Fundamentals of Data Assimilation

Tom Auligné

National Center for Atmospheric Research, Boulder, CO USA

GSI Data Assimilation Tutorial - June 28-30, 2010

Acknowledgments and References

- WRFDA Overview (WRF Tutorial Lectures, H. Huang and D. Barker)
- Data Assimilation concepts and methods (ECMWF Training Course, F. Bouttier and P. Courtier)
- Data Assimilation Research Testbed (DART) Tutorial (J. Anderson et al., http://www.image.ucar.edu/DAReS/DART)
- Analysis methods for numerical weather prediction (A.C. Lorenc, 1986, *Quart. J. R. Meteorol. Soc.*
- Data Assimilation: aims and basic concepts (Data Assimilation for the Earth System, NATO Science Series, N. Nichols, R. Swinbank)
- Atmospheric Data Analysis (R. Daley, 1991, *Cambridge University Press, 457 pp.*)
- Atmospheric Modeling, Data Assimilation and Predictability (E. Kalnay, 2003, *Cambridge University Press, 341 pp.*)

Table of contents

Motivation

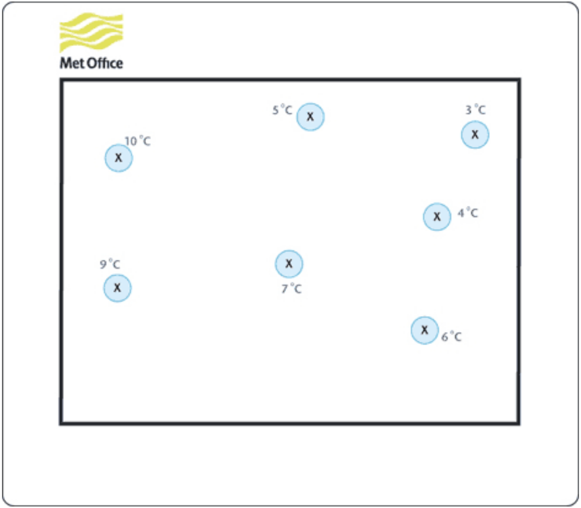
- *A sufficiently accurate knowledge of the state of the atmosphere at the initial time.*
  (Today's weather)
- *A sufficiently accurate knowledge of the laws according to which one state of the atmosphere develops from another.*
  (Tomorrow's weather)

Vilhelm Bjerknes (1904)
(Peter Lynch)

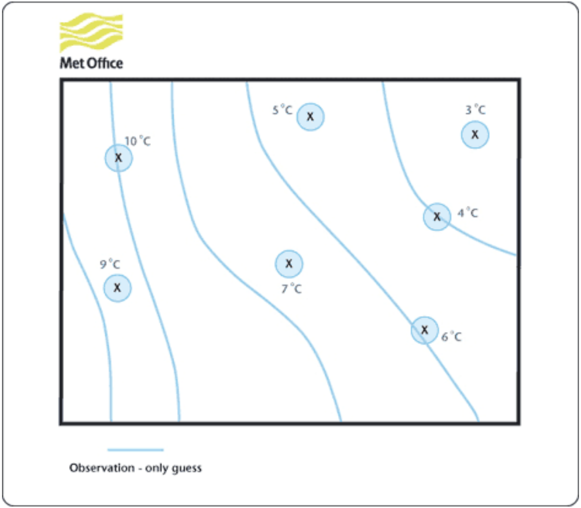
Motivation

- Initial conditions for Numerical Weather Prediction (NWP)
- Calibration and validation
- Observing system design, monitoring and assessment
- Reanalysis
- Better understanding (Model errors, Data errors, Physical process interactions, *etc*)

