Introduction to Data Assimilation, Subgrid-scale Parameterization and Predictability

Christian Franzke

Meteorological Institute

Center for Earth System Research and Sustainability

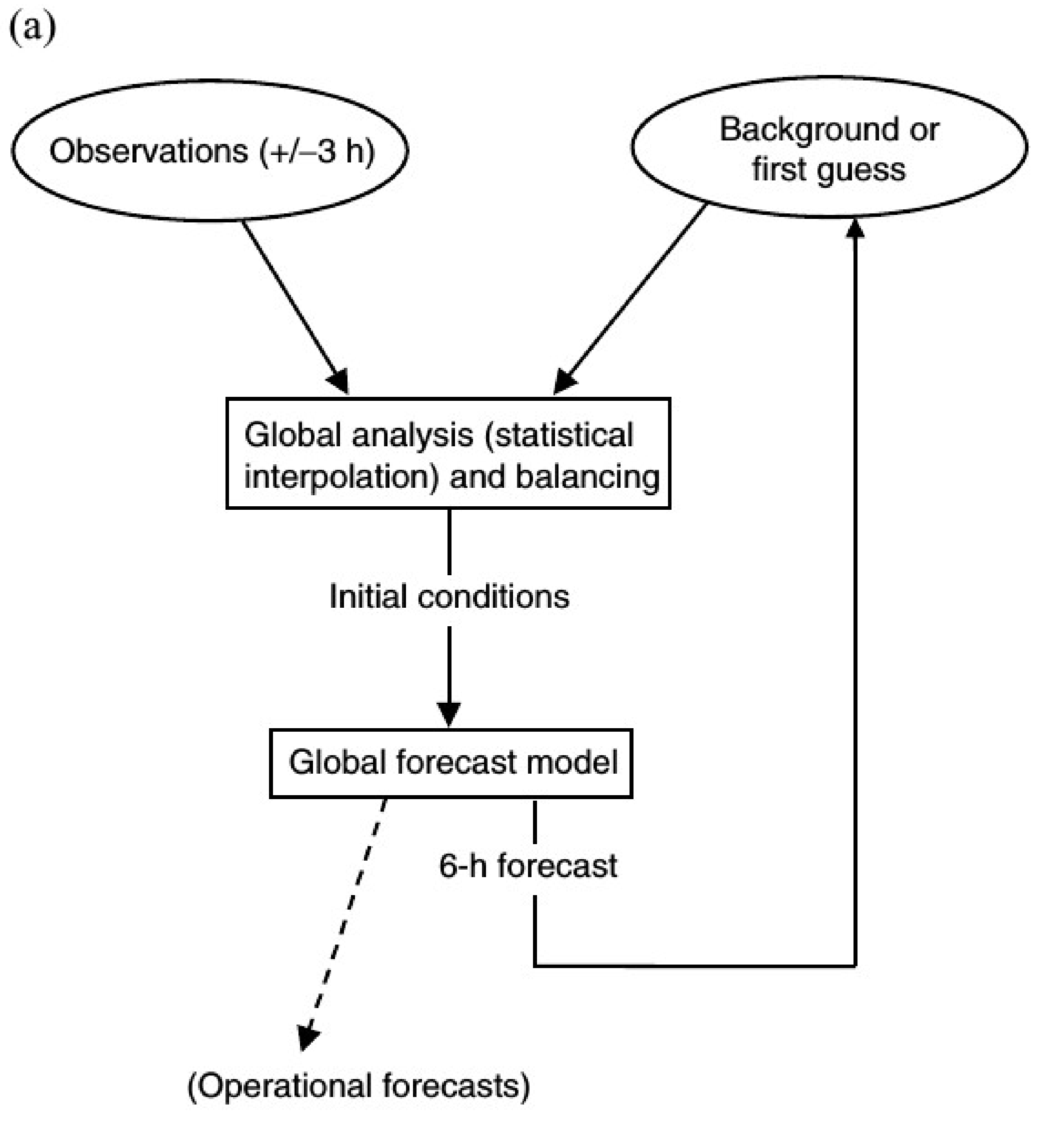
University of Hamburg

Email: christian.franzke@uni-hamburg.de

Outline

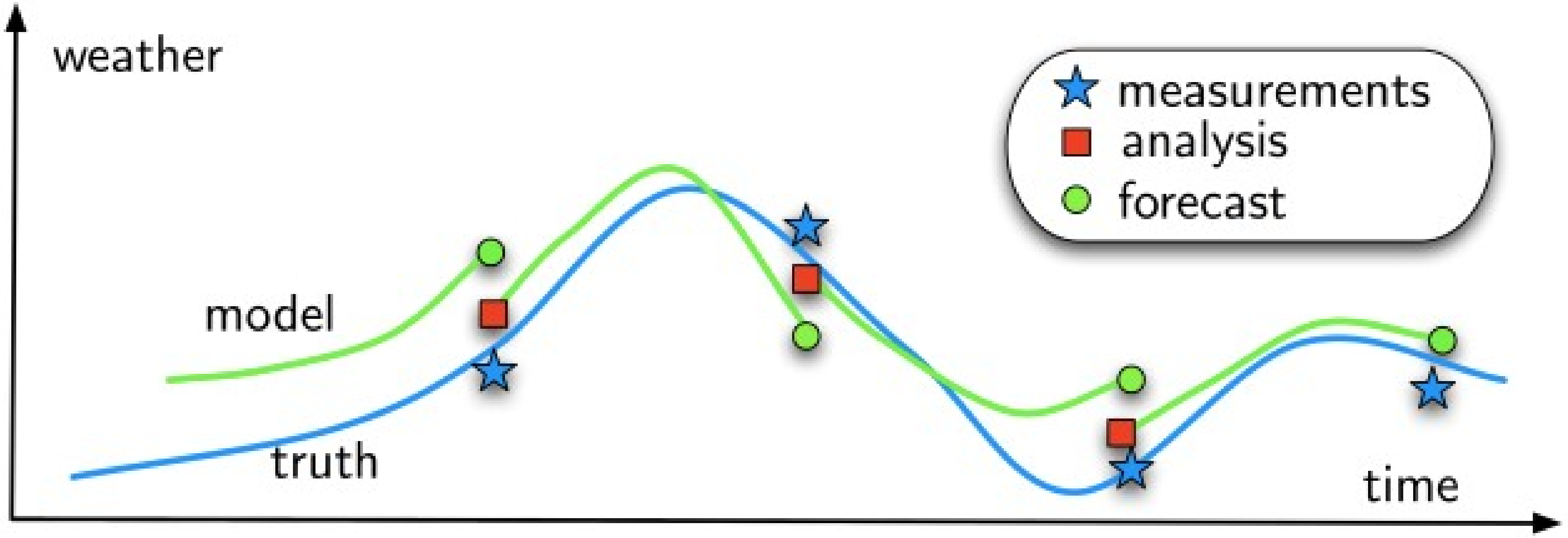
- Multivariate Data Assimilation
- Ensemble Kalman Filter

Forecast Cycle



Kalnay 2003

Data Assimilation



Olivier Thual

Optimal Interpolation

$$\mathbf{x}_t - \mathbf{x}_b = \mathbf{W}[\mathbf{y}_o - H(\mathbf{x}_b)] - \varepsilon_a = \mathbf{W}\mathbf{d} - \varepsilon_a$$

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

- Truth, analysis and background are vectors of length n
- Weights **W** are (n×p) matrix
- y_0: Observation
- x_t: Truth
- Observations are vector of length p
- Observational operator H can be nonlinear

Kalnay 2003

Optimal Interpolation

Observational increments vector:

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

We define background and analysis error as:

$$\boldsymbol{\varepsilon}_b(x, y) = \mathbf{x}_b(x, y) - \mathbf{x}_t(x, y)$$

$$\boldsymbol{\varepsilon}_a(x, y) = \mathbf{x}_a(x, y) - \mathbf{x}_t(x, y)$$

Kalnay 2003

Optimal Interpolation

Observational errors (at irregularly spaced points)

$$\boldsymbol{\varepsilon}_{oi} = \mathbf{y}_o(\mathbf{r}_i) - \mathbf{y}_t(\mathbf{r}_i) = \mathbf{y}_o(\mathbf{r}_i) - H[\mathbf{x}_t(\mathbf{r}_i)]$$

