The Abdus Salam
International Centre for Theoretical Physics

United Nations
Educational, Scientific and
Cultural Organization

IAEA
International Atomic Energy Agency

1966-3

Fall Colloquium on the Physics of Weather and Climate: Regional Weather Predictability and Modelling

29 September - 10 October, 2008

Atmospheric data assimilation

Dusanka Zupanski
Colorado State University
USA

Lecture 1: Atmospheric Data Assimilation

Dusanka Zupanski
CIRA/Colorado State University
Fort Collins, Colorado

Fall Colloquium on the Physics of Weather and Climate: Regional Weather Predictability and Modelling
29 September - 7 October, 2008, Trieste, Italy

OUTLINE

Lecture 1:

- What is data assimilation and why is it important?
- Ensemble Data Assimilation

Lecture 2:

- Maximum Likelihood Ensemble Filter (MLEF)
- Examples of data assimilation results

NOTE

If there is enough interest, we will perform simple data assimilation experiments, using Maximum Likelihood Ensemble Filter (MLEF) method and Burgers model, in one of the available Computer Labs.

Interested participants, please sign-in at the end of this class.

Suggested time for Lab exercises is today or Thursday after lunch break. Please write your suggestions next to your name when signing.

Websites

If you would like to perform the exercise on your own, downloaded the MLEF algorithm, via anonymous ftp, from the following location:

ftp://ftp.cira.colostate.edu/Zupanski/ICTP_Lecture_2008/mlef-V1.2.tar

Then, do the following (on a Linux computer or laptop):

```
tar -xvf mlef-V1.2.tar
cd mlef-V1.2
```

and read the text file called “README” for further instructions

More information about the MLEF can be found at

http://www.cira.colostate.edu/projects/ensemble

Questions

Please feel free to ask questions during and after my presentations (until Saturday, when I will be leaving).

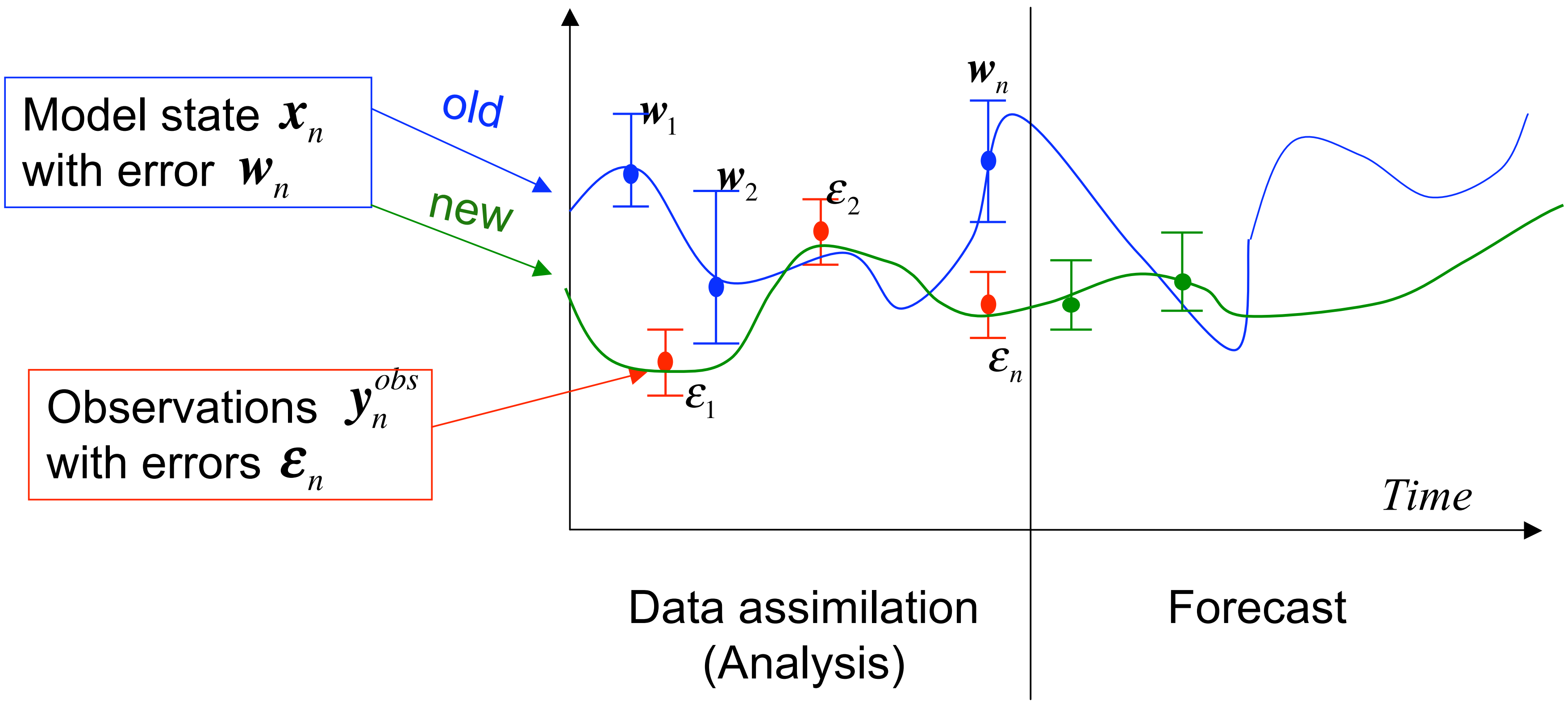
You are also welcome to send me an e-mail.

