Assimilation d'ensemble et bayésianité

Mohamed Jardak[1,2] et Olivier Talagrand[1]
1. Laboratoire de Météorologie Dynamique/IPSL
École Normale Supérieure, Paris, France
2. Meteorological Office, Exeter, Royaume-Uni

Colloque National sur l'Assimilation de données

Toulouse

3 Décembre 2014

Purpose of assimilation : reconstruct as accurately as possible the state of the atmospheric or oceanic flow, using all available appropriate information. The latter essentially consists of

- 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 latitudes. Although they basically are necessary consequences of the physical laws which govern the flow, these properties can usefully be explicitly introduced in the assimilation process.

Both observations and 'model' are affected with some uncertainty ⇒ uncertainty on the estimate.

For some reason, uncertainty is conveniently described by probability distributions (don't know too well why, but it works).

Assimilation is considered here as a problem in bayesian estimation.

Determine the conditional probability distribution for the state of the system, knowing everything we know.

Jaynes, E. T., 2003, *Probability theory: the logic of science,* Cambridge University Press

Tarantola, A., 2005, *Inverse Problem Theory and Methods for Model Parameter Estimation*, Society for Industrial and Applied Mathematics (http://www.ipgp.jussieu.fr/~tarantola/Files/Professional/Books/InverseProblemTheory.pdf)

Data of the form

$$z = \Gamma x + \zeta, \qquad \zeta \sim \mathcal{N}[\mu, S]$$

Known data vector z belongs to *data space* $\mathcal{D}$, $dim\,\mathcal{D} = m$,
Unknown state vector x belongs to *state space* $\mathcal{X}$, $dim\,\mathcal{X} = n$
Γ known (mxn)-matrix, ζ unknown 'error'

Then conditional probability distribution is

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

where

$$x^a = (\Gamma^T S^{-1} \Gamma)^{-1} \Gamma^T S^{-1} [z - \mu]$$
$$P^a = (\Gamma^T S^{-1} \Gamma)^{-1}$$

Determinacy condition : rank$\Gamma = n$. Requires $m \geq n$.

Variational form.

Conditional expectation x^a minimizes following scalar *objective function*, defined on state space $\mathcal{X}$

$$\xi \in \mathcal{X} \rightarrow \mathcal{J}(\xi) \equiv (1/2)\,[\Gamma\xi - (z-\mu)]^{\mathrm{T}}\, S^{-1}\,[\Gamma\xi - (z-\mu)]$$

Variational assimilation, implemented heuristically in many places on (not too) nonlinear data operators Γ.

$$P^a = [\partial^2 \mathcal{J} / \partial \xi^2]^{-1}$$

Conditional probability distribution

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

with

$$x^a = (\Gamma^T S^{-1} \Gamma)^{-1} \Gamma^T S^{-1} [z - \mu]$$
$$P^a = (\Gamma^T S^{-1} \Gamma)^{-1}$$

Ready recipe for determining Monte-Carlo sample of conditional pdf $P(x \mid z)$:

- Perturb data vector z according to its own error probability distribution

$$z \rightarrow z' = z + \delta, \qquad \delta \sim \mathcal{N}[0, S]$$

and compute

$$x'^a = (\Gamma^T S^{-1} \Gamma)^{-1} \Gamma^T S^{-1} [z' - \mu]$$

x'^a is distributed according to $\mathcal{N}[x^a, P^a]$

Ensemble Variational Assimilation (*EnsVar*) implements that algorithm, the expectations $x^{\text{`}a}$ being computed by standard variational assimilation (optimization)

Purpose of the present work

- Objectively evaluate EnsVar as a probabilistic estimator in nonlinear and/or non-Gaussian cases.

- Objectively compare with other existing ensemble assimilation algorithms : *Ensemble Kalman Filter* (*EnKF*), *Particle Filters* (*PF*)

- Simulations performed on two small-dimensional chaotic systems, the Lorenz'96 model and the Kuramoto-Sivashinsky equation

Experimental procedure (1)

0. Define a *reference solution* x_t^r by integration of the numerical model

1. Produce 'observations' at successive times t_k of the form

$$y_k = H_k x_k + \varepsilon_k$$

where H_k is (usually, but not necessarily) the unit operator, and ε_k is a random (usually, but not necessarily, Gaussian) 'observation error'.