Kalnay 2003

Optimal Interpolation

We define error covariance matrices:

$$\mathbf{P}_a = \mathbf{A} = E\left\{\boldsymbol{\varepsilon}_a \boldsymbol{\varepsilon}_a^T\right\}$$

$$\mathbf{P}_b = \mathbf{B} = E\left\{\boldsymbol{\varepsilon}_b \boldsymbol{\varepsilon}_b^T\right\}$$

$$\mathbf{P}_o = \mathbf{R} = E\left\{\boldsymbol{\varepsilon}_o \boldsymbol{\varepsilon}_o^T\right\}$$

Kalnay 2003

Optimal Interpolation

The nonlinear observation operator H that transforms model variables into observed variables can be linearized as:

$$H(\mathbf{x} + \delta \mathbf{x}) = H(\mathbf{x}) + \mathbf{H}\delta \mathbf{x}$$

where **H** is a p×n matrix, denoting the linear observation operator with elements

$$h_{i,j} = \partial H_i / \partial x_j$$

Kalnay 2003

Optimal Interpolation

We assume that the background is a good approximation of the truth, so that the analysis and the observations are equal to the background values plus small increments

$$\mathbf{d} = \mathbf{y}_o - H(\mathbf{x}_b) = \mathbf{y}_o - H(\mathbf{x}_t + (\mathbf{x}_b - \mathbf{x}_t))$$

$$= \mathbf{y}_o - H(\mathbf{x}_t) - \mathbf{H}(\mathbf{x}_b - \mathbf{x}_t) = \boldsymbol{\varepsilon}_o - \mathbf{H}\boldsymbol{\varepsilon}_b$$

The **H** matrix transforms vectors from model to observation space.
Its transpose or adjoint $\mathbf{H}^T$ transforms vectors from observation to model space.

Kalnay 2003

Optimal Interpolation

Best unbiased estimation: Least Squares

$$\mathbf{W} = E\left(\mathbf{x}\mathbf{y}^T\right)\left[E\left(\mathbf{y}\mathbf{y}^T\right)\right]^{-1}$$

Kalnay 2003

How to estimate W?

Multiple Regression

How to estimate W?

Multiple Regression

Assume we have two time series of vectors

$$\mathbf{x}(t) = \begin{bmatrix} x_1(t) \\ x_2(t) \\ \vdots \\ x_n(t) \end{bmatrix} \qquad \mathbf{y}(t) = \begin{bmatrix} y_1(t) \\ y_2(t) \\ \vdots \\ y_p(t) \end{bmatrix}$$

Kalnay 2003

How to estimate W?

Multiple Regression

We now derive the best linear unbiased estimation of **x** in terms of **y**, the optimal value of **W**

