Assimilation d' Ensemble et Bayésianité

Mohamed Jardak, Olivier Talagrand
and
Wael Silini

Laboratoire de Météorologie Dynamique
École Normale Supérieure
Paris, France

Colloque National sur l'Assimilation des Données

Nice, December 18th 2012

Data Assimilation

• **Purpose :** Reconstruct as accurately as possible the state of the atmosphere or oceanic flow, using all available appropriate information :

- The observations proper, which vary in nature, resolution and accuracy, and are distributed more or less regularly in space and time
- The physical laws governing the evolution of the flow, available in practice in the form of a discretized, and necessarily approximate, numerical model.
- 'Asymptotic' properties of the flow, such as, e.g., geostrophic balance of middle latitude. Although they are basically necessary consequences of the physical laws which govern the flow, these properties can be usefully explicitly introduced in the assimilation process.

• **Uncertainty :** Both observations and model are affected by uncertainty.

For some reason, uncertainty is conveniently described by probability distributions.
(see e.g. Jaynes, E.T., Probability Theory : The Logic of Science,2007)

Determine the conditional probability distribution for the state of the system, knowing everything we know. (see e.g. Tarantola 2005)

Assimilation is a problem in Bayesian estimation

Under linearity and gaussianity, the following algorithm achieves Bayesian estimation

- Given the data

$$z = \Gamma x + \zeta, \quad \zeta \in \mathcal{N}([\mu, \Sigma])$$

- The conditional posterior probability distribution is

$$P(x|z) = \mathcal{N}([x^a, P^a])$$

with

$$x^a = \left(\Gamma^T \Sigma^{-1} \Gamma\right)^{-1} \Gamma^T \Sigma^{-1}(z - \mu) \text{ and } P^a = \left(\Gamma^T \Sigma^{-1} \Gamma\right)^{-1}$$

Ready recipe for producing sample of independent realizations of posterior probability distribution :

Perturb data vector additively according to error probability distribution $\mathcal{N}([\mu, \Sigma])$, and compute analysis x^a for each perturbed data vector.

Variational Assimilation

The following algorithm produces a sample of independent realizations of the probability distribution of the state of the system, conditioned by the data x_0^b and y_k.

- Available data
  1. Background estimate at $t = 0$, $x_0^b = x_0 + \xi_0^b$, $\xi_0^b \in \mathcal{N}([0, \left[P_0^b\right]^{1/2}])$
  2. Observations at $t = 0$, $y_k = \mathrm{H}_k x_k + \epsilon_k, \epsilon_k \in \mathcal{N}([0, [\mathrm{R}_k]^{1/2}])$
  3. Model (supposed to be exact) $x_{k+1} = \mathrm{M}_k x_k$, and $k = 0, \cdots, K-1$
  4. Errors ξ_0^b and ϵ_k assumed to be unbiased and uncorrelated in time.
  5. H_k and M_k assumed linear
- The optimal state (mean of the Bayesian Gaussian pdf) at $t = 0$ minimizes the objective function

$$\begin{cases} \mathcal{J}(\xi_0) = \dfrac{1}{2}(x_0^b - \xi_0)^T \left[P_0^b\right]^{-1} (x_0^b - \xi_0) + \dfrac{1}{2}\displaystyle\sum_k (y_k - \mathrm{H}_k\xi_k)^T \mathrm{R}_k^{-1} (y_k - \mathrm{H}_k\xi_k) \\ \xi_{k+1} = \mathrm{M}_k \xi_k \end{cases}$$

What happens under nonlinearity and non-Gaussianity ?

Objectives

- Objectively evaluate the Ens/4D-Var as an ensemble estimator in the non-linear and non-Gaussian cases.
- Evaluate as far as possible the Bayesianity of the ensemble produced in the non-linear and non-Gaussian cases .
- Compare with other existant ensemble algorithm schemes (EnKF and PF) .

