Néhány gondolat az adatasszimiláció matematikai modelljének értelmezéséről

Szunyogh István

Texas A&M University
Department of Atmospheric Sciences

OMSz, Budapest, július 26, 2016

Gyuró György és Matyasovszky István, “Matyi” emlékére

Outline

Models

"*The sciences do not try to explain, they hardly even try to interpret, they mainly make models. By a model is meant a mathematical construct which, with the addition of certain verbal interpretations, describes observed phenomena. The justification of such a mathematical construct is solely and precisely that it is expected to work."*– John von Neumann

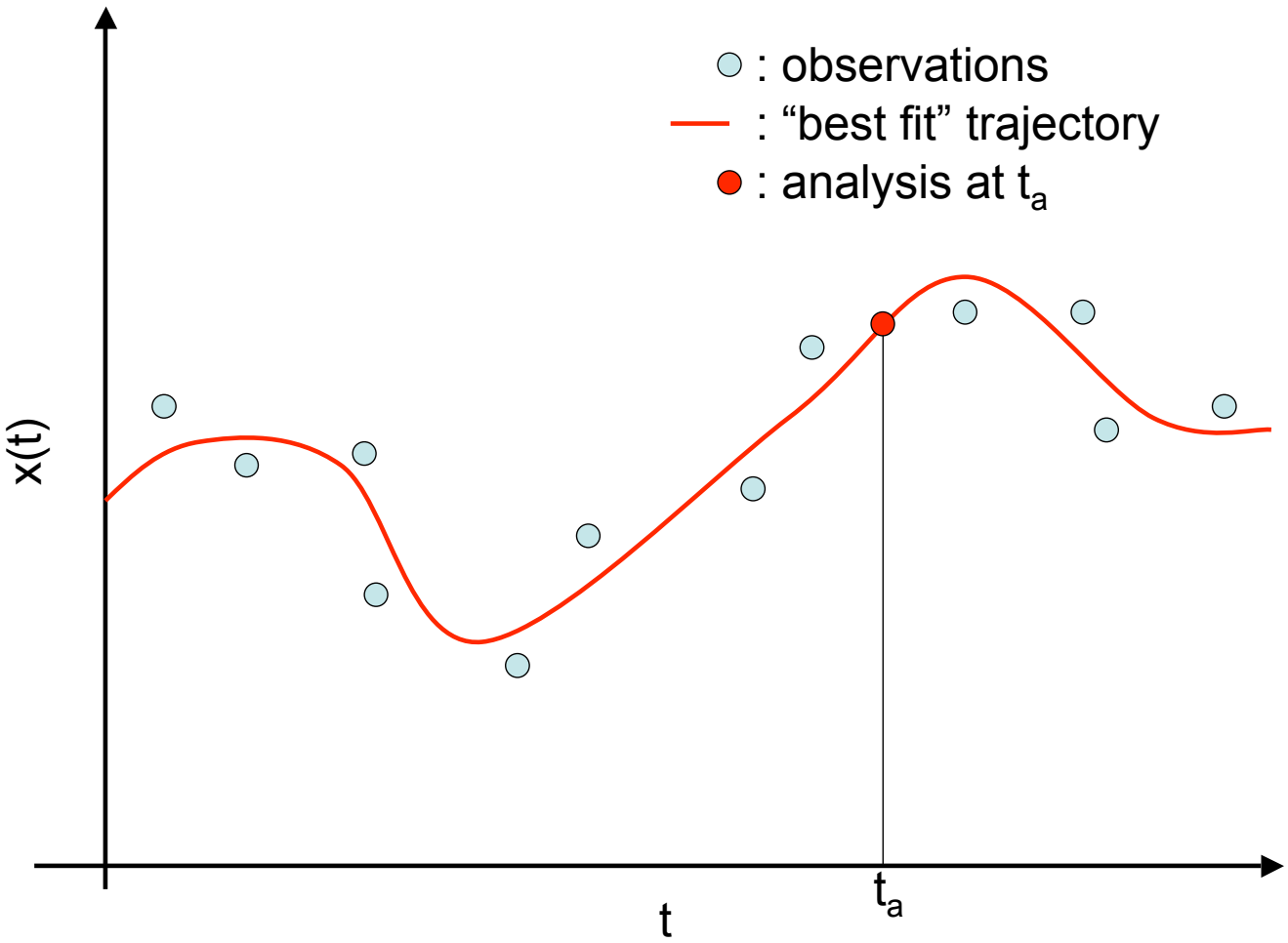
- a spot-on description of the justification of the **mathematical model of data assimilation**–this model is *"expected to work"*, because it has worked for more than 50 years, with steady improvements of the accuracy;
- the "verbal interpretations" are crucial, because they guide our intuition, but they also tend to make us forget that we are working with only a model,