$$\mathbf{x}_a(t) = \mathbf{W}\mathbf{y}(t)$$

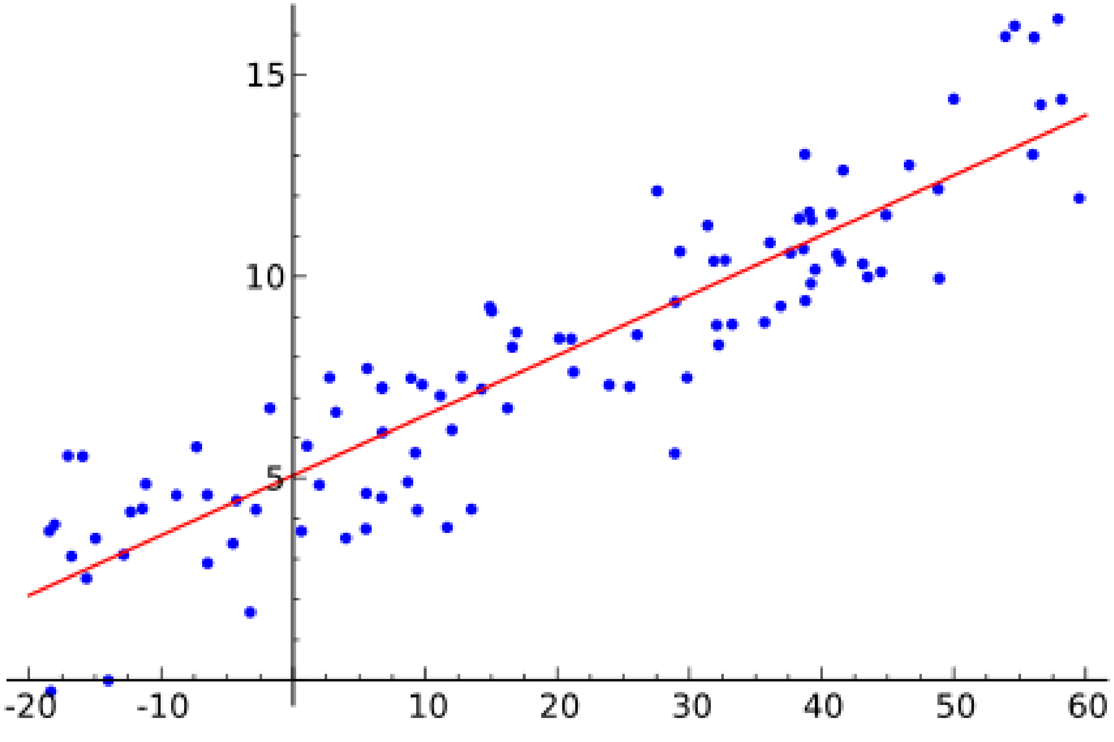
which approximates the true relationship

$$\mathbf{x}(t) = \mathbf{W}\mathbf{y}(t) - \varepsilon(t)$$

linear regression (analysis) error

Kalnay 2003

How to estimate W?



Sewaqu 2010

How to estimate W?

W minimizes mean squared error $E(\varepsilon^T \varepsilon)$

$$x_i(t) = \sum_{k=1}^{p} w_{ik} y_k(t) - \varepsilon_i(t)$$

$$\longrightarrow \quad \sum_{i=1}^{n} \varepsilon_i^2(t) = \sum_{i=1}^{n} \left[\sum_{k=1}^{p} w_{ik} y_k(t) - x_i(t) \right]^2$$

Kalnay 2003

How to estimate W?

Derivative with respect to weight matrix

$$\frac{\partial \sum_{i=1}^{n} \varepsilon_i^2(t)}{\partial w_{ij}} = 2\left[\sum_{k=1}^{p} w_{ik} y_k(t) - x_i(t)\right]\left[y_j(t)\right]$$

$$= 2\left[\sum_{k=1}^{p} w_{ik} y_k(t) y_i(t) - x_i(t) y_j(t)\right]$$

in Matrix form

$$\frac{\partial \boldsymbol{\varepsilon}^T \boldsymbol{\varepsilon}}{\partial w_{ij}} = 2\left\{\left[\mathbf{W}\mathbf{y}(t)\mathbf{y}^T(t)\right]_{ij} - \left[\mathbf{x}(t)\mathbf{y}^T(t)\right]_{ij}\right\}$$

Kalnay 2003

How to estimate W?

We get the normal equations

$$\mathbf{W}E\left(\mathbf{y}\mathbf{y}^T\right) - E\left(\mathbf{x}\mathbf{y}^T\right) = 0$$

or

$$\mathbf{W} = E\left(\mathbf{x}\mathbf{y}^T\right)\left[E\left(\mathbf{y}\mathbf{y}^T\right)\right]^{-1}$$

which gives the best unbiased estimation

$$\mathbf{x}_a(t) = \mathbf{W}\mathbf{y}(t)$$

Kalnay 2003

Optimal Interpolation

Best unbiased estimation: Least Squares

$$\mathbf{W} = E\left(\mathbf{x}\mathbf{y}^T\right)\left[E\left(\mathbf{y}\mathbf{y}^T\right)\right]^{-1}$$

