The Abdus Salam
International Centre for Theoretical Physics

United Nations
Educational, Scientific
and Cultural Organization

International Atomic
Energy Agency

Spring Colloquium on
'Regional Weather Predictability and Modeling'
April 11 - 22, 2005

1) Workshop **on Design and Use of Regional Weather Prediction Models,** *April 11 - 19*

2) Conference on **Current Efforts Toward Advancing the Skill of Regional Weather Prediction. Challenges and Outlook,** *April 20 - 22*

301/1652-3

Data Assimilation in Regional Modeling & Prediction
Lecture I
Introduction to Data Assimilation

T. Vukicevic
Cooperative Institute for Research in the Atmosphere, CSU, Ft. Collins
&
Program in Atmospheric and Oceanic Sciences, CU, Boulder, USA

Strada Costiera 11, 34014 Trieste, Italy - Tel. +39 040 2240 111; Fax +39 040 224 163 - sci_info@ictp.it, www.ictp.it

Data Assimilation in Regional Modeling and Prediction

Lecture I

Introduction to Data Assimilation

Dr. Tomislava Vukicevic

Affiliations:

Cooperative Institute for Research in the Atmosphere, CSU, Ft. Collins and Program in Atmospheric and Oceanic Sciences, CU, Boulder, USA

E-mail tomi@cira.colostate.edu

Outline

- Motivation
- History
- Methodology

Start with motivation from "big picture"

Why model and predict weather and climate?

Modeling

- Understand natural environment
  - Diagnose interactions and past states
- Evaluate resources
- Enable objective prediction of future states

Prediction

- Prevent loss of human life and material damage by severe weather
- Help planning of transport and other services
- Aid in environmental and energy planning policies

- Value of modeling and prediction with respect to societal and scientific interest critically depends on

  degree of accuracy of information

  that is contained in the modeled and predicted states

- Therefore, it is important to address the following questions

  1. How is the accuracy evaluated?
  2. What controls the accuracy?

- Regarding question 1 > It is obvious that verification of the accuracy of modeled and predicted state must involve observations

- Regarding question 2 > The accuracy in weather modeling which solves open dynamical system governing equations depends on
  - initial condition
  - boundary conditions
  - external forcing
  - "free" parameters
  - model error

Data Assimilation

- To achieve accurate modeled and predicted system states the control parameters must be evaluated accurately