Ens/4D-Var method

- for iens $= 1$: Nens
  1. perturb the data
     - $x_0^b(iens) \in \mathcal{N}(\bar{x}_0^b, \mathrm{B}^{1/2})$
     - $y_k(iens) \in \mathcal{N}(\bar{y}_k, \mathrm{R}_k^{1/2})$
  2. perform a 4D-Var to find the optimal initial ensemble member solution .

$$x_0^{opt}(iens) = \min_{x \in \mathfrak{A}} \mathfrak{J}_{iens}(X)$$

  3. find the optimal ensemble member trajectory.

$$x^{opt}(t, iens) = \underbrace{\mathfrak{M}_{t \to 0}}_{\color{red}{non-linear\ model}} (x_0^{opt}(iens))$$

- end for

How to objectively evaluate the Bayesian character of an ensemble estimation procedure ?

- Bayesianity implies reliability therefore lack of reliability implies lack of Bayesianity.

reliability is the statistical consistency between the predicted probability of occurrence and the observed frequency of occurrence.

- Consistency : Under linearity the expectation of the objective function at its minimum is half the number of observations p

$$\mathbb{E}(\mathfrak{J}(x_{opt})) = \frac{p}{2}$$

and under Gaussianity we have

$$Var(\mathfrak{J}(x_{opt})) = p.$$

Testing reliability

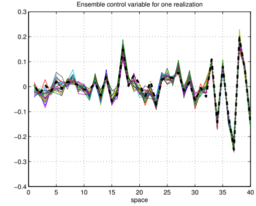
- rank histogram.
- reliability diagram.
- Brier scores :

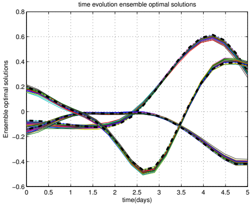
$$\mathbb{B} = \underbrace{\frac{1}{p_c(1-p_c)}}_{uncertainty} \left[\underbrace{\int_0^1 (p'-p)^2 g(p)dp}_{\mathbb{B}_{sc}} + \underbrace{\int_0^1 p'(1-p')g(p)dp}_{\mathbb{B}_{sv}} \right]$$

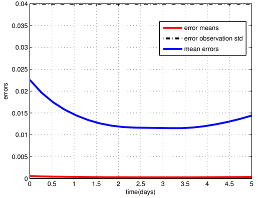
  - p predicted probability.
  - g the frequency with which p has been predicted.
  - $p'(p)$ observed frequency.
  - p_c the frequency of occurrence of the event $\mathscr{E}$ under observation.

