Basics of Data Assimilation

Jake Liu (liuz@ucar.edu)

Prediction, Assimilation, and Risk Communication Section
Mesoscale & Microscale Meteorology Laboratory
National Center for Atmospheric Research

NCAR

MPAS-JEDI Tutorial, INPE, 15-16 August, 2024

This material is based upon work supported by the National Center for Atmospheric Research, which is a major facility sponsored by the National Science Foundation under Cooperative Agreement No. 1852977.

Outline

- Scalar case
- Case with two state variables
- General n-dimensional case

What is data assimilation?

- A probabilistic method to obtain the best-possible estimate of state variables of a dynamic/physical system

- In the atmospheric sciences, DA typically involves combining a short-term model forecast (i.e., Background or Prior) and observations, along with their respective errors characterization, to produce an *analysis (Posterior)* that can initialize a numerical weather prediction model (e.g., WRF or MPAS)

Scalar Case

- State variable to estimate "x", e.g., consider this morning's 2-meter temperature at INPE, at 9 am local time, i. e., 12 UTC,

- Now we have a "background" (or "prior") information x_b of x, which is from a 6-h MONAN-v1.0 forecast initiated from 06 UTC GFS analysis.

- We also have an observation y of x at a surface station at INPE

- What is the best estimate (analysis) x_a of x?

Scalar Case

- We can simply average xb and y: $x_a = \frac{1}{2}(x_b + y)$
  - This actually means we trust equally the background and observation, giving them equal weight

- But if xb and y's accuracy are different and we have some knowledge about their errors
  - e.g., for background, we have statistics (e.g., mean and variance) of $x_b - y$ from the past
  - For observation, we have instrument error information from manufacturer

Scalar Case

- Then we can do a weighted mean: $x_a = ax_b + by$ in a least square sense, i.e.,

Minimize $$J(x) = \frac{1}{2}\frac{(x-x_b)^2}{\sigma_b^2} + \frac{1}{2}\frac{(x-y)^2}{\sigma_o^2}$$

Requires $$\frac{dJ(x)}{dx} = \frac{(x-x_b)}{\sigma_b^2} + \frac{(x-y)}{\sigma_o^2} = 0$$

Then we can easily get $$x_a = \frac{\sigma_o^2}{\sigma_b^2+\sigma_o^2}x_b + \frac{\sigma_b^2}{\sigma_b^2+\sigma_o^2}y = \frac{1}{1+\sigma_b^2/\sigma_o^2}x_b + \frac{1}{1+\sigma_o^2/\sigma_b^2}y$$

Or we can write in the form of analysis increment

Called “Innovation” or O minus B, or OMB

$$x_a - x_b = \frac{\sigma_b^2}{\sigma_b^2+\sigma_o^2}(y - x_b) = \frac{1}{1+\sigma_o^2/\sigma_b^2}(y - x_b)$$