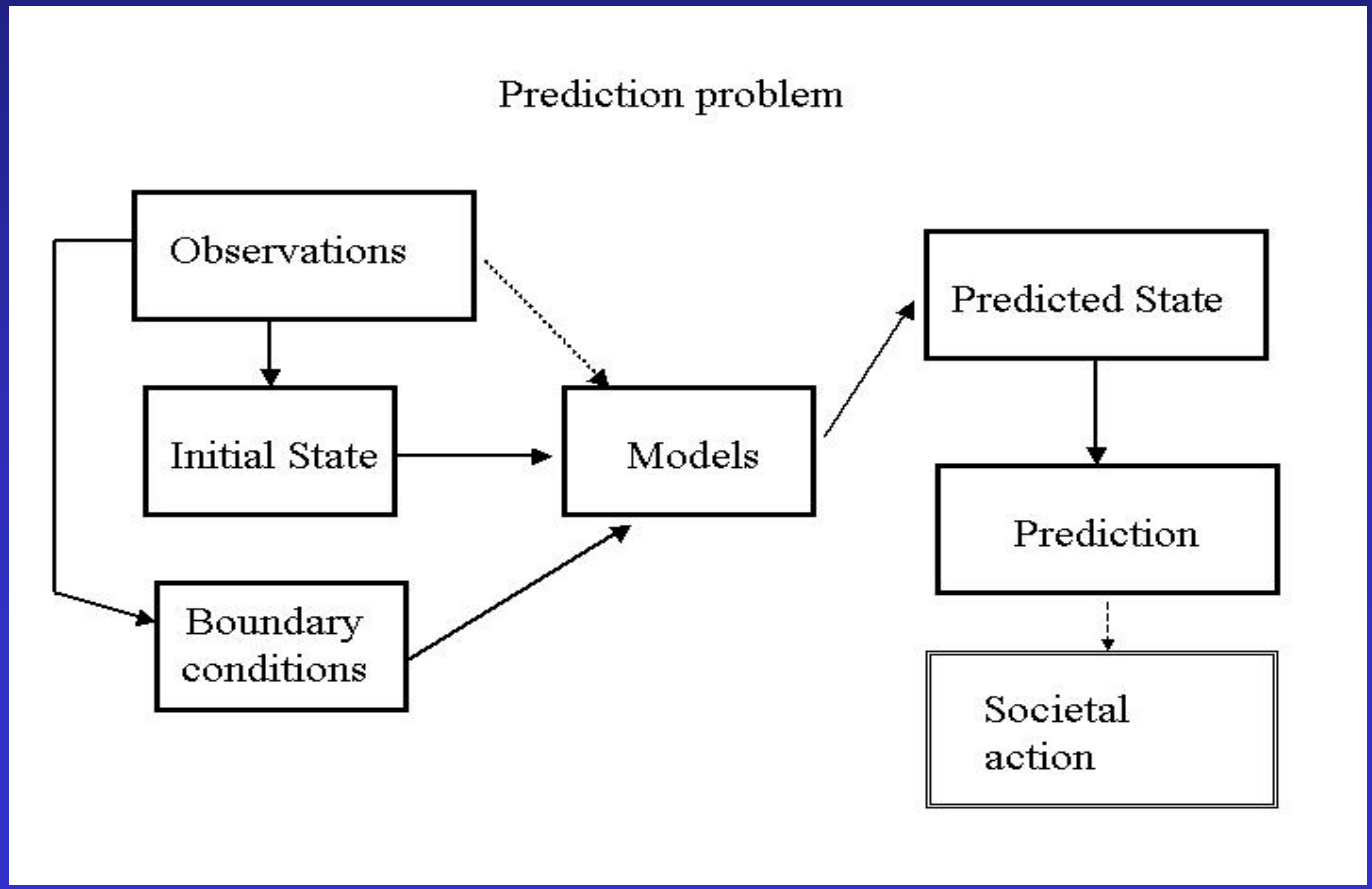
The data assimilation is

Estimation of the control parameters by "combining" information from

observations and models

In summary

Motivation for the data assimilation is desire to accurately estimate what controls the accuracy of modeled and predicted states using observations



History of the data assimilation methodology

Gauss wrote in his *Theory of motions of the heavenly bodies* (1809)

If the astronomical observations and other quantities on which the computation of orbits is based were absolutely correct, the elements also, whether deduced from three or four observations, would be strictly accurate (so far indeed as the motion is supposed to take place exactly according to the laws of Kepler) and, therefore, if other observations were used, they might be confirmed but not corrected. But since our observations are nothing more than approximations to the truth, the same must be true of all calculations resting upon them, and the highest aim of all computations made concerning concrete phenomena must be approximate, as nearly as practicable, to the truth. But this can be accomplished in no other way than by a suitable combination of more observations than the number absolutely requisite for the determination of the unknown quantities. This problem can only be properly undertaken when an approximate knowledge of the orbit has been already attained, which is afterward to be corrected so as to satisfy all of the observations in the most accurate manner possibly.

History of the data assimilation in the weather modeling and prediction is much more recent

- Weather prediction as we know it today started with Richardson's work (1965, original 1922)
- After the second world war the numerical weather prediction methods were fast developing in the US and Europe
- The data assimilation, however, started to develop just recently in late 1980-es with applications of so called estimation and control theories for dynamical systems in the weather analysis.
- The estimation and control theories were originally developed in applied mathematics for purpose of addressing engineering and signal processing problems

What was done before 1990-es to produce what we defined as the control parameters?

- The weather data analysis started since Richardson's famous first NWP attempt
- The observations are analyzed in space and sequentially in time without explicit use of the modeled dynamics
- The weather data analysis addresses only the evaluation of the initial conditions
- Other control parameters are derived mostly from variety of long term measurements and educated guesses

- It will be discussed at the end that in the current data assimilation techniques many of the control parameters are still evaluated using the same approach as in the weather data analysis

Data assimilation methodology

Modeled information

- State representation in models

$$X \equiv (x_1, x_2, \ldots, x_n)$$

Each component is a characteristic such as, for example, temperature, wind, optical depth, trace gas concentration, LAI, etc

