# A Comparison Study of Data Assimilation Algorithms for Ozone Forecasts

Lin Wu, Vivien Mallet,
Marc Bocquet, Bruno Sporstisse

CEREA, INRIA clime

Atelier Polyphemus, Oct 27, 2008

# Problem and Objective

## Data Assimilation Problem

Estimate the uncertainties for a better prediction from diverse resources, i.e. model simulations, observations and statistics information.

## Background

- Key issue in environmental problems, e.g. meteo., ocean, soil...
- Many experiences in meteorological data assimilation.

## Air Quality, Short-Range Ozone Forecasts ?

- Evaluate the performances of different data assimilation schemes.
- Help to design suitable assimilation algorithms for in realistic applications

# Air Quality Model

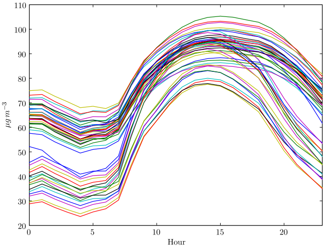
## Chemistry-transport equation for air quality model

$$\frac{\partial c_i}{\partial t} = \underbrace{-\mathrm{div}(Vc_i)}_{\text{advection}} + \underbrace{\mathrm{div}\left(\rho K \nabla \frac{c_i}{\rho}\right)}_{\text{diffusion}} + \underbrace{\chi_i(c)}_{\text{chemistry}} + \underbrace{S_i - L_i}_{\text{sources and losses}}$$

## Facts

- Nonlinear due to chemical reaction term $\chi_i(c)$
- High dimension (typically $10^6 \sim 10^7$)
- Strong uncertainties mainly due to uncertain parameters; initial conditions tend to be forgotten.

# Uncertainties




48 ensemble samples, Vivien & Bruno, JGR, 2006

## Probability density function (PDF) of model state

- PDF evolution
- Not possible (high dimension $10^6 \sim 10^7$)
- Ensemble approximations

## Uncertainties of model parameters

- Biogenic emission $\pm 100\%$
- Anthropogenic emissions $\pm 50\%$
- Boundary condition $\pm 20\%$
- Cloud attenuation $\pm 30\%$
- Deposition velocity ($O_3$, $NO_2$) $\pm 30\%$
- ...

# Assimilation Algorithms

- Model and observations at time step $k$ :

$$\left\{ \begin{array}{lll} \mathbf{x}_k = M_{k-1}[\mathbf{x}_{k-1}] + \epsilon^f_{k-1} & \text{Model} & M_{k-1} \\ \mathbf{y}_k = H_k[\mathbf{x}_k] + \epsilon^o_k & \text{Observation} & \mathbf{y}_k \end{array} \right.$$

- Minimization of a cost function $J(\mathbf{x})$ that deals with obs. :

$$\frac{1}{2}(\mathbf{x} - \mathbf{x}_k)^T \mathbf{P}_k^{-1} (\mathbf{x} - \mathbf{x}_k) + \frac{1}{2} (\mathbf{y}_k - H_k[\mathbf{x}])^T \mathbf{R}_k^{-1} (\mathbf{y}_k - H_k[\mathbf{x}])$$

# Probabilitic Formulation of Data assimilation Problem

- Model & Obs. : $\mathbf{Y}_k \equiv \{\mathbf{y}_i^o, i = 1, ..k\}$

$$\begin{cases} \mathbf{x}_k^t = M_{k-1}[\mathbf{x}_{k-1}^t] + \epsilon_{k-1}^f \\ \mathbf{y}_k^o = H_k[\mathbf{x}_k^t] + \epsilon_k^o \end{cases}$$

- Forecast (governed by dynamics) : Chapman-Kolmogorov equation, or Fokker-Planck equation (SDE)

$$p(\mathbf{x}_k^t|\mathbf{Y}_{k-1}) = \int p(\mathbf{x}_k^t|\mathbf{x}_{k-1}^t)p(\mathbf{x}_{k-1}^t|\mathbf{Y}_{k-1})d\mathbf{x}_{k-1}^t$$

- Analysis (conditioned by observations) :
  - Discrete observations : Bayes rule

$$p(\mathbf{x}_k^t|\mathbf{Y}_k) = \frac{p(\mathbf{y}_k^o|\mathbf{x}_k^t)p(\mathbf{x}_k^t|\mathbf{Y}_{k-1})}{p(\mathbf{y}_k^o|\mathbf{Y}_{k-1})}$$

  - Continuous observations : Zakai or Kushner equations

- Estimation criteria : maximum likelyhood, minimum variance

...

# Assimilation Algorithms

Neither the PDE nor the integral in Bayes formula is tractable for high dimensional geophysical systems ; all assimilation algorithms are approximations.