Scalar Case

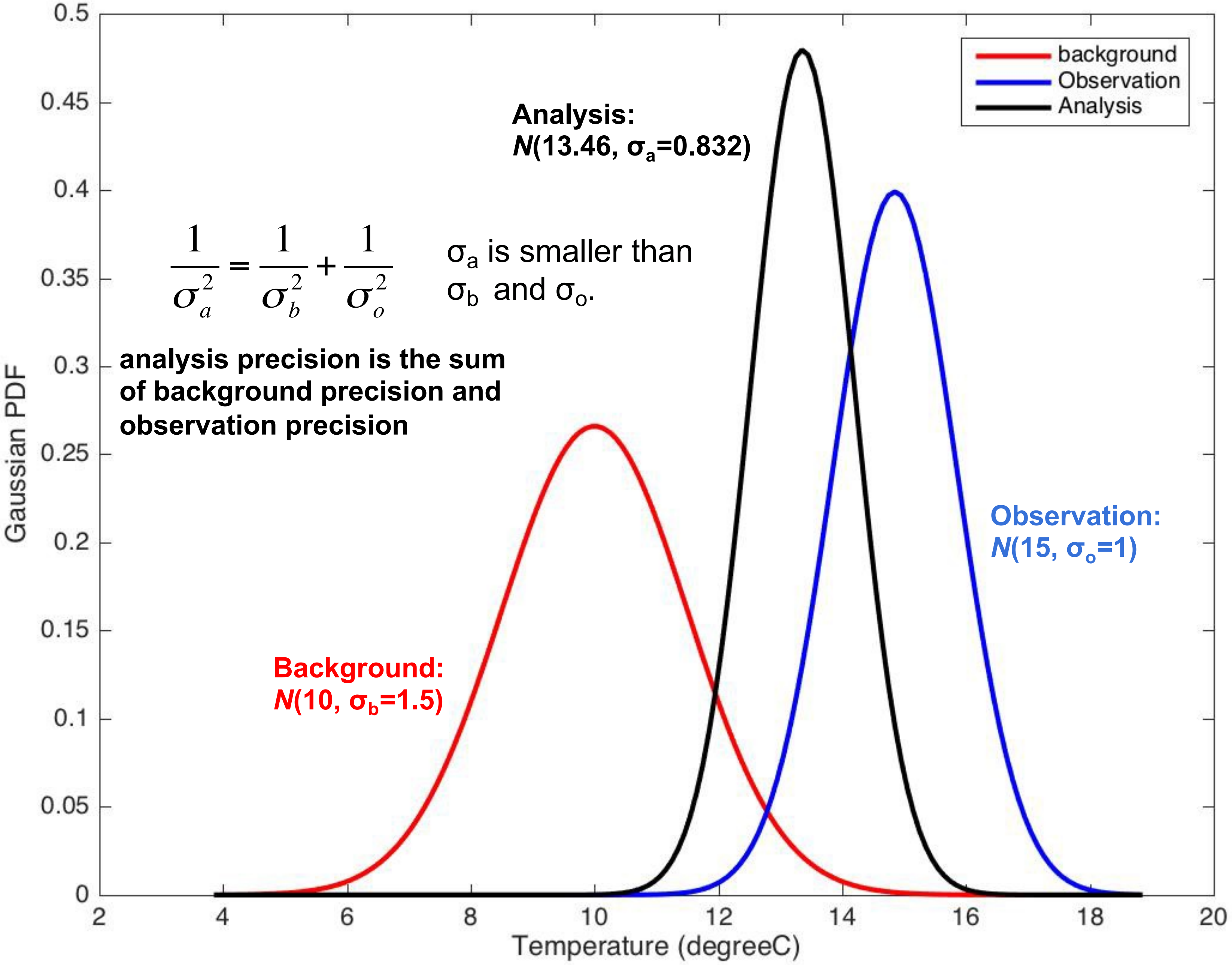
Minimize
$$J(x) = \frac{1}{2}\frac{(x-x_b)^2}{\sigma_b^2} + \frac{1}{2}\frac{(x-y)^2}{\sigma_o^2}$$

is actually equivalent to maximize a Gaussian Probability Distribution Function (PDF)

$$ce^{-J(x)}$$

Assume errors of X_b and y are unbiased

A probabilistic view of scale case



Two state variables case

- Consider two state variables to estimate: INPE and USP's 2m temperatures x_1 and x_2 at 12 UTC today.

- Background from 6-h forecast: x_1^b and x_2^b and their error covariance with correlation *c*

$$\mathbf{B} = \begin{bmatrix} \sigma_1^2 & c\sigma_1\sigma_2 \\ c\sigma_1\sigma_2 & \sigma_2^2 \end{bmatrix} = \begin{bmatrix} \sigma_1 & 0 \\ 0 & \sigma_2 \end{bmatrix} \begin{bmatrix} 1 & c \\ c & 1 \end{bmatrix} \begin{bmatrix} \sigma_1 & 0 \\ 0 & \sigma_2 \end{bmatrix}$$

- We only have an observation y_1 at the INPE station and its error variance σ_o^2