Probabilistic information

- Let us adopt the view that the information is most complete when represented as probability

$$p(X)$$

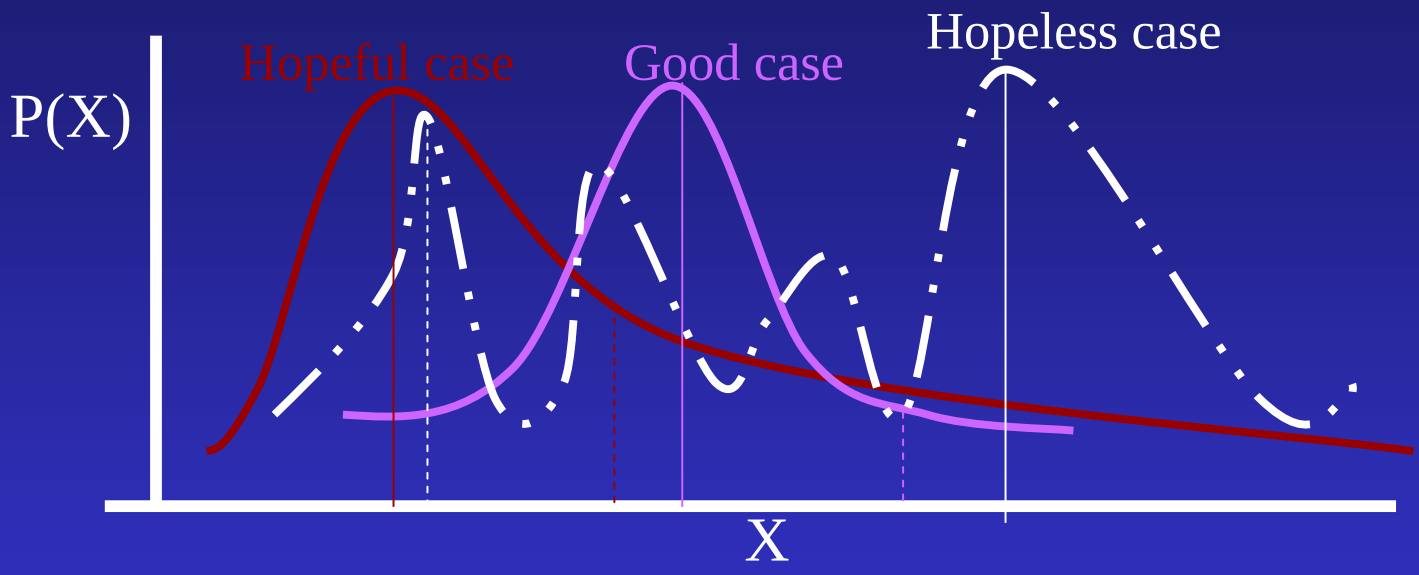
The probability of X being true

Simple example

Relationship between probability of information and accuracy

- Probability of X being true has to be high for X to be accurate
- Which implies that p(X) must have "small spread"
- P(X) should have these properties for all degrees of freedom in the modeling/prediction prblem

Generic examples



Schematic examples of possible probability distribution for X

Probabilistic information in observations

- Observations are not exactly the truth, they are another information characterized with

$$p(Y)$$

Key property

Observations are by design closer to truth than any other information

Information about the Atmospheric system

Desired state components

Temperature (N)

Wind (N)

Pressure (N)

Humidity (N)

Cloud properties (N*k)

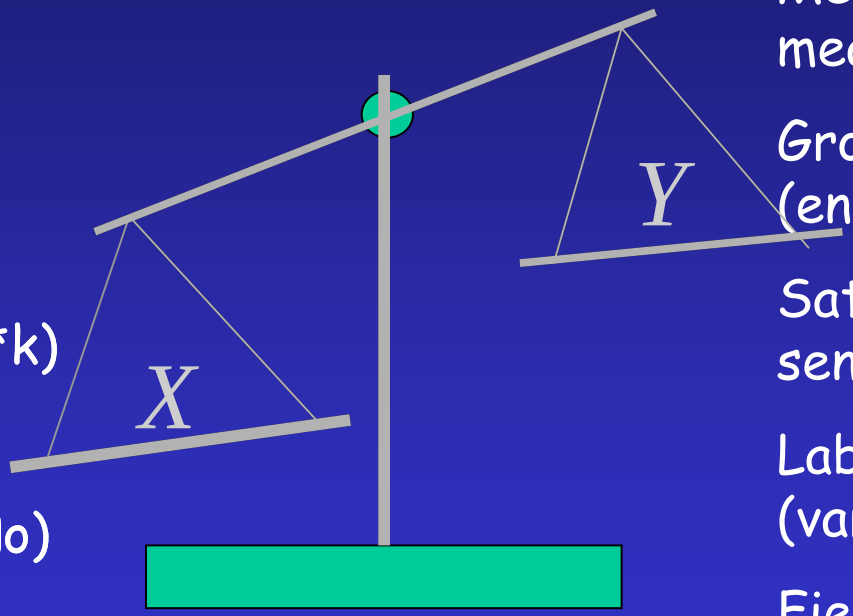
Soil properties (Ns)

