight W is given by :

$W = \mathbf{s}_b^2 (\mathbf{s}_b^2 + \mathbf{s}_o^2)^{-1}$, and the analysis error variance is :

$$\mathbf{s}_a^2 = (1 - W) \mathbf{s}_b^2$$

Comments

- The analysis is obtained by adding first guess to the innovation.
- Optimal weight is background error variance multiplied by inverse of total variance.
- Precision of analysis is sum of precisions of background and observation.
- Error variance of analysis is error variance of background *reduced* by (1- optimal weight).

Simple Assimilation Cycle

- Observation used once and then discarded.
- Forecast phase to update T_b and \mathbf{s}_b^2
- Analysis phase to update T_a and \mathbf{s}_a^2
- Obtain background as

$$T_b(t_{i+1}) = M[T_a(t_i)]$$

- Obtain variance of background as

$$\mathbf{s}_b^2(t_{i+1}) = \mathbf{s}_b^2(t_i) \quad \text{alternatively} \quad \mathbf{s}_b^2(t_{i+1}) = a\mathbf{s}_a^2(t_i)$$

Simple Kalman Filter

Analysis step as before.

$$T_t(t_{i+1}) = M[T_t(t_i)] - \mathbf{e}_m, \quad Q^2 = \langle \mathbf{e}_m^2 \rangle \quad (\text{model not biased!})$$

$$\begin{aligned} \text{Then } \mathbf{e}_{b,i+1} &= (T_b - T_t)_{i+1} = M(T_{a,i}) - M(T_{t,i}) + \mathbf{e}_m \\ &= \mathbf{M}\mathbf{e}_{a,i} + \mathbf{e}_m \quad \text{where } \mathbf{M} = \partial M / \partial T \end{aligned}$$

Forecast background error covariance is :

$$(\mathbf{s}_{b,i+1})^2 = \langle (\mathbf{e}_{b,i+1})^2 \rangle = \mathbf{M}^2 (\mathbf{s}_{a,i})^2 + Q^2$$

Multivariate Data Assimilation

Multivariate Case

state vector $\mathbf{x}(t) = \begin{pmatrix} x_1 \\ x_2 \\ \bullet \\ \bullet \\ x_n \end{pmatrix}$

observation vector $\mathbf{y}(t) = \begin{pmatrix} y_1 \\ y_2 \\ \bullet \\ y_m \end{pmatrix}$

State Vectors

\mathbf{X} state vector (column matrix)

\mathbf{X}_t true state

\mathbf{X}_b background state

\mathbf{X}_a analysis, estimate of **\mathbf{X}_t**

Ingredients of Good Estimate of the State Vector (“analysis”)

- Start from a good “first guess” (forecast from previous good analysis)
- Allow for errors in observations and first guess (give most weight to data you trust)
- Analysis should be smooth
- Analysis should respect known physical laws

Some Useful Matrix Properties

Transpose of a product : $(\mathbf{AB})^T = \mathbf{B}^T \mathbf{A}^T$

Inverse of a product : $(\mathbf{AB})^{-1} = \mathbf{B}^{-1} \mathbf{A}^{-1}$

Inverse of a transpose : $(\mathbf{A}^T)^{-1} = (\mathbf{A}^{-1})^T$

Positive definiteness for symmetric matrix \mathbf{A} :

$\forall \mathbf{x}$, the scalar $\mathbf{x} \mathbf{A} \mathbf{x}^T > 0$, unless $\mathbf{x} = \mathbf{0}$.

(this property is conserved through inversion)

Observations

- Observations are gathered into an observation vector \mathbf{y} , called the observation vector.
- Usually fewer observations than variables in the model; they are irregularly spaced; and may be of a different kind to those in the model.
- Introduce an observation operator to map from model state space to observation space.

$$\mathbf{x} \rightarrow H(\mathbf{x})$$

Errors

Variance becomes Covariance Matrix

- Errors in x_i are often correlated
 - spatial structure in flow
 - dynamical or chemical relationships
- Variance for scalar case becomes Covariance Matrix for vector case COV
- Diagonal elements are the variances of x_i
- Off-diagonal elements are covariances between x_i and x_j
- Observation of x_i affects estimate of x_j