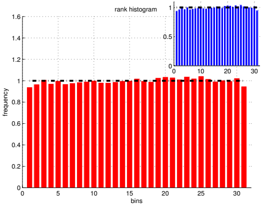
Ens/4D-Var results : the Lorenz96 model

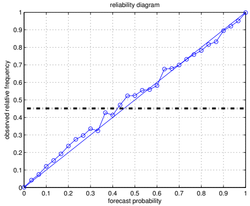
Linear case : 5 days time length

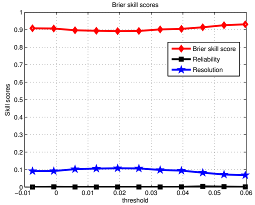




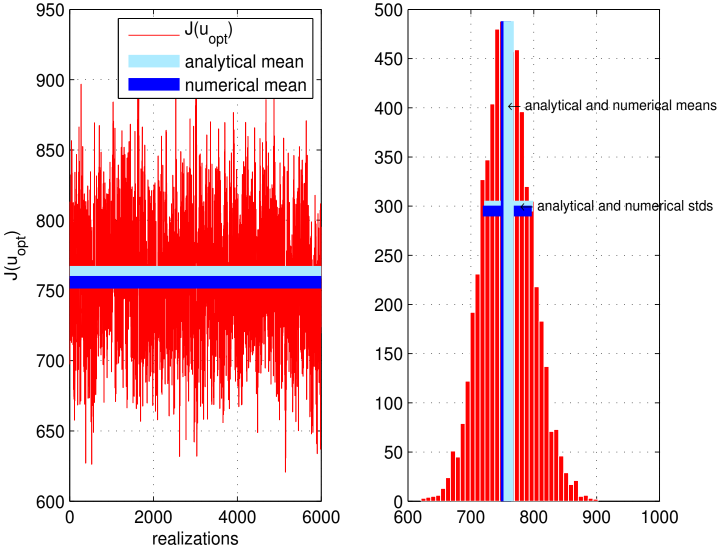








Ens/4D-Var results : consistency



The Lorenz96 model

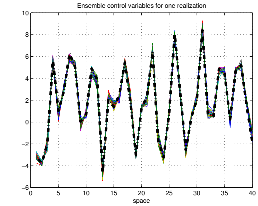
- Forward model

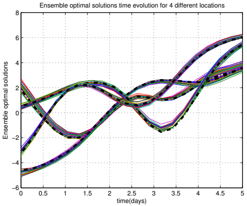
$$\frac{dx_k}{dt} = (x_{k+1} - x_{k-2})x_{k-1} - x_k + F \quad \text{for} \quad k = 1, \cdots, N$$

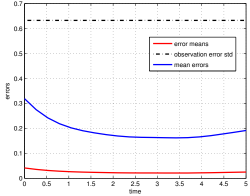
- Set-up parameters :
  1. the index k is cyclic so that $x_{k-N} = x_{k+N} = x_k$.
  2. $F = 8$, external driving force.
  3. x_k, a damping term.
  4. $N = 40$, the system size.
  5. $Nens = 30$, number of ensemble members.
  6. $\frac{1}{\lambda_{max}} \simeq 2.5 days$, λ_{max} the largest Lyapunov exponent.
  7. $\Delta t = 0.05 = 6 hours$, the time step.