Estimate of weight matrix:

$$\begin{aligned}\mathbf{W} &= E\{(\mathbf{x}_t - \mathbf{x}_b)[\mathbf{y}_o - H(\mathbf{x}_b)]^T\}(E\{[\mathbf{y}_o - H(\mathbf{x}_b)][\mathbf{y}_o - H(\mathbf{x}_b)]^T\})^{-1} \\ &= E[(-\boldsymbol{\varepsilon}_b)(\boldsymbol{\varepsilon}_o - \mathbf{H}\boldsymbol{\varepsilon}_b)^T]\{E[(\boldsymbol{\varepsilon}_o - \mathbf{H}\boldsymbol{\varepsilon}_b)(\boldsymbol{\varepsilon}_o - \mathbf{H}\boldsymbol{\varepsilon}_b)^T]\}^{-1}\end{aligned}$$

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

Kalnay 2003

Optimal Interpolation

The analysis covariance matrix is:

$$\mathbf{P}_a = E\{\boldsymbol{\varepsilon}_a \boldsymbol{\varepsilon}_a^T\} = E\{\boldsymbol{\varepsilon}_b \boldsymbol{\varepsilon}_b^T + \boldsymbol{\varepsilon}_b(\boldsymbol{\varepsilon}_o - \mathbf{H}\boldsymbol{\varepsilon}_b)^T \mathbf{W}^T$$
$$+\mathbf{W}(\boldsymbol{\varepsilon}_o - \mathbf{H}\boldsymbol{\varepsilon}_b)\boldsymbol{\varepsilon}_b^T + \mathbf{W}(\boldsymbol{\varepsilon}_o - \mathbf{H}\boldsymbol{\varepsilon}_b)(\boldsymbol{\varepsilon}_o - \mathbf{H}\boldsymbol{\varepsilon}_b)^T \mathbf{W}^T\}$$
$$= \mathbf{B} - \mathbf{B}\mathbf{H}^T\mathbf{W}^T - \mathbf{W}\mathbf{H}\mathbf{B} + \mathbf{W}\mathbf{R}\mathbf{W}^T + \mathbf{W}\mathbf{H}\mathbf{B}\mathbf{H}^T\mathbf{W}^T$$

Kalnay 2003

Optimal Interpolation

The analysis covariance matrix is:

$$\begin{aligned}\mathbf{P}_a = E\{\boldsymbol{\varepsilon}_a\boldsymbol{\varepsilon}_a^T\} = E\{&\boldsymbol{\varepsilon}_b\boldsymbol{\varepsilon}_b^T + \boldsymbol{\varepsilon}_b(\boldsymbol{\varepsilon}_o - \mathbf{H}\boldsymbol{\varepsilon}_b)^T\mathbf{W}^T \\ &+\mathbf{W}(\boldsymbol{\varepsilon}_o - \mathbf{H}\boldsymbol{\varepsilon}_b)\boldsymbol{\varepsilon}_b^T + \mathbf{W}(\boldsymbol{\varepsilon}_o - \mathbf{H}\boldsymbol{\varepsilon}_b)(\boldsymbol{\varepsilon}_o - \mathbf{H}\boldsymbol{\varepsilon}_b)^T\mathbf{W}^T\} \\ = \mathbf{B} - \mathbf{B}\mathbf{H}^T\mathbf{W}^T - \mathbf{W}\mathbf{H}\mathbf{B} + \mathbf{W}\mathbf{R}\mathbf{W}^T + \mathbf{W}\mathbf{H}\mathbf{B}\mathbf{H}^T\mathbf{W}^T&\end{aligned}$$

This can be written as:

$$\mathbf{P}_a = (\mathbf{I} - \mathbf{W}\mathbf{H})\mathbf{B}$$

Kalnay 2003

Summary of OI

Basic equations of OI:

$$\mathbf{x}_a = \mathbf{x}_b + \mathbf{W}[\mathbf{y}_o - H(\mathbf{x}_b)] = \mathbf{x}_b + \mathbf{W}\mathbf{d}$$

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

$$\mathbf{P}_a = (\mathbf{I}_n - \mathbf{W}\mathbf{H})\mathbf{B}$$

Kalnay 2003

Summary of OI

Interpretation

$$\mathbf{x}_a = \mathbf{x}_b + \mathbf{W}[\mathbf{y}_o - H(\mathbf{x}_b)] = \mathbf{x}_b + \mathbf{W}\mathbf{d}$$

The analysis is obtained by adding to the first guess the product of the optimal weight matrix and the innovation (difference between observation and first guess).