## Variational Algorithms

A block of observations. Optimal control theory applies.

- Four dimensional variational assimilation (4DVar) : time interval $k = 0, \ldots, N$.

## Sequential Algorithms

Spontaneous observations. Filtering theory applies.

- Optimal interpolation (OI) ;
- Ensemble Kalman filter (EnKF) ;
- Reduced-rank square root Kalman filter (RRSQRT) ;

# Assimilation Algorithms - 4DVar

- 4DVar (Le Dimet 1982)
  Maximum likelyhood estimation with assumptions of Markovian dynamics and Gaussian errors in the model and observations ; minimization of a cost function $J(\mathbf{x})$ that deals with a set of obs. :

$$\underbrace{\frac{1}{2}(\mathbf{x}-\mathbf{x}^b)^T\mathbf{B}^{-1}(\mathbf{x}-\mathbf{x}^b)}_{J_b}+\underbrace{\frac{1}{2}\sum_{k=0}^{N}\overbrace{(\mathbf{y}_k-H_k[\mathbf{x}_k])^T\,\mathbf{R}_k^{-1}\,(\mathbf{y}_k-H_k[\mathbf{x}_k])}^{J_{oi}}}_{J_o}$$

  - The assimilation window : $0-N$
  - $\mathbf{x}_k = M_{0\rightarrow k}[\mathbf{x}] = M_k M_{k-1}\ldots M_1[\mathbf{x}]$

# Assimilation Algorithms - 4DVar

- Efficient calculation of gradients by adjoint model
  - $\tilde{\mathbf{x}}_N = 0$ ,
  - For $k = N, \ldots, 1$, calculates $\tilde{\mathbf{x}}_{k-1} = \mathbf{M}_{k-1}^T \left(\tilde{\mathbf{x}}_k - \mathbf{H}_k^T d_k\right)$, where $d_k = \mathbf{R}_k^{-1} \left(\mathbf{y}_k - H_k(\mathbf{x}_k)\right)$ ,
  - $\tilde{\mathbf{x}}_0 := \tilde{\mathbf{x}}_0 - \mathbf{H}_0^T (d_0)$ gives the gradient of $J_o$ with respect to $\mathbf{x}$.
- Balgovind isotropic correlation function for $\mathbf{B}$. The error covariance between two points is given by :

$$f(d) = \left(1 + \frac{d}{L}\right) e^{-\frac{d}{L}} v$$

  - $L$ : the characteristic length
  - $d$ : the distance between the two points
  - $v$ : is the a priori variance

# Assimilation Algorithms - EnKF

- EnKF (Evensen 1994)
  Monte Carlo solution of Fokker-Planck equations, at analysis step assumptions of Gaussian errors and linear dynamics.
  - Initialization : given initial pdf $p(\mathbf{x}_0^t)$, an ensemble of $r$ members $\{\mathbf{x}_0^{a,i}, \quad , i = 1, \ldots, r\}$. Let the bar denote the mean over ensemble members, e.g. $\bar{\mathbf{x}}_0^a = \frac{1}{r} \sum_{i=1}^{r} \mathbf{x}_0^{a,i}$
  $$\tilde{\mathbf{P}}_0^a = \frac{1}{r-1} \sum_{i=1}^{r} \left( \mathbf{x}_0^{a,i} - \bar{\mathbf{x}}_0^a \right) \left( \mathbf{x}_0^{a,i} - \bar{\mathbf{x}}_0^a \right)^T$$
  - Forecast formula :
  $$\mathbf{x}_k^{f,i} = M_{k-1}[\mathbf{x}_{k-1}^{a,i}] + \epsilon_{k-1}^{f,i}$$
  $$\tilde{\mathbf{P}}_k^f = \frac{1}{r-1} \sum_{i=1}^{r} \left( \mathbf{x}_k^{f,i} - \bar{\mathbf{x}}_k^f \right) \left( \mathbf{x}_k^{f,i} - \bar{\mathbf{x}}_k^f \right)^T$$

# Assimilation Algorithms - EnKF

- Analysis Formula :

$$\mathbf{x}_k^{a,i} = \mathbf{x}_k^{f,i} + \tilde{\mathbf{K}}_k \left(\mathbf{y}_k^i - H_k[\mathbf{x}_k^{f,i}]\right)$$
$$\tilde{\mathbf{K}}_k = \tilde{\mathbf{P}}_k^f \mathbf{H}_k^T (\mathbf{H}_k \tilde{\mathbf{P}}_k^f \mathbf{H}_k^T + \mathbf{R}_k)^{-1}$$
$$\tilde{\mathbf{P}}_k^a = \frac{1}{r-1} \sum_{i=1}^{r} \left(\mathbf{x}_k^{a,i} - \bar{\mathbf{x}}_k^a\right) \left(\mathbf{x}_k^{a,i} - \bar{\mathbf{x}}_k^a\right)^T$$