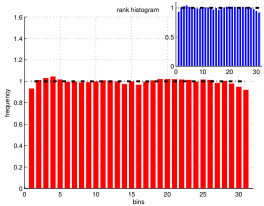
Ens/4D-Var results : the Lorenz96 model

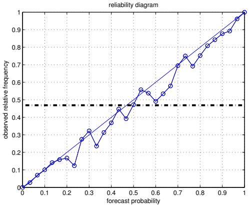
Nonlinear case : 5 days time length

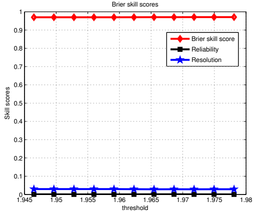






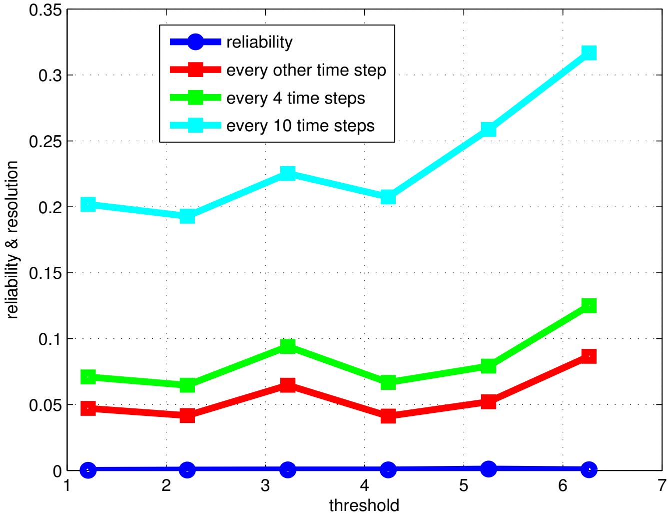




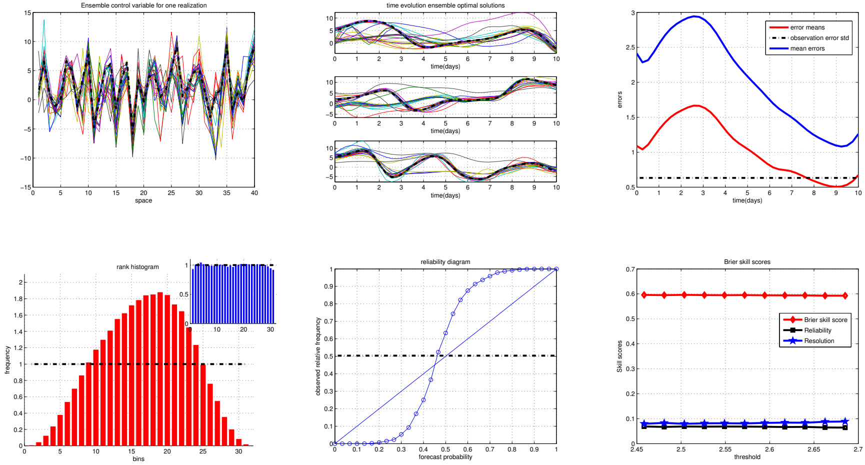


Ens/4D-Var results : observation frequency impact

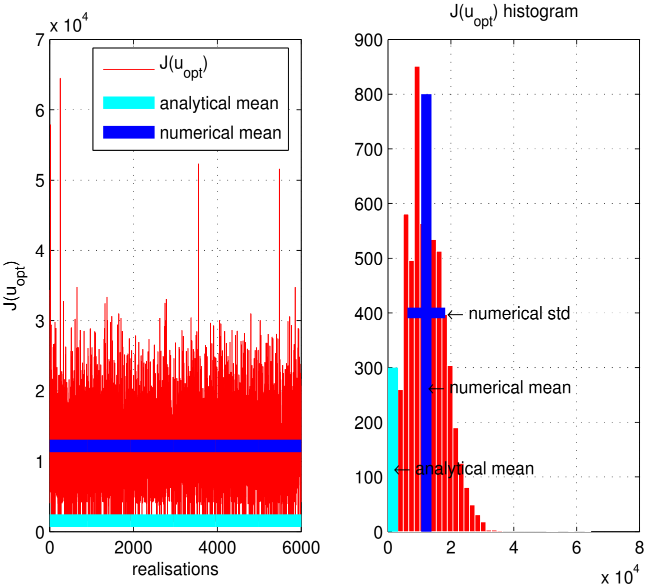
Impact on the reliability and resolution



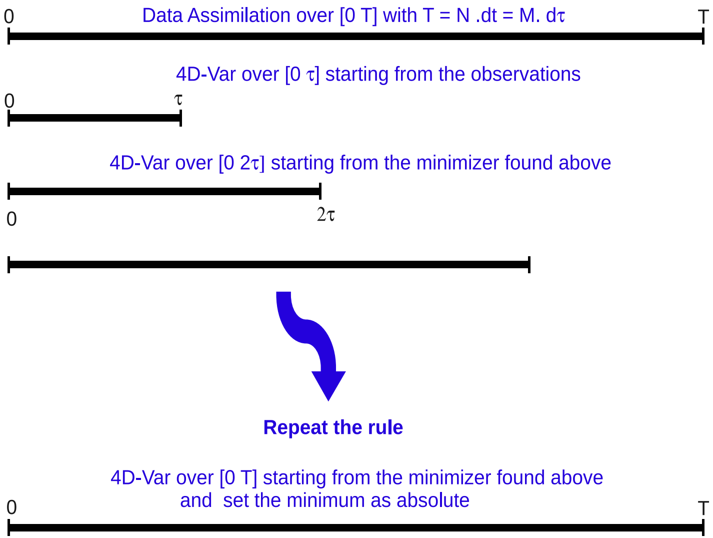
Nonlinear case : 10 days time length



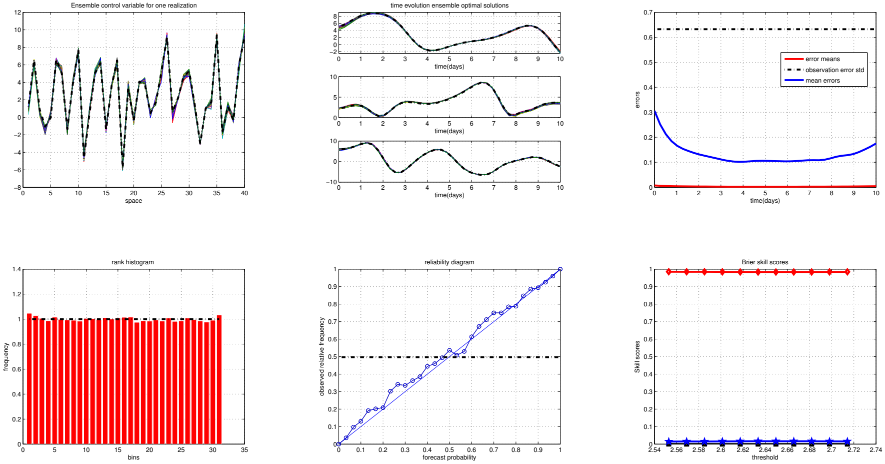
Ens/4D-Var results : loss of consistency