Kalnay 2003

Summary of OI

Interpretation

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

The optimal weight matrix is given by the background error covariance in the observation space ($\mathbf{BH}^T$) multiplied by the inverse of the total error covariance.

Kalnay 2003

Summary of OI

Interpretation

$$\mathbf{P}_a = (\mathbf{I}_n - \mathbf{WH})\mathbf{B}$$

The error covariance of the analysis is given by the error covariance of the background, reduced by a matrix equal to the identity matrix minus the optimal weight matrix.

Kalnay 2003

3D-VAR ~ OI

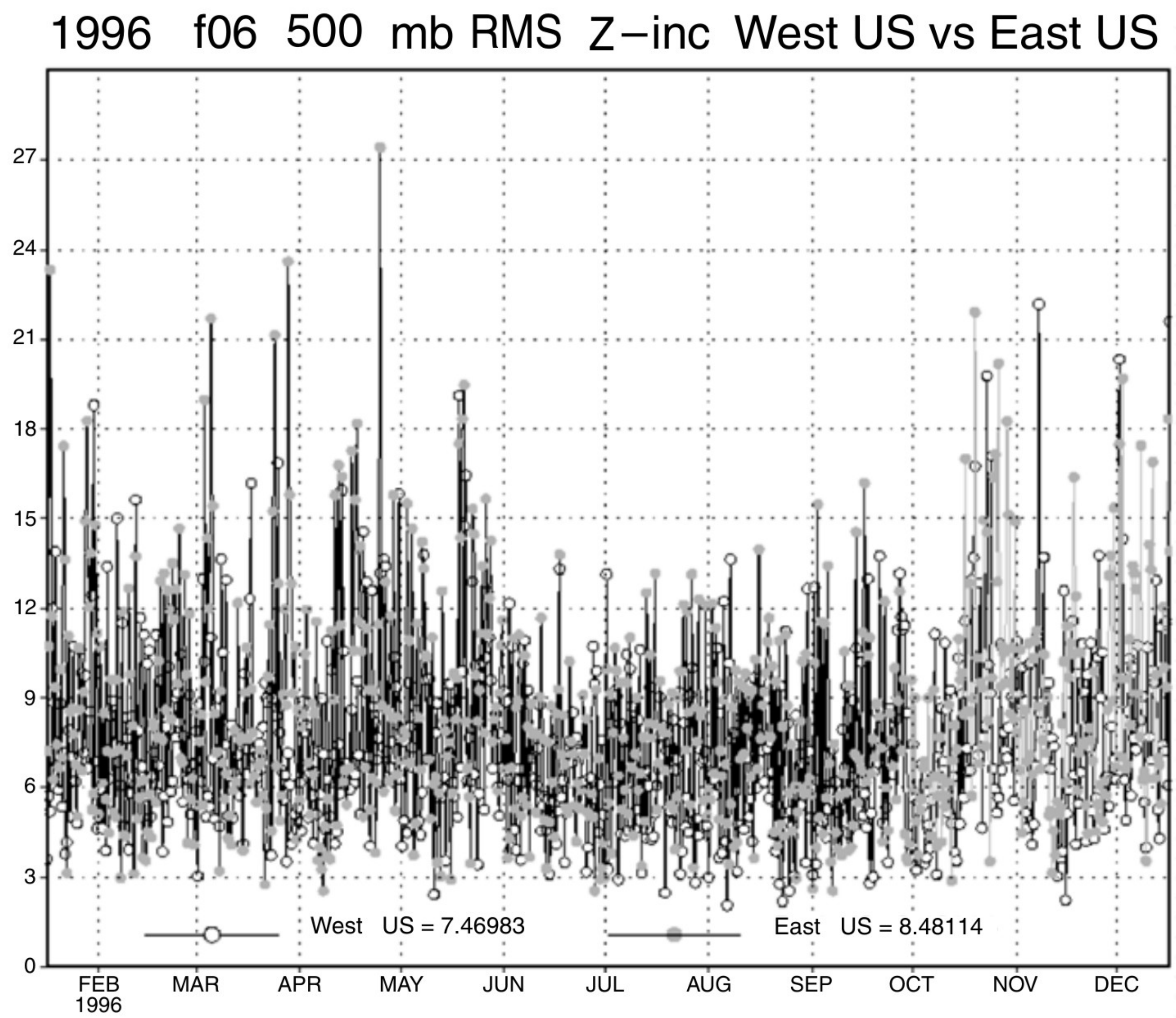
- OI: Find optimal weight matrix **W** that minimizes the analysis error covariance