Dusanka Zupanski, CIRA/CSU
Zupanski@CIRA.colostate.edu

What is Data Assimilation?

- Data assimilation (DA) started as Atmospheric DA in 1960s-1970s (you might have heard of Gandin, who was a pioneer in the atmospheric DA).
- Initially, atmospheric DA was used to define “objective” **initial conditions** for Numerical Weather Predictions (NWP) models. In these early days, the term “**objective analysis**” was used.
- Today, DA is used to define “**optimal**” initial conditions for NWP, climate, hydrological, oceanic, land, canopy, pollution transport and other bio-geophysical models (it is not only atmospheric DA anymore).
- Advanced DA methods, available today, can also estimate and correct model errors (**model error estimation**) and define optimal empirical parameters (**parameter estimation**).

What is Data Assimilation?



Model state x_n with error w_n

Observations y_n^{obs} with errors ε_n

New (posterior) model state fits the observations better than the old (prior) model state: Data Assimilation (DA) “**optimally**” combines **observations** and a **dynamical forecast model** in order to improve estimates of the current state (**analysis**) and the future state (**forecast**) of a dynamical system of interest (e.g., atmosphere, ocean, land).

How is this done?

There are several steps:

Step 1: Define forecast model time evolution equation:

$$\boxed{\boldsymbol{x}_n = M_{n,n-1}(\boldsymbol{x}_{n-1}) + \boldsymbol{w}_n}$$

n - Time step index (denoting model time steps)

M - Dynamical model for model state evolution (e.g., Eta model)

$\boldsymbol{x}$ - Model state vector of dim $Nstate$

$\boldsymbol{w}$ - Model error vector of dim $Nstate$

Step 2: Define time evolution equation for the observations:

$$\boxed{y_k = H_k(\boldsymbol{x}_k) + \boldsymbol{\varepsilon}_k}$$

k - Time step index (denoting observation times)

$\boldsymbol{y}$ - Observations vector of dim $Nobs$

$\boldsymbol{\varepsilon}$ - Observation error

H - Observation operator

Dusanka Zupanski, CIRA/CSU
Zupanski@CIRA.colostate.edu

Step 3: Combine the two equations in the probability space, by employing the Bayesian conditional Probability Density Function (PDF):

$$PDF = p\left(\boldsymbol{x}\middle|\,\boldsymbol{y}\right) = \frac{p(\boldsymbol{y}|\boldsymbol{x})p(\boldsymbol{x})}{p(\boldsymbol{y})}$$

In order to proceed with the next step, we assume that $p(\boldsymbol{x})$ and $p(\boldsymbol{y}|\boldsymbol{x})$ are known. Typical assumption is that both PDFs are Gaussian and are mutually independent, thus we can write

$$p(\boldsymbol{x}) \sim exp\left[-\frac{1}{2}\left(\boldsymbol{x}-\boldsymbol{x}_b\right)^T \boldsymbol{P}_f^{-1}\left(\boldsymbol{x}-\boldsymbol{x}_b\right)\right]$$

$\boldsymbol{P}_f$ is forecast error covariance matrix (prescribed or calculated)