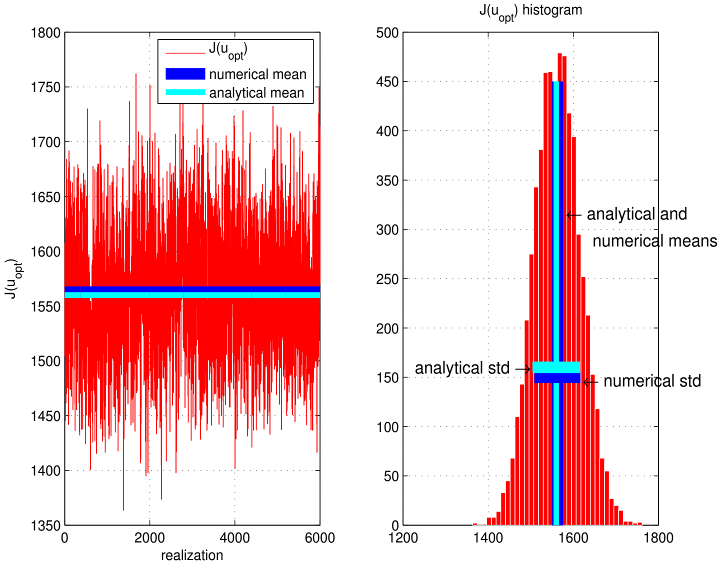
Quasi-Static Variational Assimilation (QSVA)