- Model Error : approximated by perturbing model input data and model parameters

$$\epsilon_{k-1}^{f,(i)} \simeq M_{k-1}\left(\mathbf{x}_{k-1}^{a,(i)}, \mathbf{w}^{(i)}\mathbf{d}\right) - M_{k-1}\left(\mathbf{x}_{k-1}^{a,(i)}, \mathbf{d}\right)$$

  - $\mathbf{d}$ : the vector of parameters to be perturbed
  - $\mathbf{w}^{(i)}$ : for $i-$th sample, the diagonal matrix whose elements are perturbation coefficients. For instance, for one lognormal parameter $p$ in $\mathbf{d}$, the perturbation is as :

$$\hat{p} = p \times \sqrt{\alpha}^{\gamma}$$

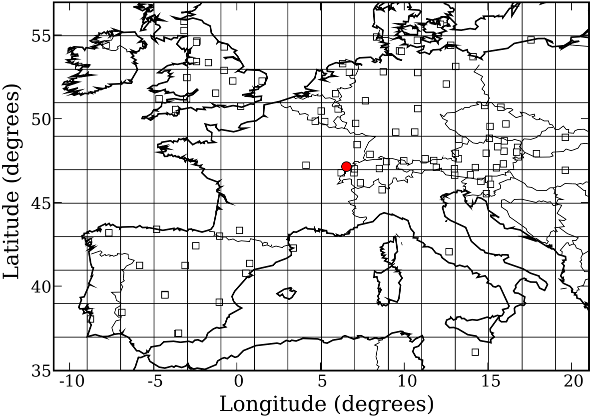
    $\alpha$ : perturbation magnitude ; $\gamma$ : quantity sampled from standard normal distribution (constant for one type of parameter).

# List of Perturbed Parameters

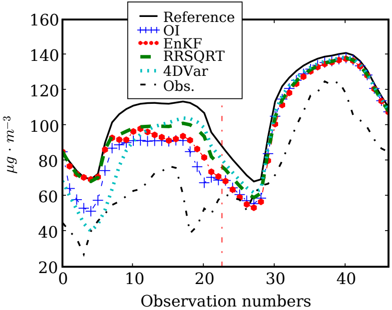
| Parameter name | distribution | $\alpha$ |
|---|---|---|
| Boundary condition | log-normal | 3. |
| Deposition velocity | log-normal | 1.5 |
| Photolysis rate | log-normal | 1.3 |
| Surface emission | log-normal | 1.5 |
| Attenuation | log-normal | 1.3 |
| Vertical diff. coef. | log-normal | 1.3 |
| Cloud height | log-normal | 1.3 |
| Vertical wind | log-normal | 1.3 |
| Specific humidity | log-normal | 1.3 |
| Rain | log-normal | 1.3 |
| Pressure | log-normal | 1.3 |
| Air density | log-normal | 1.3 |
| Merid. diff. coef. | log-normal | 1.3 |
| Zonal diff. coef. | log-normal | 1.3 |
| Temperature* | normal | 0.005 |

Tab.: The set of perturbed parameters.

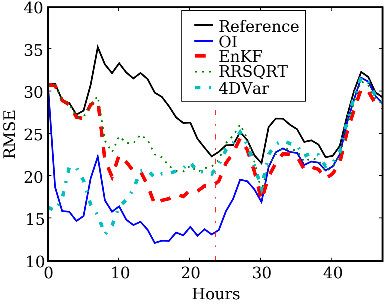
# Comparisons Results : EMEP Network



# Comparisons Results : Four Methods




Ozone concentrations for Montandon station.

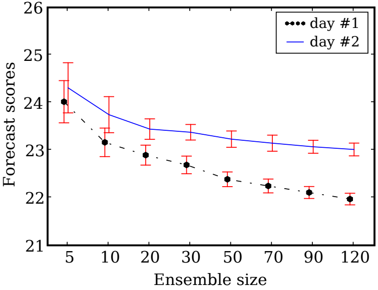



Time evolution of RMSE.

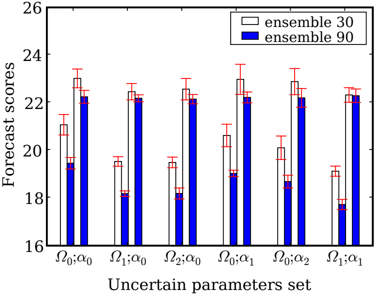
## Notes

OI : overall better performance (explicit parameterization of model error) ; EnKF : better prediction ; RRSQRT : poor overall performance (SVD truncations ?) ; 4DVar : better assimilation worst prediction.