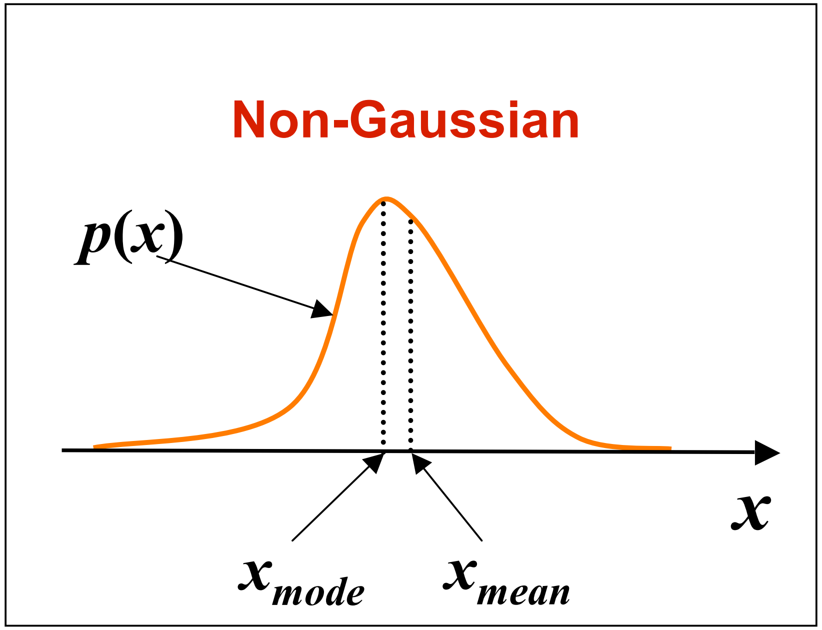
$$p(\boldsymbol{y}|\boldsymbol{x}) \sim exp\left\{-\frac{1}{2}\left[\boldsymbol{y}-H\left[M\left(\boldsymbol{x}\right)\right]\right]^T \boldsymbol{R}^{-1}\left[\boldsymbol{y}-H\left[M\left(\boldsymbol{x}\right)\right]\right]\right\}$$

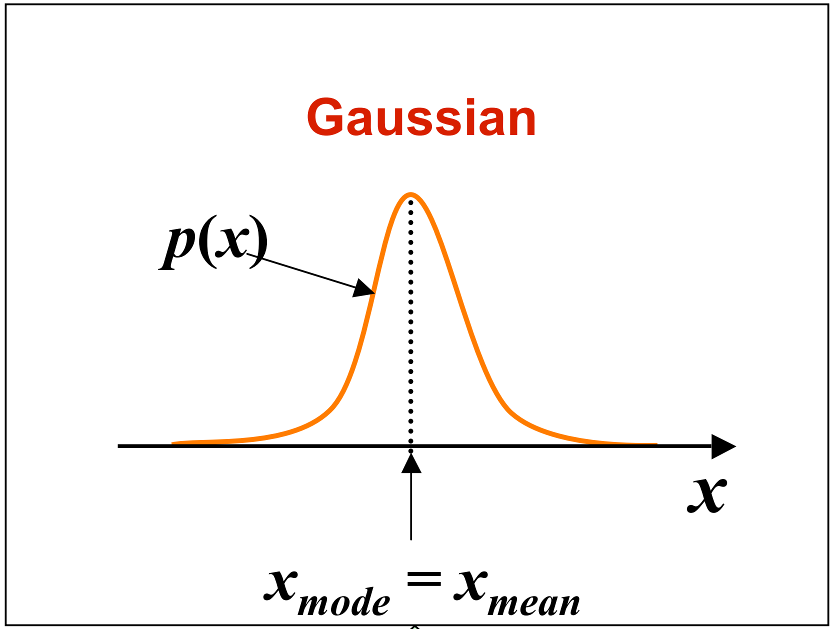
$\boldsymbol{R}$ is observation error covariance matrix (prescribed)

Step 4: Finally, obtain the minimum variance (mean) or maximum likelihood (mode) solution by minimizing the following cost function J

$$J = \frac{1}{2}[\boldsymbol{x}-\boldsymbol{x}_b]^T \boldsymbol{P}_f^{-1}[\boldsymbol{x}-\boldsymbol{x}_b] + \frac{1}{2}[H[M(\boldsymbol{x})]-\boldsymbol{y}]^T \boldsymbol{R}^{-1}[H[M(\boldsymbol{x})]-\boldsymbol{y}]$$

MEAN vs. MODE





Minimum variance estimate= Maximum likelihood estimate!