Nonlinear case : 10 days time length with QSVA

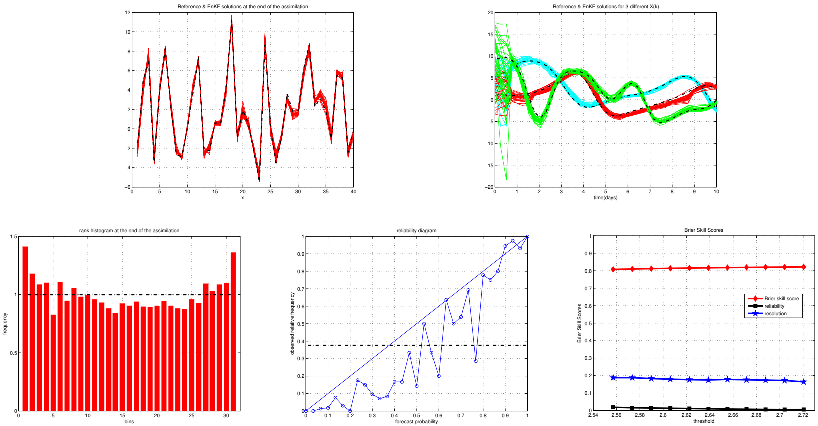


Ens/4D-Var results : consistency

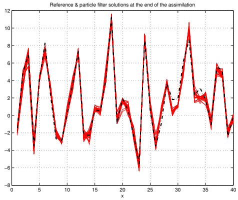


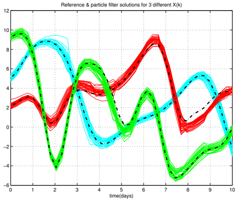
EnKF results : the Lorenz96 model

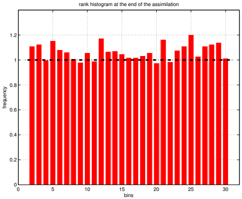
Nonlinear case : 10 days time length

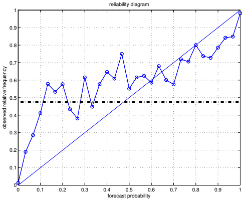


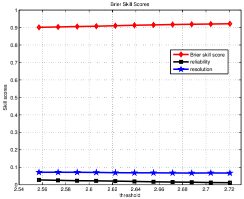
Nonlinear case : 10 days time length



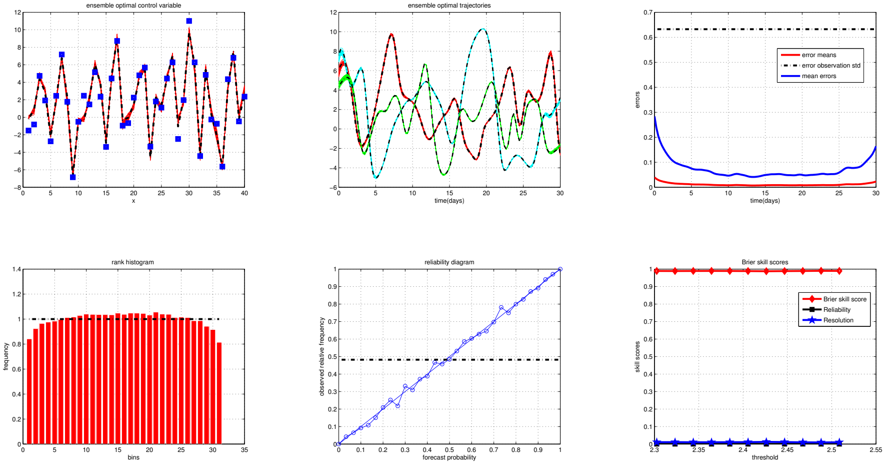






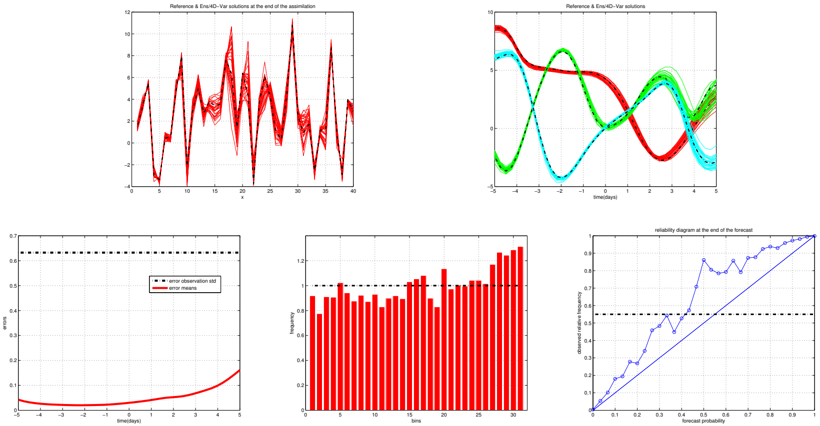


Nonlinear case : 30 days time length with QSVA



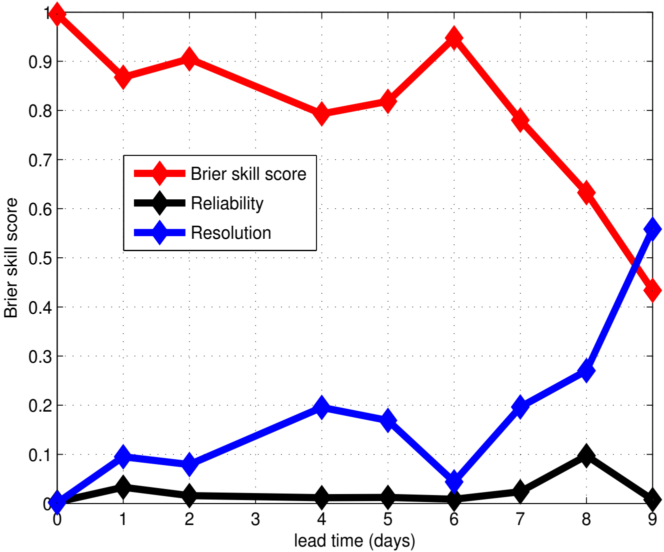
Ens/4D-Var results : 5 days forecast

Nonlinear case : forecasting

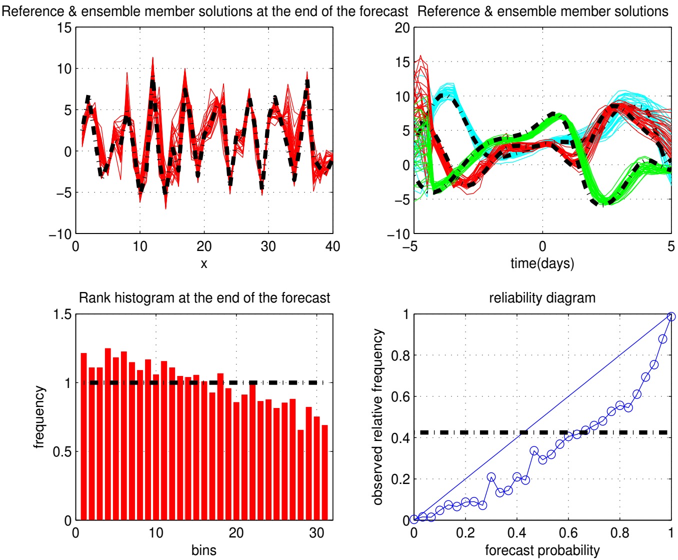


Ens/4D-Var results : forecasting

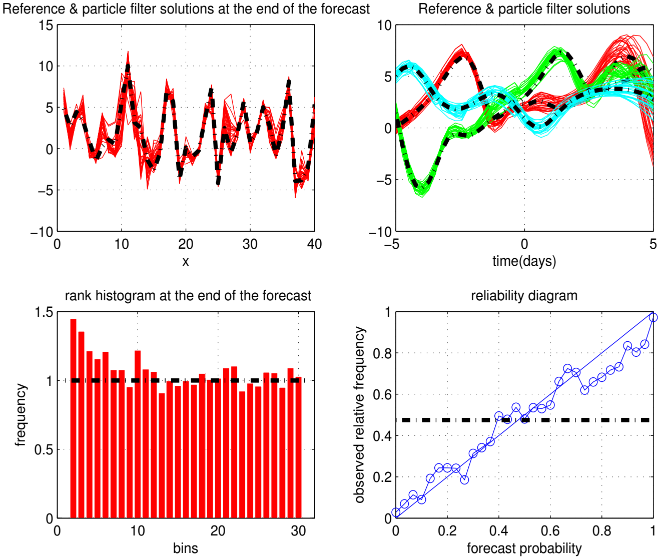
Brier Skill Scores as a function of the lead time



EnKF results : 5days forecast



PF results : 5days forecast



Weak constraint Ens/4D-Var

- define the objective function.

$$\mathfrak{J}(x, \eta_1, \eta_2, \cdots . \eta_{N-1}, \eta_N) = \frac{1}{2}\left\{(x - x_b)^T \mathrm{B}^{-1}(x - x_b)\right\} \; +$$

$$\frac{1}{2}\sum_{i=0}^{N}\left\{(y_i - H_i(x_i))^T \mathrm{R}_i^{-1}(y_i - H_i(x_i))\right\} + \frac{1}{2}\sum_{i=1}^{N}\eta_i^T \mathrm{Q}_i^{-1}\eta_i$$

1. B background error covariance matrix and R observation error covariance matrix.
2. Q model error covariance matrix.
3. $H : \mathbb{R}^{state} \longrightarrow \mathbb{R}^{obs}$ observation operator.
4. x_b background state vector and y_i observation vector at time $t = t_i$.
5. η_i model error vector at $t = t_i$ with $x(t_i) = \mathfrak{M}_{t_i \leftarrow t_{i-1}}(x(t_{i-1})) + \eta_i$