- 3D-VAR: Find optimal analysis field that minimizes a cost function

$$2J(\mathbf{x}) = (\mathbf{x} - \mathbf{x}_b)^T \mathbf{B}^{-1} (\mathbf{x} - \mathbf{x}_b) + [\mathbf{y}_o - H(\mathbf{x})]^T \mathbf{R}^{-1} [\mathbf{y}_o - H(\mathbf{x})]$$

Kalnay 2003

DA with evolving error covariance



Kalnay 2003

3D-VAR ~ 4D-VAR

3D-VAR

$$2J(\mathbf{x}) = (\mathbf{x} - \mathbf{x}_b)^T \mathbf{B}^{-1} (\mathbf{x} - \mathbf{x}_b) + [\mathbf{y}_o - H(\mathbf{x})]^T \mathbf{R}^{-1} [\mathbf{y}_o - H(\mathbf{x})]$$

4D-VAR

$$J[\mathbf{x}(t_0)] = \frac{1}{2}[\mathbf{x}(\mathrm{t}_0) - \mathbf{x}^b(t_0)]^T \mathbf{B}_0^{-1} [\mathbf{x}(\mathrm{t}_0) - \mathbf{x}^b(t_0)] + \frac{1}{2}\sum_{i=0}^{N} \left[H(\mathbf{x}_i) - \mathbf{y}_i^0\right]^T \mathbf{R}_i^{-1} \left[H(\mathbf{x}_i) - \mathbf{y}_i^o\right]$$

Kalnay 2003

4D-VAR

$$J[\mathbf{x}(t_0)] = \frac{1}{2}[\mathbf{x}(\mathrm{t}_0) - \mathbf{x}^b(t_0)]^T \mathbf{B}_0^{-1}[\mathbf{x}(\mathrm{t}_0) - \mathbf{x}^b(t_0)] + \frac{1}{2}\sum_{i=0}^{N}\left[H(\mathbf{x}_i) - \mathbf{y}_i^0\right]^T \mathbf{R}_i^{-1}\left[H(\mathbf{x}_i) - \mathbf{y}_i^o\right]$$

The cost function includes a term measuring the distance to the background at the beginning of the interval and a summation over time of the cost function for each observation increment

Kalnay 2003

4D-VAR

4D-Var seeks an initial condition such that the forecast best fits the observations within the assimilation interval.

4D-Var assumes **perfect model**
→ Big disadvantage

Kalnay 2003

Kalman Filter ~ OI

In OI the forecast error covariance is computed once and for all.
Assumption: Forecast errors are statistically stationary

Kalman filter propagates error covariance forward
→ State-dependent error covariance

Kalnay 2003

Extended Kalman Filter

Forecast step:

Nonlinear model

$$\mathbf{x}^f(t_i) = M_{i-1}[\mathbf{x}^a(t_{i-1})]$$

$$\mathbf{P}^f(t_i) = \mathbf{L}_{i-1}\mathbf{P}^a(t_{i-1})\mathbf{L}_{i-1}^T + \mathbf{Q}(t_{i-1})$$