# Sensitivity to Algorithm Parameters

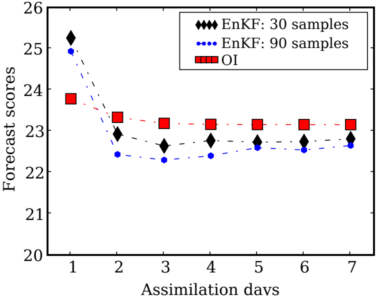



Forecast scores of EnKF against the ensemble size.

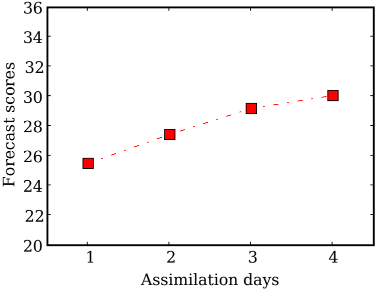



Forecast scores for EnKF against different uncertain parameter definitions.

# Sensitivity to Algorithm Parameters

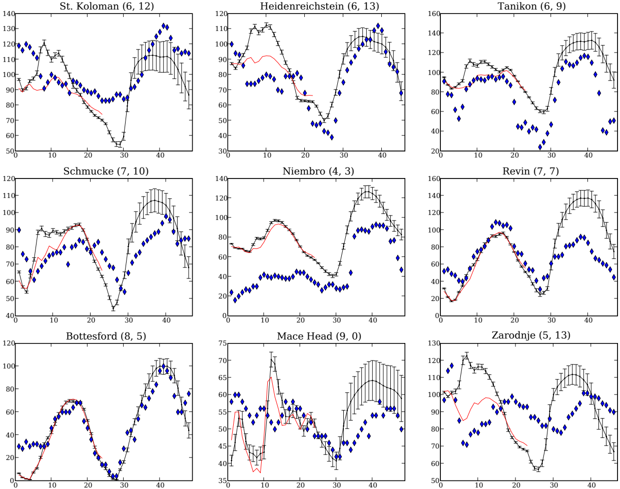



Forecast scores of OI and EnKF (with 30 and 90 members) against the number of assimilation days.

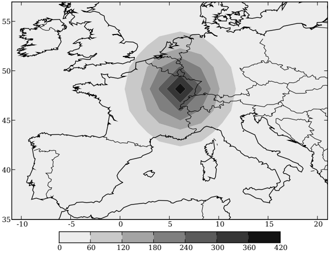



Forecast scores against the number of assimilation days for the two experiments using 4DVar.

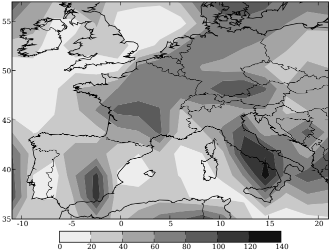
# EnKF performances at nine stations



# The Error Covariance Structure



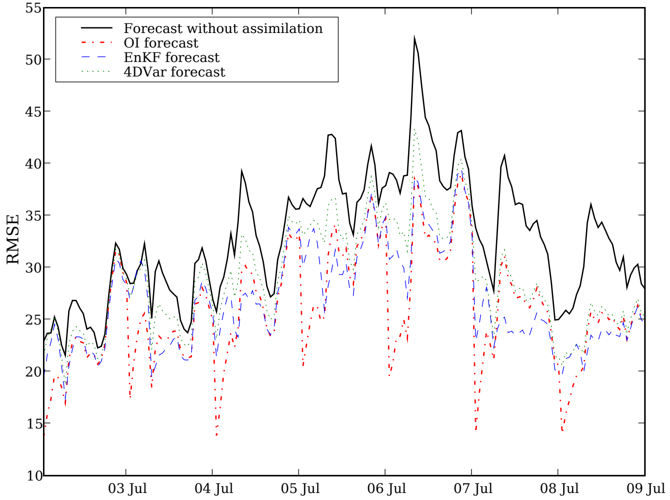

Balgovind parameterization.




Approximations by EnKF forecast ensemble.

The covariance between the error at the station Montandon and the error in all ground cells at 13 :00 UT, July 2, 2001.

# Cycling Forecast



# Conclusions and Perspectives

## Conclusions

- Assimilations improve the forecast significantly.
- Pros and cons for a set of assimilation methods ; perturbation methods for assimilation.

## Perspectives

- Methods that allows control of uncertain parameters (e.g. $K_z$), not just state.
- Better perturbations, e.g. perturbing heterogeneously in space on emissions.
- Serious studies on error covariance modeling.
- Hybridation of variational and sequential assimilation methods.