Ensemble Based Data Assimilation

Istvan Szunyogh

Texas A&M University
Department of Atmospheric Sciences

Symposium in Memory of Dezső Dévényi
Budapest, June 21, 2010

I will focus on techniques that are currently used or could be used to solve the data assimilation problem for a state-of-the-art research or operational numerical model of the atmosphere

(In essence, I am excluding ensemble particle filters from my discussion)

Outline

1. The Data Assimilation Problem
   - Why Do We Need Data Assimilation?
   - The Observations
2. Mathematical Formulation
   - General Solution Strategy
   - Sequential Data Assimilation: Extended Kalman Filter
   - Ensemble Based Kalman Filters (EnKF)
3. Concluding Remarks

Why Do We Need Data Assimilation?

The purpose of data assimilation:

- In Numerical Weather Prediction (NWP) the purpose of data assimilation is to generate initial conditions for the NWP forecasts. We call the result of data assimilation, which is a representation of the state of the atmosphere on a grid, the **analysis**.
- Analyses are also often used to study the climate (e.g., based on reanalyses) and to study dynamical processes in the atmosphere.

The Challenges in Data Assimilation

- The observation locations typically do not coincide with the model grid points.
- Observations are taken continuously in time, while analyses are typically prepared every 1, 3, or 6 hours.
- The observed physical quantities are not necessarily the same as the model variables. Most importantly, remotely sensed observations (e.g., satellite based observations and radar observations) measure quantities that depends on a complicated integral of the model variables.
- The number of observations is huge. This is a relatively new problem: in the XX. century, one of the main challenges was to obtain an estimate of the atmospheric state based on a limited number of observations.
- Tight operational time constraint

Illustration I: Surface Pressure Observations

While **surface pressure** is a state vector component in almost all models, there are large regions where surface pressure observations are not available: the state of these state vector components has to be inferred from observations of other state vector components and past observations

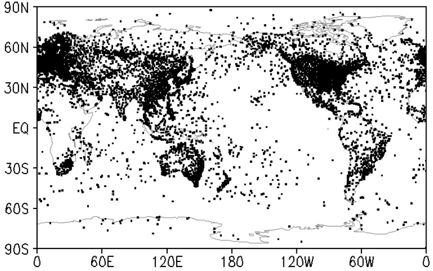


Illustration II: Wind Observations

While the two horizontal components of the **wind vector** are state vector components in all models, there are large regions where wind vector observations (derived from cloud movement) are available for a narrow layer of the atmosphere (the model atmospheres go from the surface to less than 1 hPa)

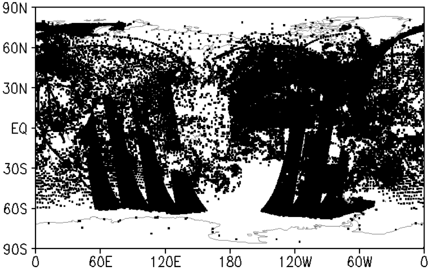
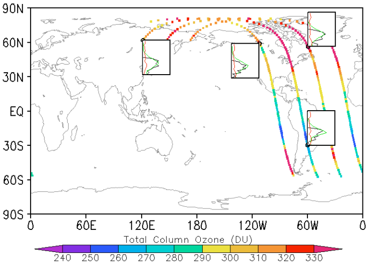


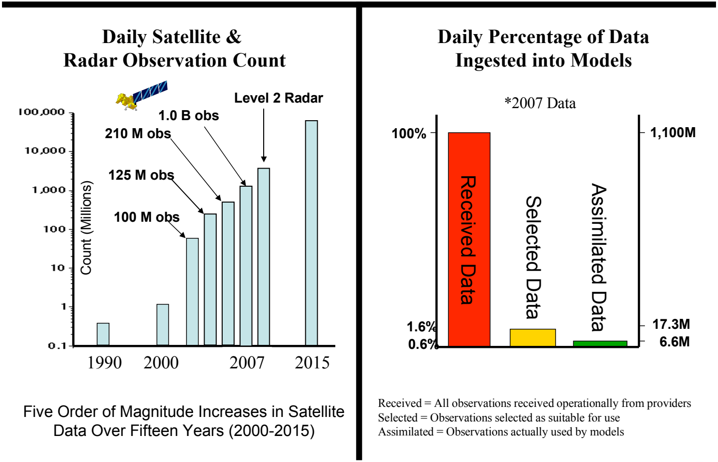
Illustration III: Total Column Ozone Observations Derived from Ultra Violate Backscattering

Ozone concentration is a state vector component in the models, but the satellite based instruments measures radiance, which is related to an integral of a set of model variables.