Ocean properties (No)

Aerosols (N*j)

Trace gases (N*i)

Process Parameters (N*p)



N >> M

Measured state components

Meteorological station direct measurements (M1)

Ground based remote sensing (energy; M2)

Satellite remote sensing(energy; M3)

Laboratory measurements (variety, small volume; M4)

Field measurements (variety, small sample; M5)

Geological direct or indirect measurements (variety; samples; M6)

"Ground truth"

- The observations do not span the desired (modeled) large space of information about the system state
- The reference truth in the observations is, for most part, in different quantities than the modeled

Combining information from models and observations

Bayes' rule

observations

Posterior $p(X^t / Y) = \dfrac{p(Y / X^t)\, p(X^t / X)}{\int p(Y / X^t)\, p(X^t / X)\, dX^t}$ Prior (model)

state after assimilation

- Difficult to evaluate explicitly but the rule provides exact formula for making progress

Data assimilation problem in probabilistic formulation

- Find best solution for (X^t / Y) , where the best is defined by one of the following criteria on the posterior probability
  - Minimum variance
  - Maximum likelihood
  - Maximum entropy reduction

Solutions for the best estimate given the criterion

- Mean of distribution is the solution for the minimum variance criterion

- Maximum P(X) is the solution for the maximum likelihood criterion

- Entropy reduction: the change in the logarithm of the number of distinct possible internal states of the system being observed, consistent with the change in knowledge of the system resulting from observations

entropy

$$S(P) = -\int P(X^t / Y) \ln\left[P(X^t / Y) / M(X)\right] dX$$

- Most commonly used criteria are minimum variance and maximum likelihood

- Next we discuss how to transform the problem that is defined in terms of the probability distributions to what is much more familiar: the control parameters in weather models (initial and boundary conditions, free parameters, model error, etc)

Assumption for probability distribution

- To apply Bayes' rule the distribution functions have to be assumed for

$p(Y / X^t)$ and $p(X^t / X)$

$p(X^t / Y)$ functional form is obviously the consequence of

$$p(X^t / Y) = \frac{p(Y / X^t)\, p(X^t / X)}{\int p(Y / X^t)\, p(X^t / X)\, dX^t}$$

- Prior/model may or may not not have significant impact depending on the strength of observational constraint and the measurements' information content

$p(X^t / X)$ could be shaped or "flat"

- Probability of observations relative to truth in the observation space is

$p(Y / X^t)$ unimodal with small variance

Most common assumption

Normal or Gaussian distribution

$$p(z) = (2\pi)^{-M/2}\left|C\right|^{-\frac{1}{2}} \exp\left[-\frac{1}{2}(z-\bar{z})^T C^{-1}(z-\bar{z})\right]$$

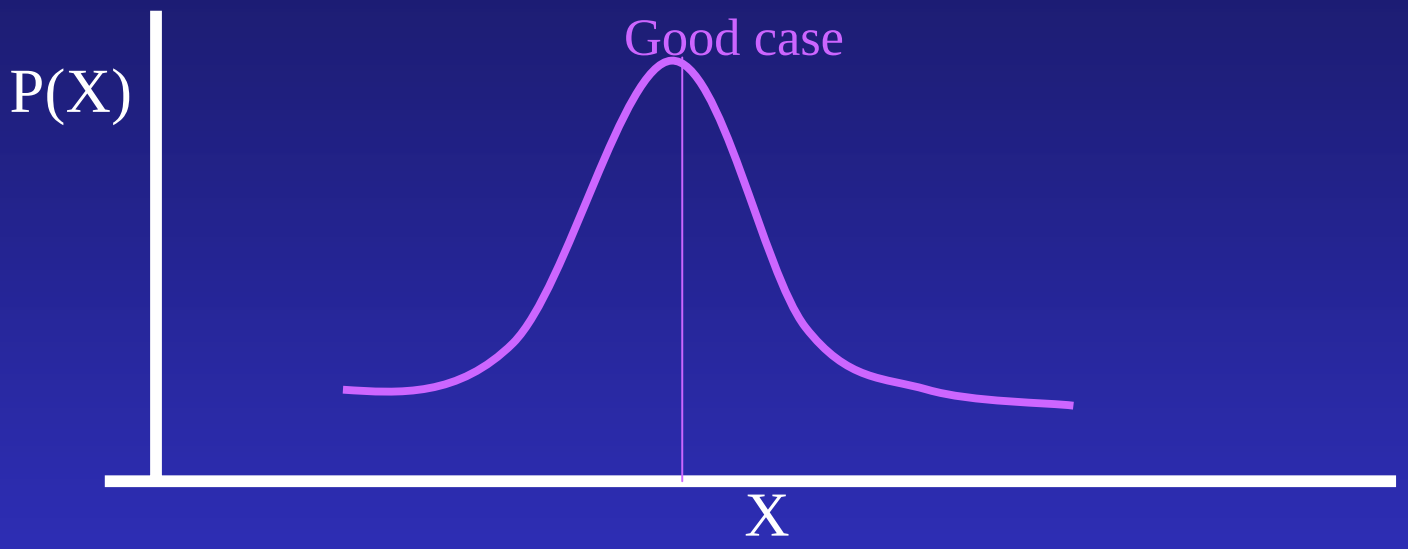
z stands for either (Y / X^t), (X^t / Y) or $(X^t / X\ )$