- Now we want to estimate T at 2 locations with obs at one location

Analysis increment for two variables

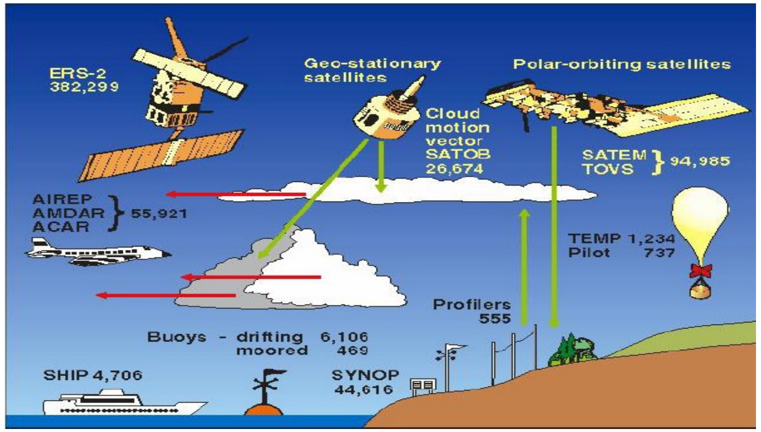
$$x_1^a - x_1^b = \frac{\sigma_1^2}{\sigma_1^2 + \sigma_o^2}(y_1 - x_1^b) \longleftarrow \text{INPE}$$

$$x_2^a - x_2^b = \frac{c\sigma_1\sigma_2}{\sigma_1^2 + \sigma_o^2}(y_1 - x_1^b) \longleftarrow \text{USP}$$

Unobserved variable x_2 gets updated through the error correlation *c* in the background error covariance.

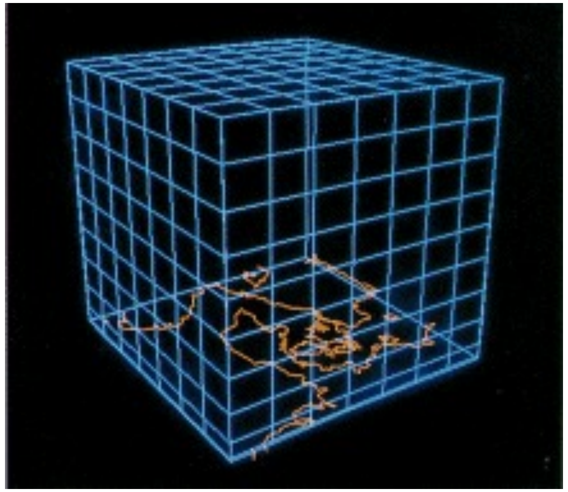
In general, this correlation can be correlation between two locations (spatial), two variables (multivariate), or two times (temporal).