Daily Data Ingest

from the presentation by S. Lord at the UMD 2007 Data Assimilation Summer School



Daily Satellite & Radar Observation Count

Five Order of Magnitude Increases in Satellite Data Over Fifteen Years (2000-2015)

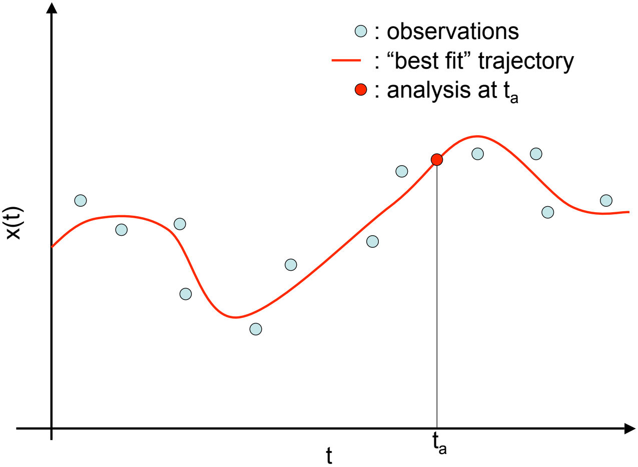
Daily Percentage of Data Ingested into Models

Received = All observations received operationally from providers
Selected = Observations selected as suitable for use
Assimilated = Observations actually used by models

Mathematical Formulation of the Data Assimilation Problem

- Let $\mathbf{x}$ be the m-dimensional vector representing the state of the model at a given time. The time evolution of the model is represented by a trajectory $\{\mathbf{x}(t)\}$ in the m-dimensional state space
- Suppose we are given a **set of noisy observations** of the atmosphere made at various times
- We want to determine which trajectory $\{\mathbf{x}(t)\}$ of the system fits the observations “best”
- The state estimate for a given time t along the trajectory is called the **analysis**

Illustration for m=1: model with one variable



Assumptions

- Assume that the observations are the result of measuring quantities that depend on the system state in a known way, with Gaussian measurement errors.
- An observation at time t_j is a triple $(\mathbf{y}_j^o, \mathcal{H}_j, \mathbf{R}_j)$, where $\mathbf{y}_j^o$ is a vector of observed values, and $\mathcal{H}_j$ and $\mathbf{R}_j$ describe the relationship between $\mathbf{y}_j^o$ and $\mathbf{x}(t_j)$:

$$\mathbf{y}_j^o = \mathcal{H}_j(\mathbf{x}(t_j)) + \varepsilon_j,$$

where ε_j is a Gaussian random variable with mean $\mathbf{0}$ and covariance matrix $\mathbf{R}_j$.

The Least-Square Problem to be Solved

The trajectory of the system that best fits the observations at times $t_1 < t_2 < \cdots < t_n$ can be obtained by solving a least-square problem, an idea originally proposed by **Gauss** (around 1795) and **Legendre** (1805):

- The likelihood of a trajectory $\mathbf{x}(t)$ is proportional to

$$L[\mathbf{x}(t)] = \prod_{j=1}^{n} \exp\left(-\frac{1}{2}[\mathbf{y}_j^o - \mathcal{H}_j(\mathbf{x}(t_j))]^T \mathbf{R}_j^{-1} [\mathbf{y}_j^o - \mathcal{H}_j(\mathbf{x}(t_j))]\right).$$

- The most likely trajectory is the one that maximizes this expression, or equivalently minimizes the “cost function”

$$J^o(\{\mathbf{x}(t)\}) = \sum_{j=1}^{n} [\mathbf{y}_j^o - \mathcal{H}_j(\mathbf{x}(t_j))]^T \mathbf{R}_j^{-1} [\mathbf{y}_j^o - \mathcal{H}_j(\mathbf{x}(t_j))].$$

Potential Approaches to Solve the Formal Problem

- **Variational Approach:** A direct minimization of the cost function
- **Sequential Estimation of the State** (Extended Kalman Filter)
  - A new state estimate is obtained at each analysis time
  - The model dynamics is used to propagate the state estimate between observation times
  - It was originally proposed in a pair of papers by Kalman (1960) and Kalman and Bucy (1961) for linear systems and was extended to nonlinear systems in the late 60's

Rudolf Kalman Receives the National Medal of Science And Technology on October 7, 2009

Sequential Algorithms

All practical methods are sequential schemes, which propagate the state estimate between two analysis times (e.g. 0000 UTC and 0600 UTC, 0600 UTC and 1200 UTC, 1200 UTC and 1800 UTC) with integrating the model:

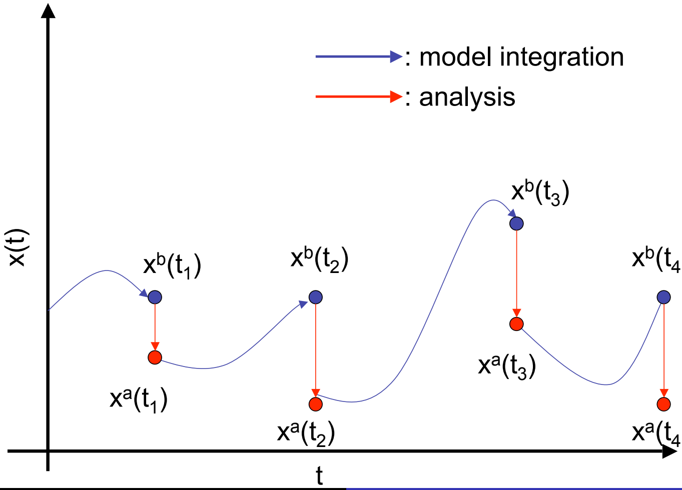
$$\mathbf{x}_n^b = \mathcal{M}_{t_{n-1},t_n}(\mathbf{x}_{n-1}^a),$$

and are equivalent to minimizing the cost function

$$J_{t_n}^o(\mathbf{x}) = [\mathbf{x}-\mathbf{x}_n^b]^T(\mathbf{P}_n^b)^{-1}[\mathbf{x}-\mathbf{x}_n^b] + [\mathbf{y}_n^o - \mathcal{H}_n(\mathbf{x})]^T\mathbf{R}_n^{-1}[\mathbf{y}_n^o - \mathcal{H}_n(\mathbf{x})].$$