Tangent linear

Analysis step:

$$\mathbf{x}^a(t_i) = \mathbf{x}^f(t_i) + \mathbf{K}_i\mathbf{d}_i$$

$$\mathbf{P}^a(t_i) = (\mathbf{I} - \mathbf{K}_i\mathbf{H}_i)\mathbf{P}^f(t_i)$$

$$\mathbf{d}_i = \mathbf{y}_i^o - H[\mathbf{x}^f(t_i)]$$

Kalnay 2003

Kalman Filter

1) Define ensemble of observations

$y^o_j = y^o + \varepsilon_j$

Kalman Filter

1) Define ensemble of observations

$y^o_j = y^o + \varepsilon_j$

2) Define ensemble covariance matrix

$R_e = \overline{\varepsilon\varepsilon^T}$

Kalnay 2003

Kalman Filter

1) Define ensemble of observations

$$y^o_j = y^o + \varepsilon_j$$

2) Define ensemble covariance matrix

$$R_e = \overline{\varepsilon\varepsilon^T}$$

3) Analysis step

$$y^a_j = y^f_j + P^f_e H^T (H P^f_e H^T + R_e)^{-1} (y^o_j - H y^f_j)$$

Kalnay 2003

Kalman Filter

3) Analysis step

$$y^a_j = y^f_j + P^f_e H^T (H P^f_e H^T + R_e)^{-1} (y^o_j - H y^f_j)$$

4) Analysis and ensemble mean are identical

$$\overline{y}^a_j = \overline{y}^f_j + P^f_e H^T (H P^f_e H^T + R_e)^{-1} (\overline{y}^o_j - H \overline{y}^f_j)$$

5) Kalman gain (Optimal weight) matrix

$$K_e = P^f_e H^T (H P^f_e H^T + R_e)^{-1}$$

Kalnay 2003

Kalman Filter

5) Kalman gain (Optimal weight) matrix

$$K_e = P^f_e H^T (H P^f_e H^T + R_e)^{-1}$$

6) Error covariance

$$P^a_e = (I - K_e H) P^f_e$$

Kalnay 2003

Kalman Filter

5) Kalman gain (Optimal weight) matrix

$$K_e = P^f_e H^T (H P^f_e H^T + R_e)^{-1}$$

6) Error covariance

$$P^a_e = (I - K_e H) P^f_e$$

7) Model error covariance Q

Kalnay 2003

Kalman Filter

5) Kalman gain (Optimal weight) matrix

$$K_e = P^f_e H^T (H P^f_e H^T + R_e)^{-1}$$

6) Error covariance

$$P^a_e = (I - K_e H) P^f_e$$

7) Model error covariance Q

8) Ensemble error covariance evolves as

$P^{k+1}_e = F P^k_e F^T + Q + \text{n.l.}$ (F: tangent linear operator)

Kalnay 2003

Ensemble Kalman Filter

- An ensemble of K data assimilation cycles is carried out simultaneously
- All cycles assimilate the same observations
- Different perturbations are added to observations → ensemble forecasts
- This ensemble is used to estimate forecast error covariance

Ensemble mean

$$\mathbf{P}^f \approx \frac{1}{K-1}\sum_{k=1}^{K}\left(\mathbf{x}_k^f - \overline{\mathbf{x}}^f\right)\left(\mathbf{x}_k^f - \overline{\mathbf{x}}^f\right)^T$$

Kalnay 2003

Ensemble Kalman Filter

- Evolve each ensemble member forward using the nonlinear model perturbed by noise
  $y^a = M(y^f) + \varepsilon$
- Compute ensemble mean and covariance
  $\overline{y^f} = 1/N\boldsymbol{\Sigma} y^f$
  $P^f = y^f (y^f)^T = 1/(N-1)\boldsymbol{\Sigma}(y^f - \overline{y^f})(y^f - \overline{y^f})^T$
- Update analysis
  $y^a = y^f + P^f K(y^o - Hy^f)$

  $K = P^f H^T (HP^f H^T + R)^{-1}$
  Where R is the covariance from the observations

Kalnay 2003

Excercise

Write a Ensemble Kalman Filter for Lorenz 96