For Gaussian PDFs mean=mode, thus the solution is easy to obtain. Or is it?

Dusanka Zupanski, CIRA/CSU
Zupanski@CIRA.colostate.edu

Even under the Gaussian error assumption, there are different ways to obtain the optimal solution (analysis):

1. Variational methods
   - 3d-var
   - 4d-var

2. Kalman Filter - like methods
   - Classical Kalman Filter (KF)
   - Extended Kalman Filter (EKF)
   - Ensemble Kalman Filter (EnKF)

The most significant difference between the two groups of DA methods is in the way how $\boldsymbol{P}_f$ is defined:

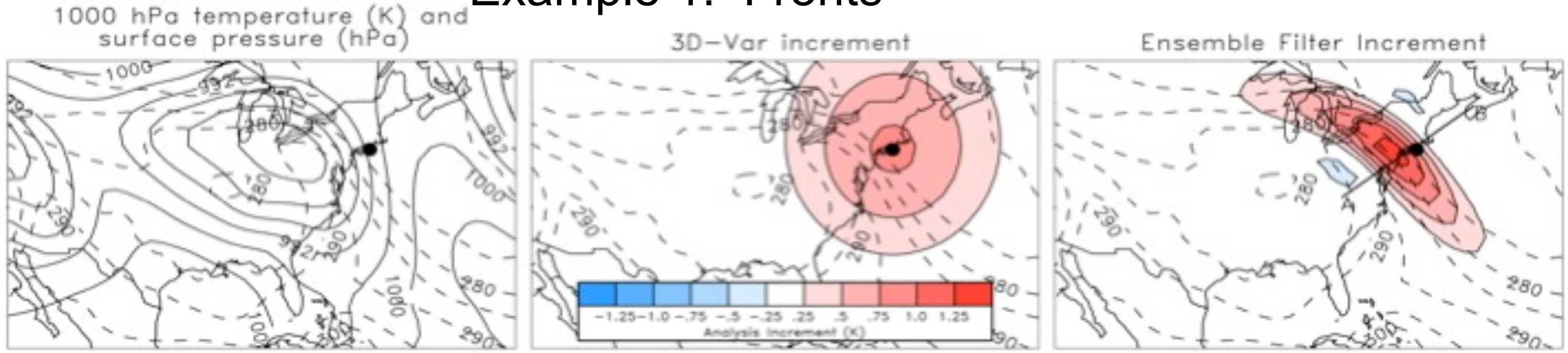
1. In variational methods $\boldsymbol{P}_f$ is prescribed.

2. In KF-like methods, $\boldsymbol{P}_f$ is evolving in time according to model dynamics.

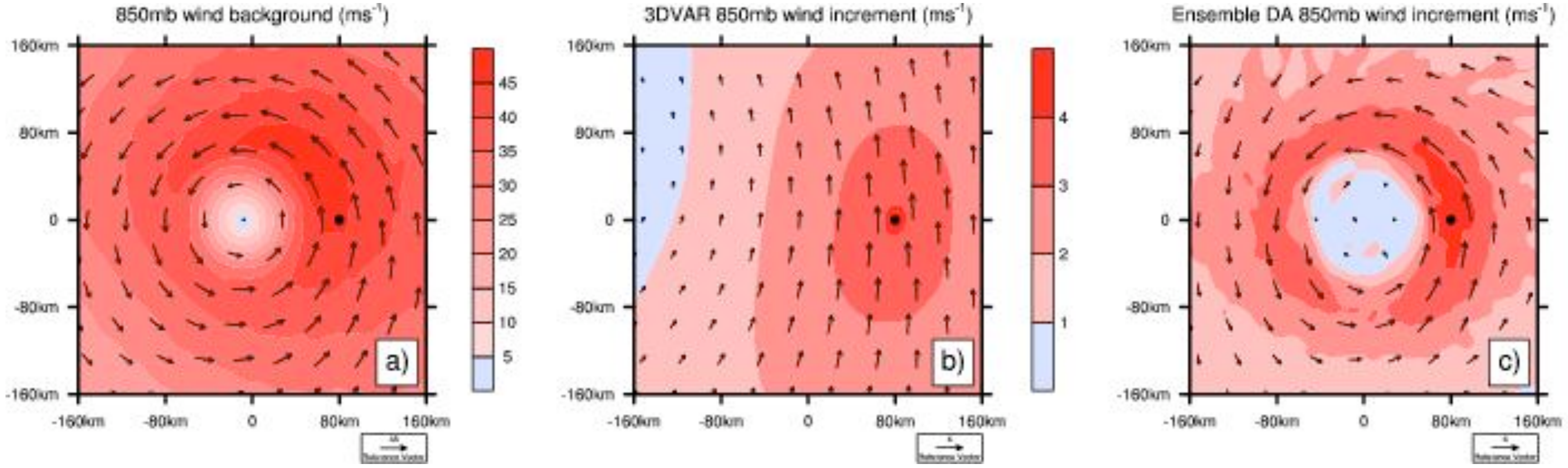
Why is it important to have time evolving P_f?