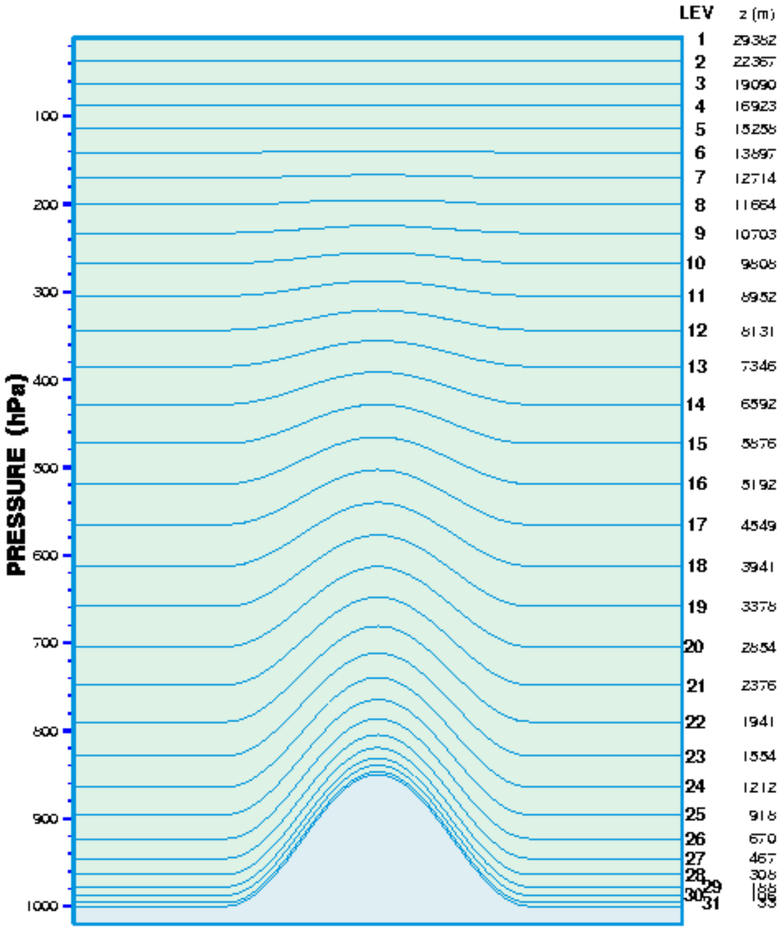
General Case



Observations
y^o, ~10^5-10^6

Model state
x, ~10^7





Vertical resolution of the DMI-HIRLAM system

General Case: vector and matrix notation

state vector

$$x = \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_m \end{bmatrix}$$

observation vector

$$y = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix}$$

background error covariance

$$\mathbf{B} = \begin{bmatrix} \sigma_1^2 & c_{12}\sigma_1\sigma_2 & \cdots & \cdots \\ c_{12}\sigma_1\sigma_2 & \sigma_2^2 & \cdots & \cdots \\ \cdots & \cdots & \ddots & \cdots \\ \cdots & \cdots & \cdots & \sigma_m^2 \end{bmatrix} = \sigma\, \mathrm{C}\, \sigma$$

Correlation matrix

m x m

Observation error covariance

$$\mathbf{R} = \begin{bmatrix} \sigma_{o1}^2 & 0 & \cdots & 0 \\ 0 & \sigma_{o2}^2 & \cdots & 0 \\ \vdots & \cdots & \ddots & \vdots \\ 0 & \cdots & \cdots & \sigma_{on}^2 \end{bmatrix}$$

n x n

General Case: cost function

1 x 1   1 x m   m x m   m x 1   1 x n   n x n   n x 1

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

H maps x to y space, e. g., interpolation.
Terminology in DA: observation operator
Superscript ‘T’: transpose of a vector or matrix,
Superscript ‘-1’: inverse of a symmetric covariance matrix

Minimize J(x) is equivalent to maximize a multi-dimensional Gaussian PDF

Constant * $e^{-J(x)}$

General Case: analytical solution

Again, minimize J requires its gradient (a vector) with respect to x equal to zero:

$$\nabla J_{\mathbf{x}}(\mathbf{x}) = \mathbf{B}^{-1}(\mathbf{x}-\mathbf{x}_{\mathrm{b}}) - \mathbf{H}^{\mathrm{T}}\mathbf{R}^{-1}[\mathbf{y}-\mathbf{H}\mathbf{x}] = 0$$

m x 1

This leads to analytical solution for the analysis increment:

$$x^a - x^b = \boxed{\mathbf{BH}^{\mathrm{T}}(\mathbf{HBH}^{\mathrm{T}} + \mathbf{R})^{-1}}\,\boxed{[y - \mathbf{H}x^b]}$$

Kalman gain matrix **Innovation or OMB vector**

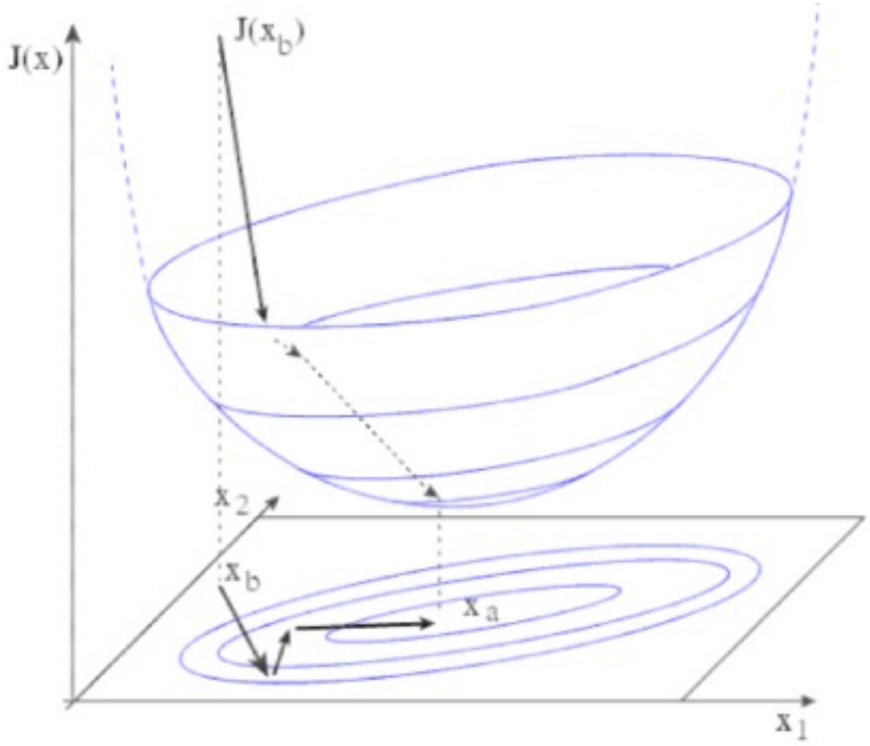
$\mathbf{HBH}^{\mathrm{T}}$: background error covariance projected into observation space