- $\mathbf{x}_{n-1}^a$ is the analysis at time t_{n-1} and
- $\mathbf{x}_n^b$ is the **background** , the **background error covariance matrix $\mathbf{P}_n^b$** represents the uncertainty in the state estimate $\mathbf{x}_n^b$

Illustration of the Sequential Data Assimilation for One Model Variable



Extended Kalman Filter

- **State Update Equation:**

$$\mathbf{x_n^a} = \mathbf{x_n^b} + \mathbf{K}(\mathbf{y_n^o} - \mathbf{H_n x_n^b}).$$

- where the **Kalman Gain Matrix K** is

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

- and the **Background Error Covariance** matrix is evolved with

$$\mathbf{P}_n^b = \mathbf{M}_{t_{n-1},t_n}\mathbf{P}_{n-1}^a, \mathbf{M}_{t_{n-1},t_n}^T.$$

where $\mathbf{M}_{t_{n-1},t_n}$ is the linearization of the dynamics $\mathcal{M}_{t_{n-1},t_n}$

- The error in the analysis is described by the **Analysis Error Covariance Matrix**

$$\mathbf{P}^a = (\mathbf{I} - \mathbf{K}\mathbf{H})\mathbf{P}^b$$

Main Obstacles to Implement a Straight Kalman Filter on a High-Dimensional System of Multiple Timescales

- Prohibitive computational cost of evolving the background error covariance matrix (was considered a show-stopper until about a decade ago)
- Unbounded error growth at the fast time scales in the linearized model (problematic, but can be handled in practice)
- Labor intensive development and maintenance of the linearized model (painful, but doable)
- **Remark:** Practical implementations of the variational approach also face the last two problems (and can handle them)

Practical Approach to Kalman Filtering: Ensemble-Based Kalman Filtering (EnKF)

Ensemble representation of the probability distribution (which is still considered Gaussian)

- Instead of a single analysis an **ensemble of analyses** is prepared, whose mean is $\mathbf{x}^a_{n-1}$ and is consistent with the analysis error covariance matrix

$$\mathbf{P}^a_{n-1} = (\mathbf{I} + \mathbf{P}^b_{n-1}\mathbf{H}^T_{n-1}\mathbf{R}^{-1}_{n-1}\mathbf{H}_{n-1})^{-1}\mathbf{P}^b_{n-1}.$$

- The ensemble members are evolved with the nonlinear model to obtain the **background ensemble** for the next analysis time, which provides the estimates of $\mathbf{x}^b_n$ and $\mathbf{P}^b_n$.

Some Attractive Properties of the EnKF Schemes

- They do not require the development and maintenance of a linearized version of a dynamics, which makes them easy to develop and maintain and they are easily portable (these properties made them the clear favorite of the academic community)
- Although not obvious from what have been said in this talk, they do not require a linearized version of the observation operator (this property makes them very attractive to developers of complex observation operators, e.g., those based on a radiative transfer model)
- These schemes scale well on massively parallel computers (in fact, these schemes would not be practical without such computers)
- They provide a set of ensemble initial conditions,

Some Notable Examples for EnKF Schemes

- **First Published Attempt:** Evensen (1994); **First Correct Formulation:** Houtekamer and Mitchell (1998)
- **Serial Schemes Using Perturbed Observations:** e.g., papers by Houtekamer, Evensen, Snyder, F. Zhang and co-authors
- **Serial Schemes Using a Square-Root Filter:** e.g.,, papers by Whitaker, Hamill, J. Anderson
- **Parallel Estimation of State Vector Components Using a Square-Root Filter:** papers by Ott, Hunt, Szunyogh and coathors—The latest product is the Local Ensemble Transform Kalman Filter (LETKF)
- The current versions of these schemes provide about the same analysis accuracy, but there are important differences in the computational efficiency

Most Important Active Areas of Research in EnKF

- Estimation of model errors with the method of state augmentation and the optimal use of the model error information in the formulation of the analysis equations
- Estimation of observation bias (important for remotely sensed observations
- Accounting for effects of nonlinearity in the dynamics and the observation operator in the EnKF

Concluding Remarks

- EnKF has become a mature technology
- All leading operational centers are testing some flavor of the EnKF and are in the process of implementing hybrid Var-Ensemble schemes (in these schemes the ensemble provides all or part of the background error estimates
- The distinction between variational and ensemble based schemes is becoming rather artificial