(From Whitaker et al., THORPEX web-page)

Example 1: Fronts



Example 2: Hurricanes



Ensemble methods produce realistic P_f. Nevertheless, covariance localization is sometimes needed to eliminate spurious long-distance correlations (covariance localization was applied in this example).

More specifically

❑ **3d-var** method employs a prescribed forecast error covariance and never evolves it in time.

❑ **4d-var** method evolves the forecast error covariance in time, but only until the end of a data assimilation interval. In each new data assimilation interval it starts with the same, prescribed forecast error covariance.

❑ **Kalman Filter** (KF) Does evolve the forecast error covariance in time and from one data assimilation cycle to another, but it is computationally too expensive for applications to complex atmospheric models.

⇓

Ensemble data assimilation (e.g., EnKF) is a practical alternative to KF, applicable to most complex atmospheric models (Eta, WRF).

⇓

A bonus benefit: The ensemble data assimilation is easier to deal with, since we do not need to define $\boldsymbol{P}_f$; the ensembles will do it for us.

More reasons for ensemble data assimilation

❑ Ensemble data assimilation provides initial ensemble perturbations for ensemble forecasting.

❑There is no need to develop adjoint models.

❑ As will be shown in Lecture 2, the ensemble data assimilation methods can deal with non-linear and discontinuous (on/off) functions better than variational methods

There are many different versions of EnKF

- Monte Carlo EnKF (Evensen 1994; 2003)
- EnKF (Houtekamer et al. 1995; 2005; First operational version)
- Hybrid EnKF (Hamill and Snyder 2000)
- EAKF (Anderson 2001)
- ETKF (Bishop et al. 2001)
- EnSRF (Whitaker and Hamill 2002)
- LEKF (Ott et al. 2004)
- LETKF (Hunt et al. 2007)
- SEEK filter (Pham et al. 1998)

Minimum variance solution

- MLEF (Zupanski 2005; Zupanski and Zupanski 2006)

Maximum likelihood solution

Why maximum likelihood solution? Because it is more adequate for non-linear and discontinuous models (Zupanski et al. 2008) and for problems involving non-Gaussian PDFs (e.g., Fletcher and Zupanski 2006).

Kalman filter solution

Analysis step:

$$\boldsymbol{x}_a = \boldsymbol{x}_b + \boldsymbol{P}_f \boldsymbol{H}^T (\boldsymbol{H}\boldsymbol{P}_f \boldsymbol{H}^T + \boldsymbol{R})^{-1} \left[\boldsymbol{y} - H(\boldsymbol{x}_b) \right]$$

$\boldsymbol{x}_a$ - Optimal estimate of $\boldsymbol{x}$ (analysis)

$\boldsymbol{x}_b$ - Background (prior) estimate of $\boldsymbol{x}$

$$\boldsymbol{P}_a = [\boldsymbol{I} - \boldsymbol{P}_f \boldsymbol{H}^T (\boldsymbol{H}\boldsymbol{P}_f \boldsymbol{H}^T + \boldsymbol{R})^{-1} \boldsymbol{H}] \boldsymbol{P}_f = \left(\boldsymbol{I} - \boldsymbol{K}\boldsymbol{H} \right) \boldsymbol{P}_f$$

$\boldsymbol{P}_a$ - Analysis (posterior) error covariance matrix (*Nstate* x *Nstate*)