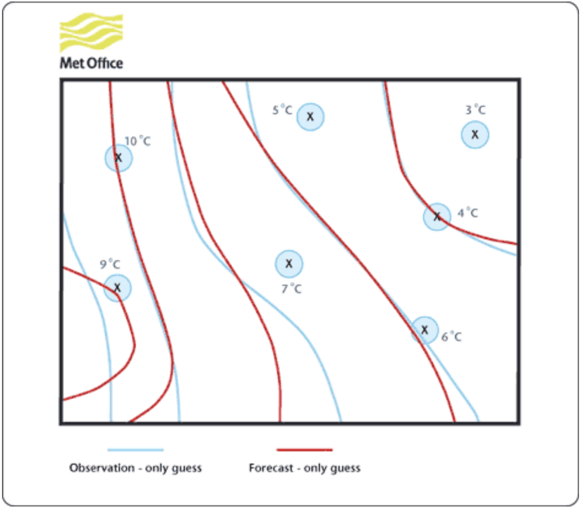
Motivation



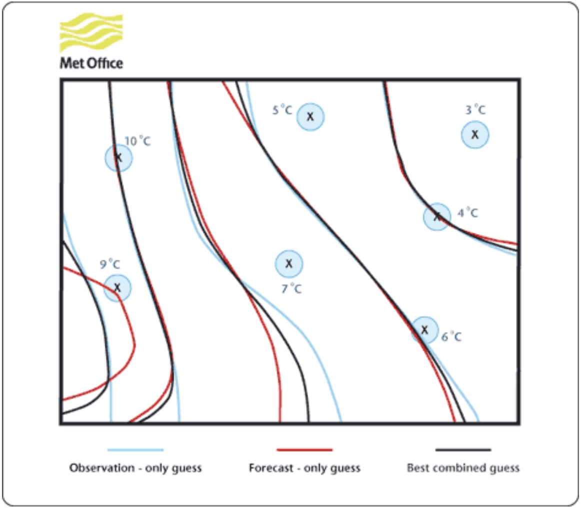
Motivation



Motivation


Met Office
10 °C
5 °C
3 °C
4 °C
9 °C
7 °C
6 °C
Observation - only guess
Forecast - only guess

Motivation


Met Office
10 °C
9 °C
5 °C
7 °C
3 °C
4 °C
6 °C
Observation - only guess
Forecast - only guess
Best combined guess

From Empirical to Statistical methods

- Successive Correction Method (SCM, *Cressman 1959*)
  Each observation within a radius of influence L is given a weight w varying with the distance r to the model grid point: $w(r) = \frac{L^2 - r^2}{L^2 + r^2} (r \leq L)$
- Nudging
- Physical Initialization (PI), Latent Heat Nudging (LHN)

However...

- Relaxation functions are somewhat arbitrary
- **Good** forecast can be replaced by **bad** observations
- Noisy observations can create unphysical analysis

So...

Modern DA techniques are usually statistical

What is the temperature in this room?

Notations

- x_t: "True" state
- x_o: Observation
- x_b: Background information
- $d = x_o - x_b$: Innovation or *Departure*
- x_a: Analysis ("optimal" in RMSE sense)

Hypotheses

- Observation and Background errors are uncorrelated, unbiased, normally distributed, with variance σ_o^2 and σ_b^2
- Linear Analysis: $x_a = \alpha x_o + \beta x_b = x_b + \alpha(x_o - x_b)$

Best Linear Unbiased Estimate

The analysis value is $x_a = x_b + \alpha(x_o - x_b)$ and its error variance:

$$\sigma_a^2 = \overline{(x_a - x_t)(x_a - x_t)} = (1-\alpha)^2\sigma_b^2 + \alpha^2\sigma_o^2$$

$$\frac{\partial \sigma_a^2}{\partial \alpha} = 2\alpha(\sigma_b^2 + \sigma_o^2) - 2\sigma_b^2 = 0 \quad \Rightarrow \quad \alpha = \frac{\sigma_b^2}{\sigma_b^2 + \sigma_o^2}$$

Best Linear Unbiased Estimate (BLUE)

$x_a = x_b + B(B+R)^{-1}(x_o - x_b)$ and $A^{-1} = B^{-1} + R^{-1}$

with $A = \sigma_a^2, \quad B = \sigma_b^2, \quad R = \sigma_o^2$

Statistically, the analysis is better than:

- the observation ($A < R$),
- the background ($A < B$).

Variational Cost Function

This solution is equivalent to minimizing the cost function:

$$J(x) = \frac{1}{2}(x - x_b)^T B^{-1}(x - x_b) + \frac{1}{2}(x - x_o)^T R^{-1}(x - x_o) = \mathbf{J_b} + \mathbf{J_o}$$

Proof:

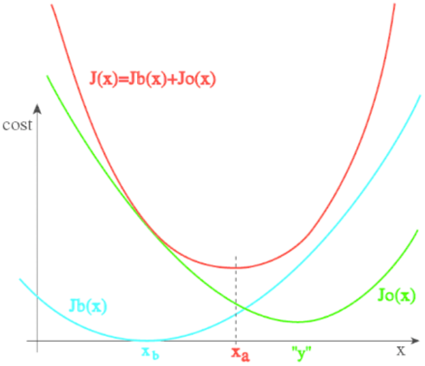
$$\nabla J = B^{-1}(x - x_b) + R^{-1}(x - x_o) = 0$$

$$\begin{aligned} \Rightarrow x_a &= x_b + \frac{\sigma_b^2}{\sigma_b^2 + \sigma_o^2}(x_o - x_b) \\ &= x_b + K(x_o - x_b) \end{aligned}$$

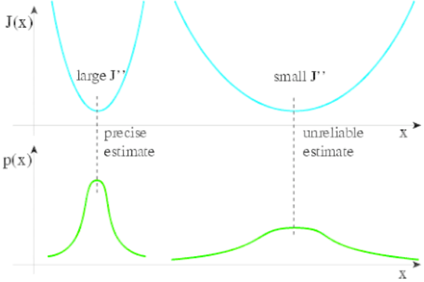
with K being the Kalman Gain:

$$K = B(B + R)^{-1}$$

Analysis Accuracy



from Bouttier and Courtier 1999



Quality of the Analysis

The precision is defined by the convexity or **Hessian** $A = J''^{-1}$

Conditional Probabilities

According to Bayes Theorem, the joint pdf of x and x_o is:

$$P(x \wedge x_o) = P(x|x_o)P(x_o) = P(x_o|x)P(x)$$

Since $P(x_o) = 1$, $P(x|x_o) = P(x_o|x)P(x)$

We assumed the background and observation errors were Gaussian:

$P(x) = \lambda_b e^{[\frac{1}{2\sigma_o^2}(x_b - x)^2]}$ and $P(x_o|x) = \lambda_o e^{[\frac{1}{2\sigma_o^2}(x_o - x)^2]}$

$$\Rightarrow P(x|x_o) = \lambda_a e^{[\frac{1}{2\sigma_o^2}(x_o - x)^2 + \frac{1}{2\sigma_o^2}(x_b - x)^2]} = \lambda_a e^{-J(x)}$$

Maximum Likelihood

The minimum of the cost function J is also the estimator of x_t with the maximum likelihood

Partial Conclusions

Under the aforementioned hypotheses, the BLUE:

- can be determined analytically through the Kalman gain K
- is also the minimum of a cost function $J = J_b + J_o$
- is optimal for minimum variance **and** maximum likelihood

Sequential Data Assimilation

Forecast model $M_{i \rightarrow i+1} = M$ from step i to $i+1$

$$x_{i+1}^t = M(x_i^t) + q_i$$

where q_i is the model error. As q_i is unknown and x_i^a is the best estimate of x_i^t, usually: $x_{i+1}^f = M(x_i^a)$

Forecast error

$$x_{i+1}^f - x_{i+1}^t = M(x_i^a) - M(x_i^t) - q_i \approx \mathbf{M}_i(x_i^a - x_i^t) - q_i$$

M is called the **Tangent-Linear** code of the non-linear model M

Forecast error covariance matrix

$$P_{i+1}^f \approx \mathbf{M}_i \overline{(x_i^a - x_i^t)(x_i^a - x_i^t)^T} \mathbf{M}_i + \overline{q_i q_i^T} = \mathbf{M}_i P_i^a \mathbf{M}_i^T + Q_i$$

Sequential Data Assimilation

We can use the forecast as background for the BLUE calculation

$$K_i = P_i^f (P_i^f + R)^{-1}$$

$$x_i^a = x_i^f + K(x_i^o - x_i^f)$$

$$(P_i^a)^{-1} = (P_i^f)^{-1} + R^{-1} \Rightarrow P_i^a = (I - K_i)P_i^f$$

Finally, we can distinguish the model space x from the observation space y and introduce an Observation Operator $H : x \mapsto y$, which is linearized: $H(x_i^a) - H(x_i^t) \approx \mathbf{H}(x_i^a - x_i^t)$

$$K_i = P_i^f \mathbf{H}_i^T (\mathbf{H}_i P_i^f \mathbf{H}_i^T + R)^{-1}$$

$$x_i^a = x_i^f + K(y_i^o - x_i^f)$$

$$P_i^a = (I - K_i \mathbf{H}_i)P_i^f$$

The Extended Kalman Filter Algorithm

Analysis step i:

$$K_i = P_i^f \mathbf{H}_i^T [\mathbf{H}_i P_i^f \mathbf{H}_i^T + R]^{-1} \quad (1)$$

$$x_i^a = x_i^f + K_i [y^o - H x_i^f] \quad (2)$$

$$P_i^a = [I - K_i \mathbf{H}_i] P_i^f \quad (3)$$

Forecast step from i to $i+1$:

$$x_{i+1}^f = M(x_i^a) \quad (4)$$

$$P_{i+1}^f = \mathbf{M}_i P_i^a \mathbf{M}_i^T + Q_i \quad (5)$$

Hypotheses

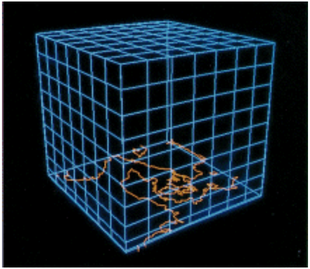
- Gaussian distributions of errors
- **M**: Linearization around non-linear Model M
- **H**: Linearization around non-linear Observation Operator H

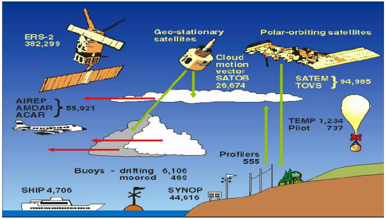
From scalar to vector: dimensions

$x \rightarrow \mathbf{x}$

Number of grid points $\approx 10^7$

Dimension of P^f, $P^a \approx 10^7 \times 10^7$





$y^o \rightarrow \mathbf{y^o}$

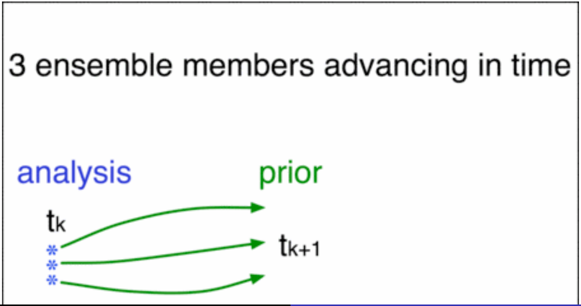
Number of observations $\approx 10^6$

Dimension of $R \approx 10^6 \times 10^6$

Ensemble Kalman Filter (EnKF)

Hypotheses

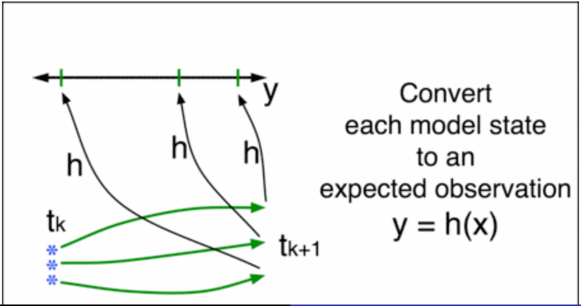
- Monte Carlo approximation to pdfs
- Gaussian distributions used for computing update
- Localization in space: for each model grid point, only a few observations are used to compute the analysis increment.



Ensemble Kalman Filter (EnKF)

Hypotheses

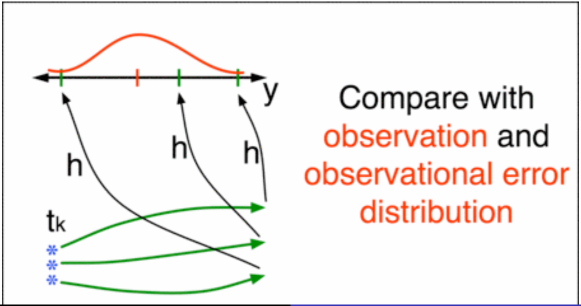
- Monte Carlo approximation to pdfs
- Gaussian distributions used for computing update
- Localization in space: for each model grid point, only a few observations are used to compute the analysis increment.



Ensemble Kalman Filter (EnKF)

Hypotheses

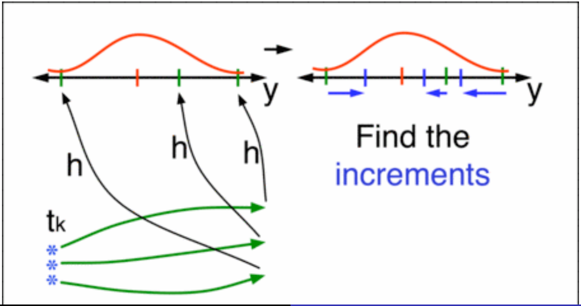
- Monte Carlo approximation to pdfs
- Gaussian distributions used for computing update
- Localization in space: for each model grid point, only a few observations are used to compute the analysis increment.