–IS

The Mathematical Model of Data Assimilation

- Let **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



First Assumption

- 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 $\mathbf{x}(t_j)$ by

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

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

Because under this assumption the observations describe the the atmospheric state by a probability distribution, *"best fit"* will refer to *"best fit in a statistical sense"*

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

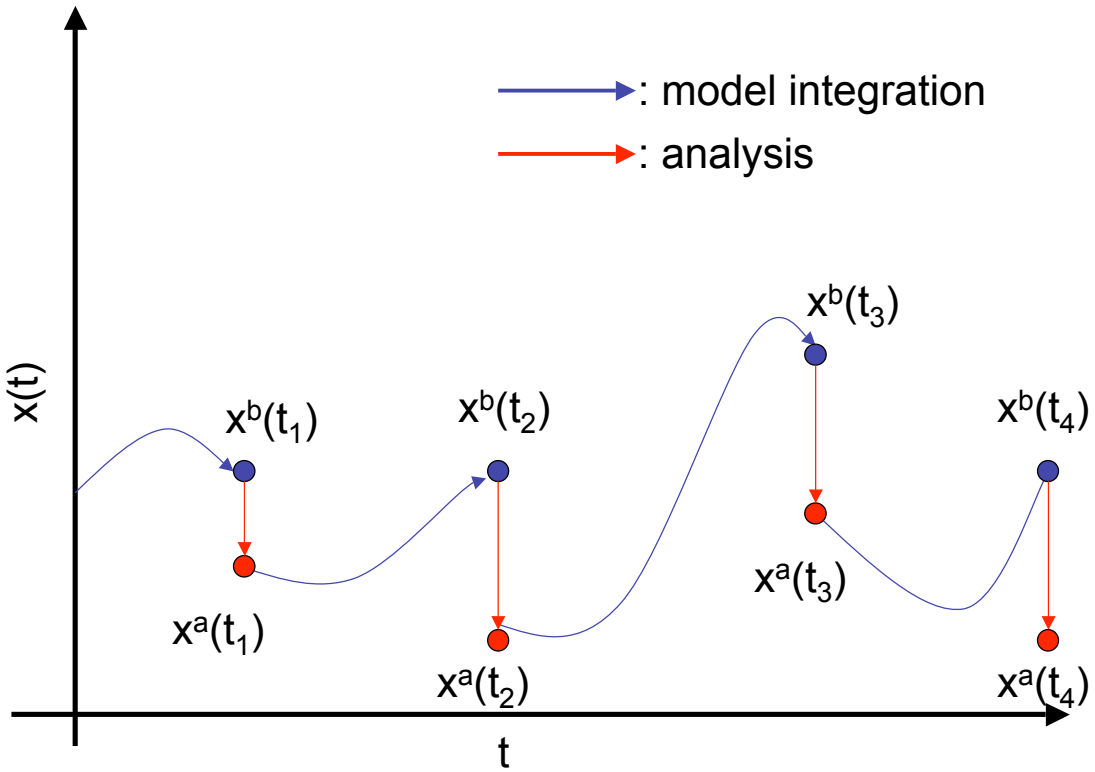
$$L[\mathbf{x}(t)] = \prod_{j=1}^{n} \exp\left(-\frac{1}{2}[\mathbf{y}_j^o - \mathcal{H}_j(\mathbf{x}(t_j))]^T \mathbf{R}_j^{-1} [\mathbf{y}_j^o - \mathcal{H}_j(\mathbf{x}(t_j))]\right).$$

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

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

Approaches to Solve the Formal Problem

- **Long-Time-Window 4D-Var** (no practical implementation exists)
- **Sequential Estimation of the State**–(Extended) **Kalman Filter** (all practical DA systems)



Rudolf Kalman (May 19, 1930–July 2, 2016) Receives the National Medal of Science And Technology (October 7, 2009)