$\boldsymbol{K}$ - Kalman gain matrix (*Nstate* x *Nobs*)

Forecast step:

$$\boldsymbol{x}_0 = \boldsymbol{x}_a \quad ; \quad \boldsymbol{x}_n = M_{n,n-1}(\boldsymbol{x}_{n-1}) + \boldsymbol{w}_n$$

Often neglected

$$\boldsymbol{P}_f = \boldsymbol{M}\boldsymbol{P}_a \boldsymbol{M}^T + \boldsymbol{Q}$$

- Update of forecast error covariance

Ensemble Kalman Filter (EnKF) solution

EnKF was first introduced by Evensen (1994) as a Monte Carlo filter.

Equations following Evensen (2003)

Analysis step:

Analysis solution defined for each ensemble member i:

$$\boldsymbol{x}_a^i = \boldsymbol{x}_b^i + \boldsymbol{P}_f^e \boldsymbol{H}^T (\boldsymbol{H}\boldsymbol{P}_f^e \boldsymbol{H}^T + \boldsymbol{R}^e)^{-1}(\boldsymbol{y}^i - H(\boldsymbol{x}_b^i))$$

Mean analysis solution:

$$\overline{\boldsymbol{x}_a} = \overline{\boldsymbol{x}_b} + \boldsymbol{P}_f^e \boldsymbol{H}^T (\boldsymbol{H}\boldsymbol{P}_f^e \boldsymbol{H}^T + \boldsymbol{R}^e)^{-1}(\overline{\boldsymbol{y}} - H(\overline{\boldsymbol{x}_b}))$$

Analysis ensemble perturbations:

$$p_a^i = \boldsymbol{x}_a^i - \overline{\boldsymbol{x}_a}$$

Analysis error covariance in ensemble subspace:

$$\left(\boldsymbol{P}_a\right)^{1/2} = \left[p_a^1 \quad p_a^2 \quad . \quad p_a^{Nens} \right]$$

$$\boldsymbol{P}_a = \frac{1}{N_{ens} - 1}\left(\boldsymbol{P}_a\right)^{1/2}\left[\left(\boldsymbol{P}_a\right)^{1/2}\right]^{\boldsymbol{T}}$$

Dusanka Zupanski, CIRA/CSU
Zupanski@CIRA.colostate.edu

Ensemble Kalman Filter (EnKF) solution

Forecast step:

Ensemble forecasts employing a **non-linear** model M

$$\boxed{\boldsymbol{x}_n^j = M_{n,n-1}(\boldsymbol{x}_{n-1}^j)}$$

Non-linear forecast perturbations

$$\boxed{p_f^i = M(\boldsymbol{x}_a^i) - M(\overline{\boldsymbol{x}_a})}$$

Forecast error covariance calculated using ensemble perturbations:

$$\boxed{\left(\boldsymbol{P}_f\right)^{1/2} = \left[p_f^1 \quad p_f^2 \quad . \quad p_f^{Nens}\right]}$$

$$\boxed{\boldsymbol{P}_f = \frac{1}{N_{ens} - 1}\left(\boldsymbol{P}_f^e\right)^{1/2}\left[\left(\boldsymbol{P}_f^e\right)^{1/2}\right]^{\boldsymbol{T}}}$$

Dusanka Zupanski, CIRA/CSU
Zupanski@CIRA.colostate.edu

MLEF solution

Analysis step:

Analysis solution $\boldsymbol{x}_a$ obtained by minimizing the cost function

$$J = \frac{1}{2}[\boldsymbol{x}-\boldsymbol{x}_b]^T \boldsymbol{P}_f^{-1}[\boldsymbol{x}-\boldsymbol{x}_b] + \frac{1}{2}[H(\boldsymbol{x})-\boldsymbol{y}]^T \boldsymbol{R}^{-1}[H(\boldsymbol{x})-\boldsymbol{y}]$$