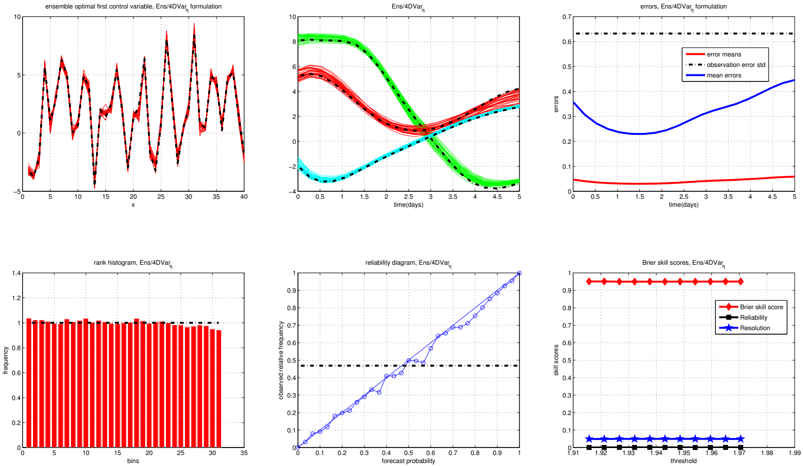
- find the optimal control variable $(x_0^{opt}, \eta_1^{opt}, \eta_2^{opt}, \cdots . \eta_N^{opt})$ and the optimal trajectory x^{opt}.

$$(x_0^{opt}, \eta_1^{opt}, \eta_2^{opt}, \cdots . \eta_N^{opt}) = \min_{x, \eta_1, \eta_2, \cdots . \eta_N \in \mathfrak{A}} \mathfrak{J}(x, \eta_1, \eta_2, \cdots . \eta_N)$$

$$x^{opt}(t_i) = \mathfrak{M}_{t_i \leftarrow t_{i-1}}(\mathfrak{M}_{t_{i-1} \leftarrow t_{i-2}} \cdot\cdot (\mathfrak{M}_{t_2 \leftarrow t_1}(\mathfrak{M}_{t_1 \leftarrow t_0}(x_0^{opt}) + \eta_1^{opt}) + \eta_2^{opt}) \cdot\cdot + \eta_{i-1}^{opt}) + \eta_i^{opt}$$

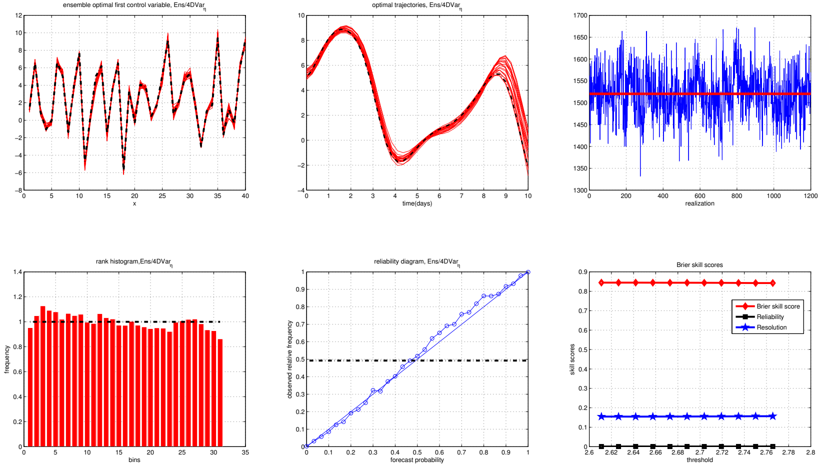
Weak constraint Ens/4D-Var results : the Lorenz96 model

Ens/4DVar$_\eta$: 5 days time length



Weak constraint Ens/4D-Var results : the Lorenz96 model

Ens/4DVar$_\eta$: 10 days time length



Summary

- Under non-linearity and non-Gaussianity the Ens/4D-Var is a reliable and consistent ensemble estimator (provided the QSVA is used for long DA windows) .
- Ens/4D-Var is at least as good an estimator as EnKF and PF.
- Similar results have been obtained for the Kuramuto-Sivashinsky model.

Pros

- Easy to implement when having a 4D-Var code
- Highly parallelizable
- No problems with algorithm stability (i.e. no ensemble collapse, no need for localization and inflation, no need for weight resampling)
- Propagates information in both ways and takes into account temporally correlated errors

Cons

- Costly (Nens 4D-Var assimilations)
- Emperical

Future work

- Evaluate the performance of Ens/4D-Var on a simple geophysical and meteorological model (Shallow water on the sphere model) (in progress).
- Investigate the 4D-VarAUS method.

Merci