$\mathbf{BH}^{\mathrm{T}}$: background error covariance projected into cross background-observation space

Iterative algorithm to find minimum of cost function

- **Descending algorithms**
  - **Descending direction: γ_n (N-dimensional vector)**
  - **Descending step: μ_n**

$$\boxed{x_{n+1} = x_n + \mu_n \gamma_n}$$



from Bouttier and Courtier 1999

Precision of Analysis with optimal B and R

$$\mathbf{A}^{-1} = \mathbf{B}^{-1} + \mathbf{H}^{\mathrm{T}}\mathbf{R}^{-1}\mathbf{H}$$

Generalization of scalar case $\dfrac{1}{\sigma_a^2} = \dfrac{1}{\sigma_b^2} + \dfrac{1}{\sigma_o^2}$

Or in another form: $\mathbf{A} = (\mathbf{I} - \mathbf{K}\mathbf{H})\mathbf{B}$

With

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

called Kalman gain matrix

Precision of analysis: more general formulation

$$\mathbf{A} = (\mathbf{I} - \mathbf{KH})\mathbf{B}_t(\mathbf{I} - \mathbf{KH})^{\mathrm{T}} + \mathbf{KR}_t\mathbf{K}^{\mathrm{T}}$$

where B_t and R_t are “true” background and observation error covariances.

This formulation is valid for any given gain matrix K, which could be suboptimal (e.g., due to incorrect estimation/specification of B and R).

Analysis increment with a single humidity observation

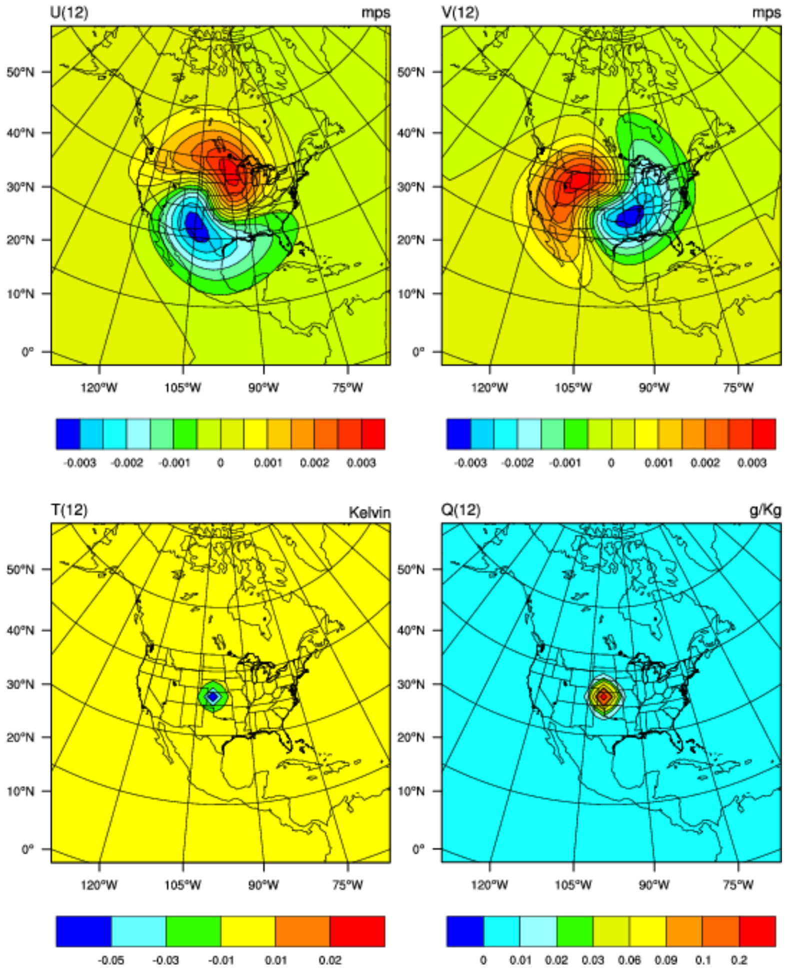
$$x^a - x^b = \mathbf{B}\mathbf{H}^{\mathrm{T}}(\mathbf{H}\mathbf{B}\mathbf{H}^{\mathrm{T}} + \mathbf{R})^{-1}[y - \mathbf{H}x^b]$$