The Error Covariance Matrix

$$\mathbf{e} = \begin{pmatrix} e_1 \\ e_2 \\ \cdot \\ \cdot \\ \cdot \\ e_n \end{pmatrix}$$

$$\mathbf{e}^T = (e_1 \quad e_2 \quad \cdot \quad \cdot \quad \cdot \quad e_n)$$

$$\langle e_i e_i \rangle = \mathbf{s}_i^2$$

$$\mathbf{P} = \langle \mathbf{e} \mathbf{e}^T \rangle = \begin{pmatrix} \langle e_1 e_1 \rangle & \langle e_1 e_2 \rangle & \cdot & \cdot & \cdot & \langle e_1 e_n \rangle \\ \langle e_2 e_1 \rangle & \langle e_2 e_2 \rangle & \cdot & \cdot & \cdot & \langle e_2 e_n \rangle \\ \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\ \langle e_n e_1 \rangle & \langle e_n e_2 \rangle & \cdot & \cdot & \cdot & \langle e_n e_n \rangle \end{pmatrix}$$

Background Errors

- They are the estimation errors of the background state:

$$\mathbf{e}_b = \mathbf{x}_b - \mathbf{x}_t$$

- average (bias) $\langle \mathbf{e}_b \rangle$
- covariance

$$\mathbf{B} = \langle (\mathbf{e} - \langle \mathbf{e}_b \rangle)(\mathbf{e} - \langle \mathbf{e}_b \rangle)^T \rangle$$

Observation Errors

- They contain errors in the observation process (instrumental error), errors in the design of H , and “representativeness errors”, i.e. discretization errors that prevent \mathbf{x}_t from being a perfect representation of the true state.

$$\mathbf{e}_o = \mathbf{y} - H(\mathbf{x}_t) \quad \langle \mathbf{e}_o \rangle$$

$$\mathbf{R} = \langle (\mathbf{e}_o - \langle \mathbf{e}_o \rangle)(\mathbf{e}_o - \langle \mathbf{e}_o \rangle)^T \rangle$$

Control Variables

- We may not be able to solve the analysis problem for all components of the model state (e.g. cloud-related variables, or need to reduce resolution)
- The work space is then not the model space but the sub-space in which we correct \mathbf{X}_b , called control-variable space

$$\mathbf{X}_a = \mathbf{X}_b + \mathbf{dx}$$

Innovations and Residuals

- Key to data assimilation is the use of differences between observations and the state vector of the system
- We call $\mathbf{y} - H(\mathbf{x}_b)$ the innovation
- We call $\mathbf{y} - H(\mathbf{x}_a)$ the analysis residual

Give important information

Analysis Errors

- They are the estimation errors of the analysis state that we want to minimize.

$$\mathbf{e}_a = \mathbf{x}_a - \mathbf{x}_t$$

Covariance matrix \mathbf{A}

Using the Error Covariance Matrix

Recall that an error covariance matrix \mathbf{C} for the error in \mathbf{x} has the form:

$$\mathbf{C} = \langle \mathbf{e}\mathbf{e}^T \rangle$$

If $\mathbf{y} = \mathbf{H}\mathbf{x}$ where \mathbf{H} is a matrix, then the error covariance for \mathbf{y} is given by:

$$\mathbf{C}_y = \mathbf{H}\mathbf{C}\mathbf{H}^T$$

BLUE Estimator

- The BLUE estimator is given by:

$$\mathbf{x}_a = \mathbf{x}_b + \mathbf{K}(\mathbf{y} - H(\mathbf{x}_b))$$

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

- The analysis error covariance matrix is:

$$\mathbf{A} = (\mathbf{I} - \mathbf{K}\mathbf{H})\mathbf{B}$$

- Note that:

$$\mathbf{B}\mathbf{H}^T (\mathbf{H}\mathbf{B}\mathbf{H}^T + \mathbf{R})^{-1} = (\mathbf{B}^{-1} + \mathbf{H}^T \mathbf{R}^{-1} \mathbf{H})^{-1} \mathbf{H}^T \mathbf{R}^{-1}$$

Statistical Interpolation with Least Squares Estimation

- Called *Best Linear Unbiased Estimator (BLUE)*.
- Simplified versions of this algorithm yield the most common algorithms used today in meteorology and oceanography.