Ensemble Kalman Filter (EnKF)

Hypotheses

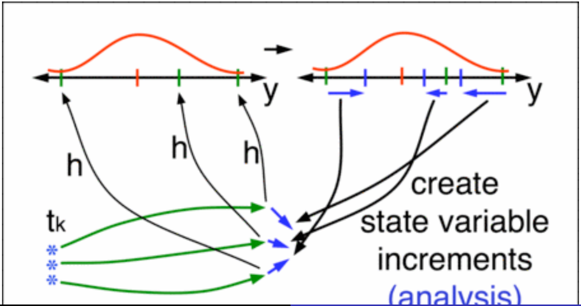
- Monte Carlo approximation to pdfs
- Gaussian distributions used for computing update
- Localization in space: for each model grid point, only a few observations are used to compute the analysis increment.



Ensemble Kalman Filter (EnKF)

Hypotheses

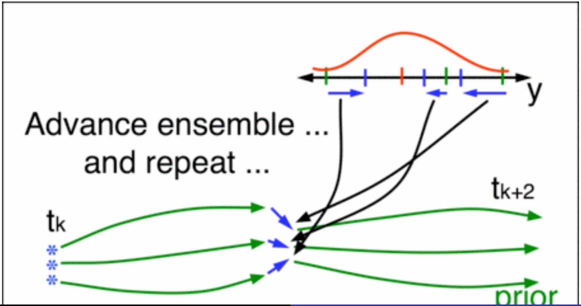
- Monte Carlo approximation to pdfs
- Gaussian distributions used for computing update
- Localization in space: for each model grid point, only a few observations are used to compute the analysis increment.



Ensemble Kalman Filter (EnKF)

Hypotheses

- Monte Carlo approximation to pdfs
- Gaussian distributions used for computing update
- Localization in space: for each model grid point, only a few observations are used to compute the analysis increment.



Ensemble Kalman Filter (EnKF)

Hypotheses

- Monte Carlo approximation to pdfs
- Gaussian distributions used for computing update
- Localization in space: for each model grid point, only a few observations are used to compute the analysis increment.

Advantages

- Easy to implement and provides estimate of Analysis Accuracy
- H and M need not be linearized

Drawbacks

Localization avoids degeneracy from under-sampling and reduces spurious noise, but it affects model internal balance

Optimal Interpolation (OI)

Hypotheses

- $P^f \approx B$ defined climatologically via empirical transforms (*e.g.* balance constraints, autocorrelation functions)
- Localization in space
- Analytical Kalman gain $K = B\mathbf{H}^T(\mathbf{H}B\mathbf{H}^T + R)^{-1}$

Advantages

Cheap and easy implementation

Drawbacks

- $B\mathbf{H}^T$ becomes difficult for complexe observation operators
- Possible incoherence of analysis between scales

3D Variational Data Assimilation (3DVar)

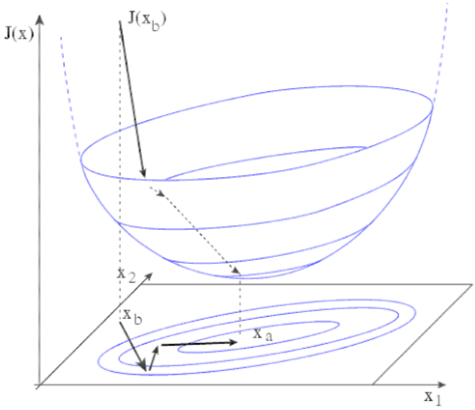
Hypotheses

Avoid calculating K by solving the equivalent minimization problem defined by the cost function:
$J(x) = \frac{1}{2}(x - x_b)^T B^{-1}(x - x_b) + \frac{1}{2}(y^o - H(x))^T R^{-1}(y^o - H(x))$

$$\nabla J(x) = B^{-1}(x - x_b) - \mathbf{H}^T R^{-1}[y - H(x)]$$

$\mathbf{H}^T$ is called the **Adjoint** of the linearized observation operator

3D Variational Data Assimilation (3DVar)