The Model Equations–Extended Kalman Filter

The **analysis** $\mathbf{x}_j^a = \mathbf{x}^a\left(t_j\right)$ is the mean of a multivariate normal distribution with covariance matrix $\mathbf{P}_j^a$:

$$\mathbf{x}_j^a = \mathbf{x}_j^b + \mathbf{K}_j \delta\mathbf{y}_j, \qquad \mathbf{P}_j^a = (\mathbf{I} - \mathbf{K}_j)\mathbf{P}_j^b$$

where

$$\begin{aligned}
\mathbf{x}_j^b &= \mathcal{M}_{j,j-1}\left(\mathbf{x}_{j-1}^a\right), \\
\delta\mathbf{y}_j &= \mathbf{y}_j^o - \mathcal{H}_j\left(\mathbf{x}_j^b\right), \\
\mathbf{P}_j^b &= \mathbf{M}_{j,j-1}\mathbf{P}_j^a\mathbf{M}_{j,j-1}^T, \\
\mathbf{K}_j &= \mathbf{P}_j^b\mathbf{H}_j^T\left(\mathbf{H}_j\mathbf{P}_j^b\mathbf{H}_j^T + \mathbf{R}_j\right)^{-1};
\end{aligned}$$

where $\mathbf{M}_{j,j-1}$ is the linearization of $\mathcal{M}_{j,j-1}$ about $\mathbf{x}_{j-1}^a$, and $\mathbf{H}_j$ is the linearization of $\mathcal{H}_j$ about $\mathbf{x}_j^b$

The Sources of Errors in the Statistical Model

A particular scheme must be robust to the following types of errors

- observations errors that do not satisfy the assumptions
- model errors (differences between the "true" atmospheric dynamics and $\mathcal{M}_{j,j-1}$)
- errors of the observation function $\mathcal{H}_j$
- errors of the linearization that produces $\mathbf{M}_{j,j-1}$ and $\mathbf{H}_j$
- nonlinear effects that limit the accuracy of $\mathbf{M}_{j,j-1}$ and $\mathbf{H}_j$

Robust parameters (statistics) are obtained by replacing the parameters (statistics) that would be optimal for clean input data by parameters (statistics) that lead only to slight degradations of the accuracy for clean data but make the model robust for contaminated data

Formal Definition of Robust Statistics

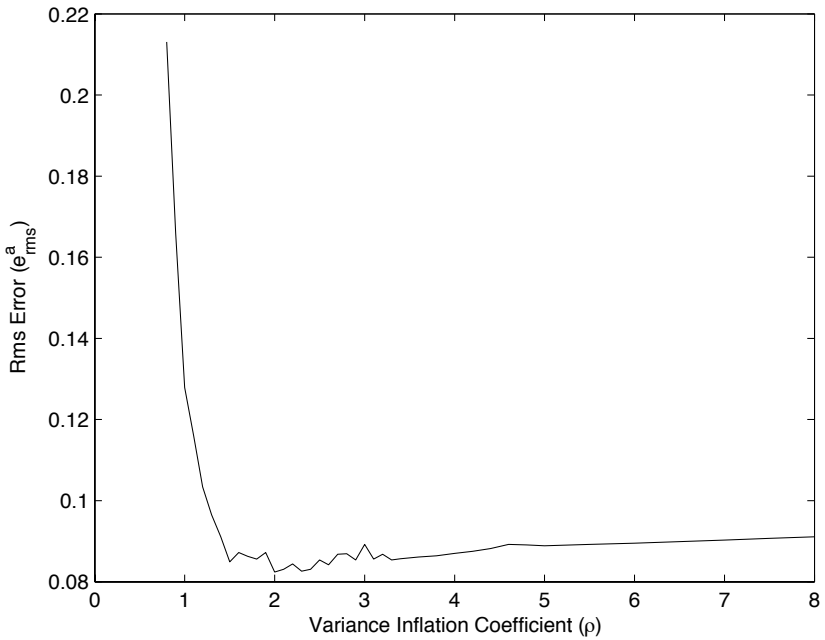
Robust statistics must satisfy the following criteria (Huber and Ronchetti 2009):

- **efficiency**–for **clean input data** (data that satisfy the assumptions of the original statistical model), the results are almost as good as for the original statistics (perfect model experiments)
- **stability**–**small errors** in the assumptions lead to small errors in the (state) estimates
- **breakdown**–**gross errors** in the input data do not lead to catastrophic breakdown

Example 1: Variance Inflation (from Szunyogh, 2014: Applicable Atmospheric Dynamics)

Assimilation of simulated **observations of the Henon Mapping** by an **Extended Kalman Filter**