Analysis error covariance in ensemble subspace:

$$(\boldsymbol{P}_a)^{1/2} = \boldsymbol{P}_f^{1/2}\left[\boldsymbol{I}+(\boldsymbol{Z}(\boldsymbol{x}_a))^T \boldsymbol{Z}(\boldsymbol{x}_a)\right]^{-1/2}$$

$$\boldsymbol{Z}(\boldsymbol{x}) = \left[\boldsymbol{z}_1(\boldsymbol{x}) \quad \boldsymbol{z}_2(\boldsymbol{x}) \quad \cdot \quad \cdot \quad \boldsymbol{z}_{Nens}(\boldsymbol{x})\right] \; ; \; \boldsymbol{z}_i(\boldsymbol{x}) = \boldsymbol{R}^{-1/2}\left[H(\boldsymbol{x}+\boldsymbol{p}_i^f)\text{-}H(\boldsymbol{x})\right]$$

$$\boldsymbol{P}_a = (\boldsymbol{P}_a)^{1/2}\left[(\boldsymbol{P}_a)^{1/2}\right]^T$$

Dusanka Zupanski, CIRA/CSU
Zupanski@CIRA.colostate.edu

MLEF solution

Forecast step:

Ensemble forecasts employing a **non-linear** model M

$$\boldsymbol{x}_n^j = M_{n,n-1}(\boldsymbol{x}_{n-1}^j)$$

Non-linear ensemble forecast perturbations

$$p_f^i = M(\boldsymbol{x}_a + p_a^i) - M(\boldsymbol{x}_a)$$

Forecast error covariance calculated using ensemble perturbations:

$$\left(\boldsymbol{P}_f\right)^{1/2} = \left[p_f^1 \quad p_f^2 \quad . \quad p_f^{Nens}\right]$$

$$\boldsymbol{P_f} = \left(\boldsymbol{P}_f\right)^{1/2}\left[\left(\boldsymbol{P}_f\right)^{1/2}\right]^{\boldsymbol{T}}$$

Dusanka Zupanski, CIRA/CSU
Zupanski@CIRA.colostate.edu

Connections between MLEF, KF and 3d-var

(Zupanski et al. 2007)

- **MLEF=KF, if PDFs are Gaussian, observation operator H is linear and differentiable and the MLEF uses a full-rank P_f (Full-rank means *Nens*=*Nstate*).**

$$\boxed{\boldsymbol{x} = \boldsymbol{x}_b + \alpha \boldsymbol{P}_f \boldsymbol{H}^T (\boldsymbol{H}\boldsymbol{P}_f^T \boldsymbol{H}^T + R)^{-1} [\boldsymbol{y} - H(\boldsymbol{x}_b)]} \quad \text{; for } \alpha=1$$

***MLEF is also applicable to Non-Gaussian PDFs** (e.g., Fletcher and Zupanski 2006).*

- **MLEF=3d-var, if observation operator is differentiable and P_f is full-rank, but not updated from one data assimilation cycle to another. In such case the same cost function is minimized:**

$$\boxed{J(\boldsymbol{x}) = \frac{1}{2}[\boldsymbol{x} - \boldsymbol{x}_b]^T \boldsymbol{P}_f^{-1} [\boldsymbol{x} - \boldsymbol{x}_b] + \frac{1}{2}[\boldsymbol{y} - H(\boldsymbol{x})]^T \boldsymbol{R}^{-1} [\boldsymbol{y} - H(\boldsymbol{x})]}$$

***MLEF is also applicable to non-differentiable observation operators** (e.g., Zupanski et al. 2008).*

Dusanka Zupanski, CIRA/CSU
Zupanski@CIRA.colostate.edu

Current status of EnKF applications

- ➢EnKF is operational in Canada, since January 2005 (Houtekamer et al. 2005).
- ➢ Pseudo operational EnKF at University of Washington (Torn and Hakim 2008).
- ➢ EnKF is better than 3d-var (Meng and Zhang 2008; Whitaker et al., 2008).
- ➢ Superior performance in application to non-hydrostatic, cloud resolving models (Caya et al. 2005; Xue et al. 2006; Carrio et al. 2008).
- ➢ Superior performance in applications to non-linear and discontinuous (with on/off switches) models (Zupanski et al. 2008).
- ➢ Superior performance for ocean (Keppenne et al. 2008), climate (Karspeck and Anderson 2007), and soil hydrology models (Reichle et al. 2007).

Theoretical advantages of ensemble-based DA methods are getting confirmed in an increasing number of practical applications.

There is still many possibilities for further improvements of the ensemble based DA methods.. For example, covariance localization is one of the critical areas that need further improvements.

A hint on how to run the MLEF algorithm

Dusanka Zupanski, CIRA/CSU
Zupanski@CIRA.colostate.edu