$$x_l^a - x_l^b = \frac{c_{lk}\sigma_l\sigma_k}{\sigma_k^2 + \sigma_{ok}^2}(y_k - x_k^b)$$

It is generalization of previous two variables case:

$$x_1^a - x_1^b = \frac{\sigma_1^2}{\sigma_1^2 + \sigma_o^2}(y_1 - x_1^b)$$

$$x_2^a - x_2^b = \frac{c\sigma_1\sigma_2}{\sigma_1^2 + \sigma_o^2}(y_1 - x_1^b)$$

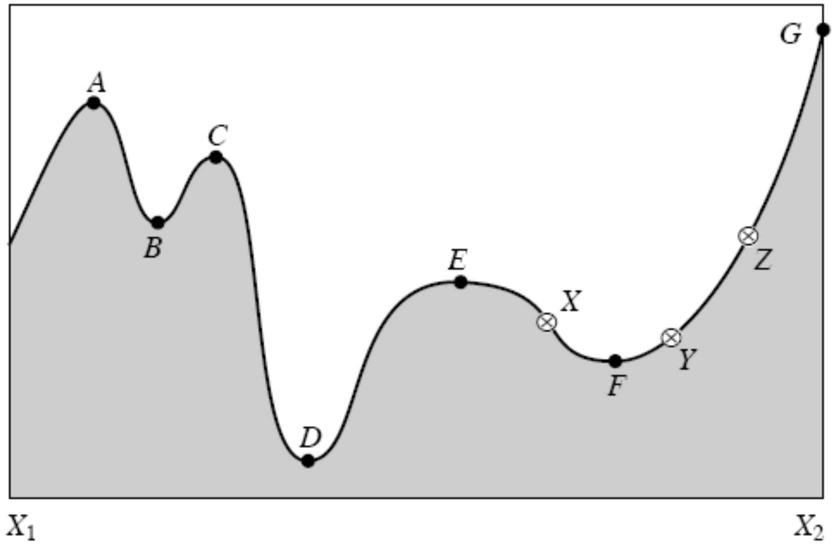


cv_options=6 in WRFDA

Other Remarks

- Observation operator H() can be non-linear and thus analysis error PDF is not necessarily Gaussian

- J(x) can have multiple local minima. Final solution of least square depends on starting point of iteration, e.g., choose the background x_b as the first guess.



.

Other Remarks

- **B** matrix is of very large dimension, explicit inverse of **B** is impossible, substantial efforts in data assimilation were given to the estimation and modeling of **B**.

- **B** shall be spatially-varied and time-evolving according to weather regime.

- Analysis can be sub-optimal if using inaccurate estimate of **B** and **R**.

- Could use non-Gaussian PDF
  - Thus not a least square cost function
  - Difficult (usually slow) to solve; could transform into Gaussian problem via variable transform

Variational vs. Ensemble DA

- They are solving the same cost function, by using different techniques

- These days, combining both techniques are common at operational centers
  - NOAA/NCEP: hybrid-4DEnVar + LETKF
  - ECMWF: ensemble of 4DVar
  - UKMO: hybrid-4DVar + LETKF

Further reading