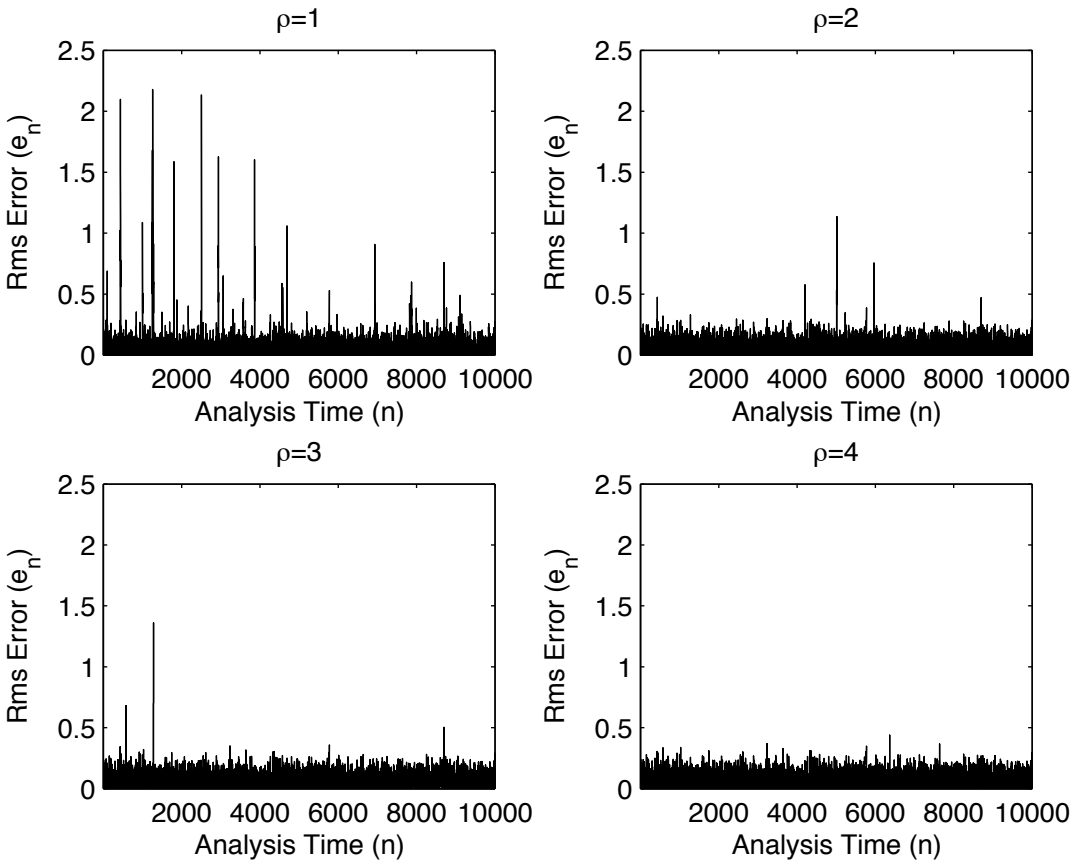
- The only sources of error in the statistical model are nonlinear effects that limit the accuracy of $\mathbf{M}_{j,j-1}$



Example 1 (Continued)

Variance inflation (replacing $\mathbf{P}_j^b$ by $\rho\mathbf{P}_j^b$, $\rho > 1$) reduces the magnitude and the frequency of **error bursts**

For more on the dynamics of error bursts see *Baek, Hunt, Szunyogh, Zimin, and Ott, 2004, Chaos, 14, 1042–1049.*



Example 2: Compensating for the Effects of Model Bias by Modifying **R**

From *Holt, Szunyogh, Gyarmati, Leidner, and Hoffman., 2015, MWR,* **143,** *3956–3980*

- The model has a single state variable and the background x^b is biased by b, and we have a direct observation y^o of x
- The analysis error has minimum variance, but not minimum rms
- The Kalman gain that minimizes the rms error is

$$\hat{K} = \left(P^b + b^2\right)\left(P^b + b^2 + R\right)^{-1}$$

rather than $K = \left(P^b\right)\left(P^b + R\right)^{-1}$

- The same effect can be achieved by using K and replacing R by

$$\hat{R} = R(1 + b^2/P^b)^{-1}$$

Example 2: Continued

Assume that

- the data assimilation system uses $(P^b)^{1/2} = 4\ hPa$ for the SLP in a TC
- the data assimilation system uses $(R)^{1/2} = 5\ hPa$ for a TCVitals SLP observation
- x^b is biased with $b = 40\ hPa$

Using $\hat{R}$ (0.45 hPa^2) rather than R (5 hPa^2)