Assumptions Used in *BLUE*

- Linearized observation operator:

$$H(\mathbf{x}) - H(\mathbf{x}_b) = \mathbf{H}(\mathbf{x} - \mathbf{x}_b)$$

- \mathbf{B} and \mathbf{R} are positive definite.

- Errors are unbiased:

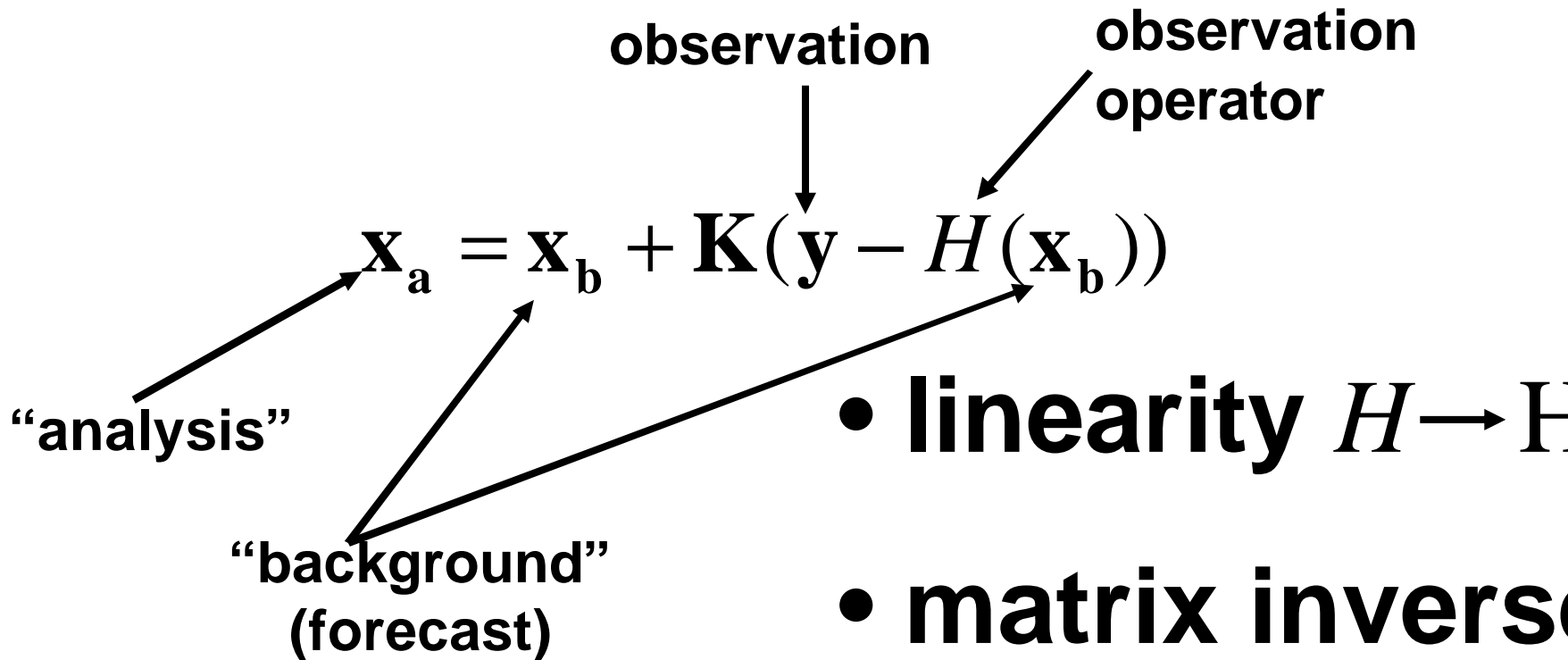
$$\langle \mathbf{x}_b - \mathbf{x}_t \rangle = \langle \mathbf{y} - H(\mathbf{x}_t) \rangle = 0$$

- Errors are uncorrelated:

$$\langle (\mathbf{x}_b - \mathbf{x}_t)(\mathbf{y} - H(\mathbf{x}_t))^T \rangle = 0$$