Consequence

- Given the Gaussian probability distribution (X^t / Y) is found by maximizing the posterior probability which is the same as minimizing the variance and

the solution is exactly the mean of $p(X^t / Y)$

Recall from Generic example



- The Gaussian distribution has the same mean and maximum
- It is characterized with only 2 parameters: mean and covariance matrix

How is the solution derived in terms of control parameters given the Gaussian distribution?

There are two commonly used techniques:

1. Variational (1+N-DVAR)
   1 stands for time and N for spatial dimensions
2. Kalman Filter
   linear, extended and ensemble

In either technique the following applies

$$X(\tau^{n+1}) = \mathrm{M}\left[X(\tau^{n}), \alpha, X_b(\tau^{n+1})\right] + \varepsilon_X\left[X(\tau^{n+1})\right]$$

$$Y(\tau^{k}) = H\left[X(\tau^{k})\right] + \varepsilon_y$$

$X(\tau^k)$ model state at time instance k, the observation time

$Y(\tau^k)$ observation state at the same time

$\varepsilon_X, \varepsilon_Y$ model and observation errors, respectively

H mapping for the model into observation space

M model operator

α and X_b free parameters and boundary conditions, respectively

Under the assumption of the Gaussian probabilities in

$$p(X^t/Y) = \frac{p(Y/X^t)\,p(X^t/X)}{\int p(Y/X^t)\,p(X^t/X)\,dX^t}$$

the product on the rhs will be maximized if the following is minimized

$$F = \frac{1}{2}(H(X^t)-y)^T R^{-1}(H(X^t)-y) + \frac{1}{2}(\zeta^t-\zeta)^T B^{-1}(\zeta^t-\zeta) + \varepsilon_X^T Q^{-1}\varepsilon_X$$

Where R, B and Q are observation, prior and model error covariance matrices, respectively

- In this expression the initial and boundary conditions and free parameters are folded into one prior vector ζ

- In the variational techniques the minimization of F is solved as the control theory problem using algorithms for finding minimum of the function F subject to the control by the initial conditions, boundary conditions, free parameters or model error over a prescribed time period.

- In the techniques based on the Kalman Filter approach the minimum of F is found by updating the initial condition and model error mean state as well as the associated covariance matrices, sequentially over time

There is no time in this lecture to derive the data assimilation algorithms in detail. Please refer to the following references

Le Dimet, F. X, and O. Talagrand, 1986: variational algorithms for analysis and assimilation of meteorological observations. Tellus, **38A**, 97-110.

•Cohn, S. E., 1997: An introduction to estimation theory. *J. Meteor. Soc. Japan,* **75**, 257-288.

•van Leeuwen, P., 2001: An Ensemble Smoother with Error Estimates. Mon. Wea. Rev., **129**, 709–728.

Summary and Notes

- The purpose of data assimilation is to improve accuracy of control parameters in weather models
- The data assimilation methodology always includes information from the modeled time evolution
- Currently used techniques are variational and Kalman Filter based, which both assume that probabilities associated with the modeled/predicted state and observations are Gaussian
- The Gaussian assumption implies that propagation of the state and associated errors in time is quasi-linear
- Model error has been introduced just recently in the set of control factors