- increases the standard deviation of the analysis error from 3.12 hPa to 4.92 hPa, but reduces the rms error of the analysis from 24.59 hPa to 4.96 hPa
- A huge reduction of the analysis bias at the price of a small increase of the analysis error variance
- Can be used, if there is no reason to believe that the analysis with a smaller bias would upset the model

Example 3: Coping with Gross Observation Errors and/or Good Observations that May Shock the Model

Roh, Genton, Jun, Szunyogh, and Hoteit, 2013: *Observation Quality Control with a Robust Ensemble Kalman Filter, MWR,* **141,** 4414–4428

- The analysis update equation can be Huberized as

$$\mathbf{x}^a = \mathbf{x}^b + \mathbf{K}\mathbf{G}\left(\delta\mathbf{y}\right),$$

where $\mathbf{G}\left(\delta\mathbf{y}\right)$ is the **Huber function,**

- For instance, a potential choice for the Huber function is

$$\mathbf{G}\left(\delta\mathbf{y}\right) = \begin{cases} \delta\mathbf{y} & \text{if} |\delta\mathbf{y}| < c \\ c & \text{if} \delta\mathbf{y} \geq c \\ -c & \text{if} \delta\mathbf{y} \geq -c \end{cases}$$

where c is a prescribed clipping innovation

Illustration of Examples 2 and 3 for TC Observations

- Based on Holt et al., 2015, MWR, **143,** 3956–3980
- **Models:** NCEP GFS at resolution T62L28, RSM at resolution 48 km and 28 levels (a glorified toy system)
- **Data assimilation:** LETKF
- **Regular observations:** all operationally assimilated non-radiance observations
- **TC observations:** TCVitals SLP ($R^{1/2} = 0.5$ hPa), drospsondes from DOTSTAR, QuikSCAT (both with Huberized innovation)