from Bouttier and Courtier 1999

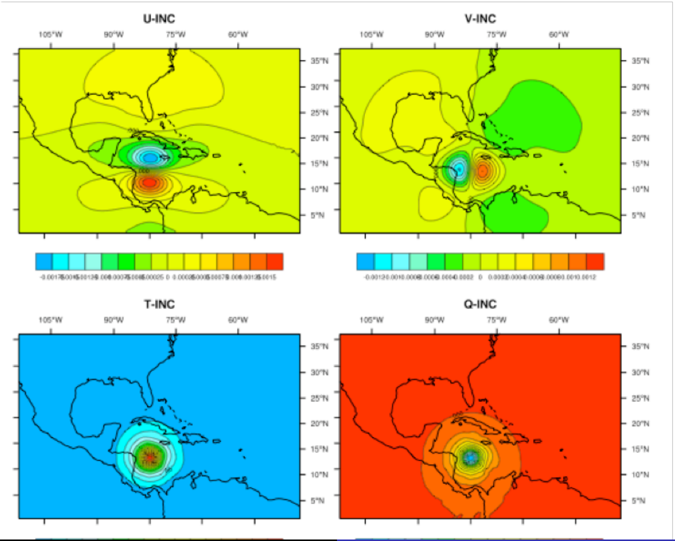
Minimization Algorithm

- Iterative minimizer
  $\rightarrow$ several simulations
- Steepest Descent, Quasi-Newton, Conjugate Gradient, *etc*

Preconditioning

- Improve Condition Nb
- Faster convergence

Single Observation Experiment



3D Variational Data Assimilation (3DVar)

Hypotheses

- Avoid calculating K by solving the equivalent minimization problem defined by the cost function:
  $J(x) = \frac{1}{2}(x - x_b)^T B^{-1}(x - x_b) + \frac{1}{2}(y^o - H(x))^T R^{-1}(y^o - H(x))$

Advantages

- Easy to use with complex observation operators
- Can add external weak or *penalty* constraints J_c

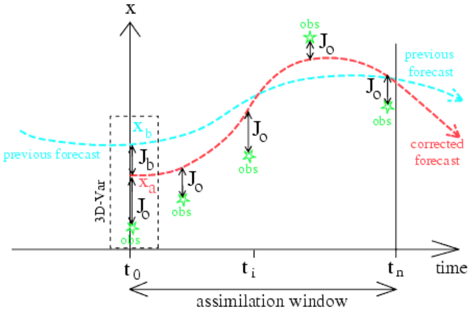
Drawbacks

- Sub-optimal for strongly non-linear observation operators
- All observations are assumed to be instantaneous

4D Variational Data Assimilation (4DVar)

Hypotheses

- Generalization of 3DVar for observations distributed in time
- Analysis variable x defined at the **beginning** of time window
- Find model trajectory minimizing the distance to observations



4D Variational Data Assimilation (4DVar)

Hypotheses

- Generalization of 3DVar for observations distributed in time
- Analysis variable x defined at the **beginning** of time window
- Find model trajectory minimizing the distance to observations

The Cost Function becomes:

$$J(x) = \frac{1}{2}(x-x_b)^T B^{-1}(x-x_b)+\frac{1}{2}(y^o-HM(x))^T R^{-1}(y^o-HM(x))$$

$$\nabla J(x) = B^{-1}(x-x_b) - \mathbf{M}^T\mathbf{H}^T R^{-1}[y-HM(x)]$$

$\mathbf{M}^T$ is called the **Adjoint** of the linearized forecast model

4D Variational Data Assimilation (4DVar)

Hypotheses

- Generalization of 3DVar for observations distributed in time
- Analysis variable x defined at the **beginning** of time window
- Find model trajectory minimizing the distance to observations

Advantages

Model internal balance is more prone to be respected

Drawbacks

- The development and maintenance of the Adjoint model $\mathbf{M}^T$ can be cumbersome
- Limitation of the "perfect model" assumption

Conclusion

- Observations y^o
- Background x_b
- Observation Operator H

(Extended) Kalman Filter (quasi-)linear statistical DA algorithm

Simplifications for practical implementation

- Ensemble methods: EnKF
- Optimal Interpolation: OI
- Variational methods: 3DVar, 4DVar

Challenges for Data Assimilation

- Modeling of Background and Observation error covariances
- Correction of systematic errors in the Model and Observations
- Processing of data to approximate Gaussianity (Quality Control)
- Accounting for Model errors and Non-Linearities
- ...