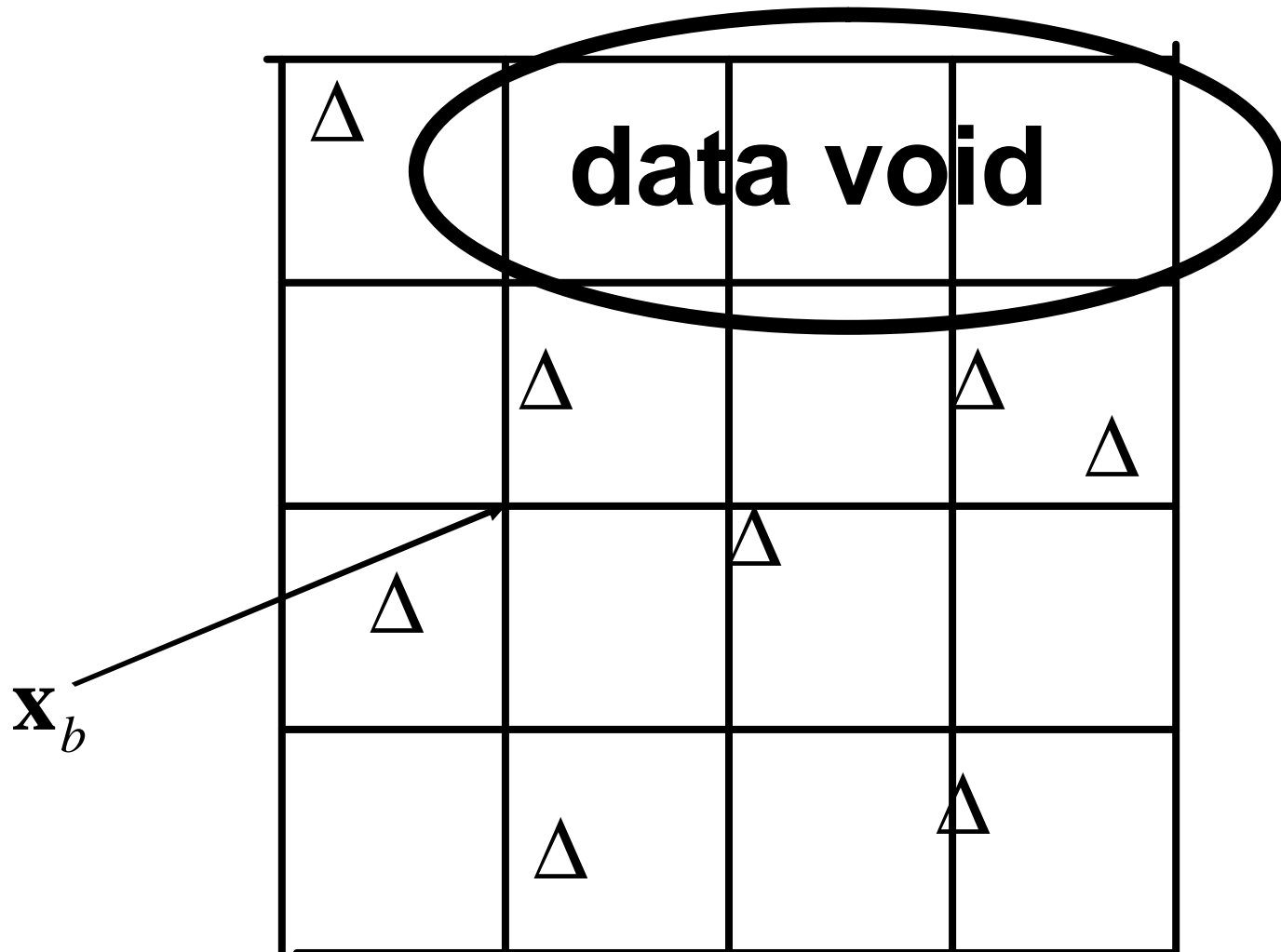
- Linear analysis: corrections to background depend linearly on (background – obs.).
- Optimal analysis: minimum variance estimate.

Optimal Interpolation



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

$$\Delta = \mathbf{y} - H(\mathbf{x}_b) \leftarrow \text{at obs. point}$$



END