Illustration: Sinlaku Analyses

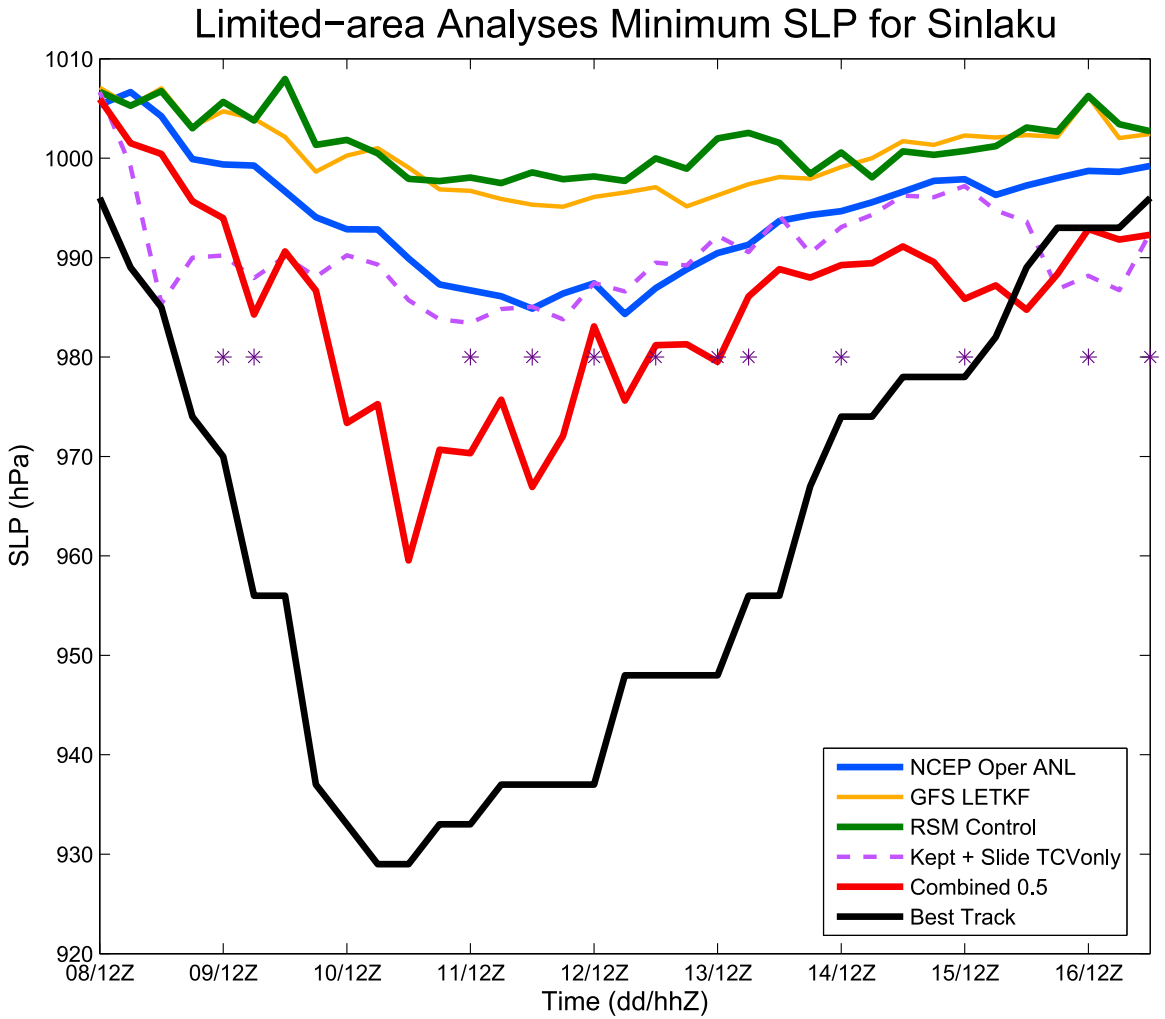
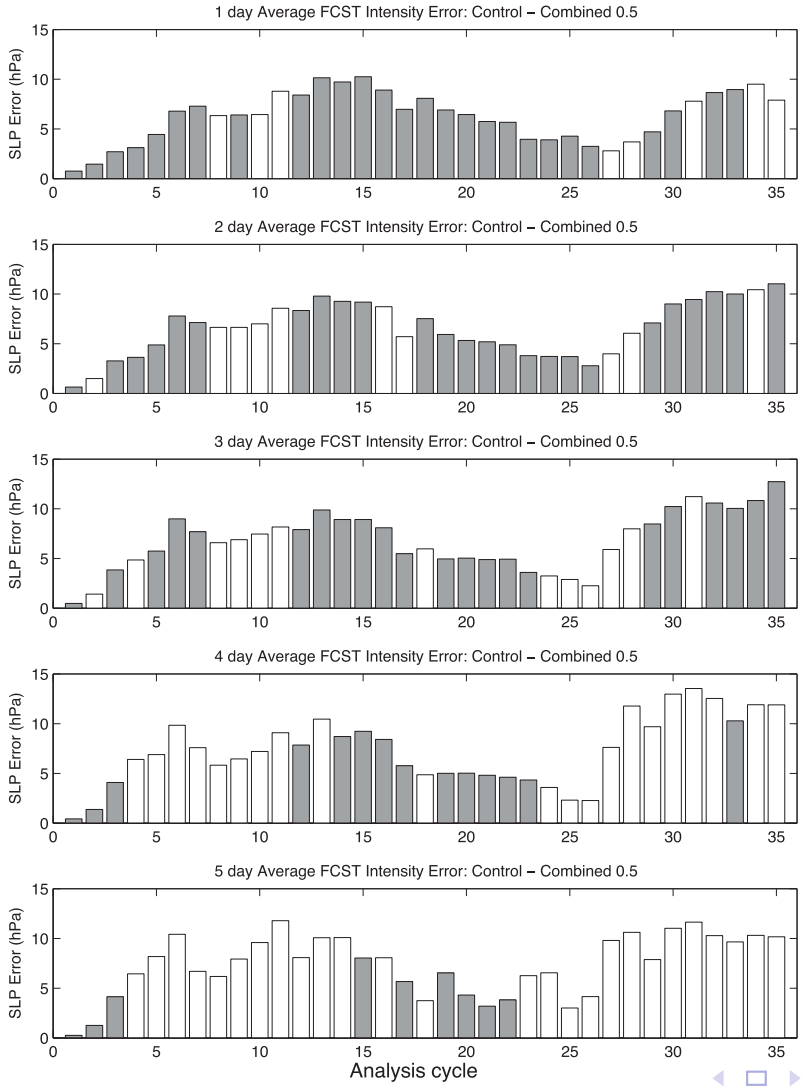


Illustration: Sinlaku Forecasts



Concluding Remarks

- People have always been working hard on making their data assimilation systems robust
- But, they do not like to talk about the adjustments they make to the error statistics, because they feel that these are hard to defend (reviewers make sure that they feel that way!)
- Keep in mind that the need for such adjustments is fully expected, as the mathematical model of data assimilation is not more than an extremely useful but imperfect model