Experimental procedure (2)

2. For given observations y_k, repeat N_{ens} times the following process

- 'Perturb' the observations y_k as follows

$$y_k \rightarrow \; z_k = y_k + \delta_k$$

where δ_k is an independent realization of the probability distribution which has produced ε_k.

- Assimilate the 'perturbed' observations z_k by variational assimilation

This produces N_{ens} (=30) model solutions over the assimilation window, considered as making up a tentative sample of the conditional probability distribution for the state of the observed system over the assimilation window.

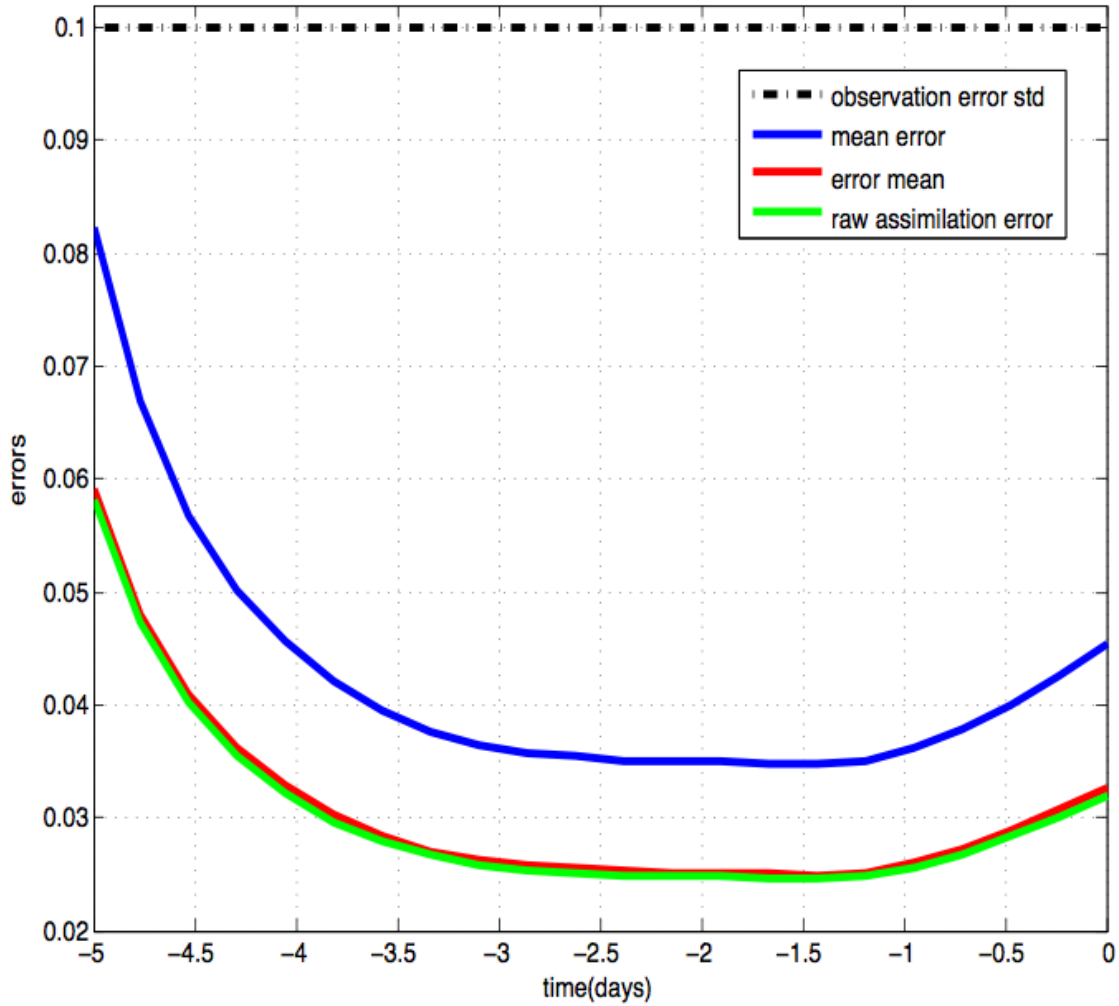
The process 1-2 is then repeated over N_{real} successive assimilation windows. Validation is performed on the set of N_{real} (=9000) ensemble assimilations thus obtained.

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.
  8. frequency of observations : every 12 hours.
  9. number of realizations : 9000 realizations.



Linearized Lorenz'96. 5 days

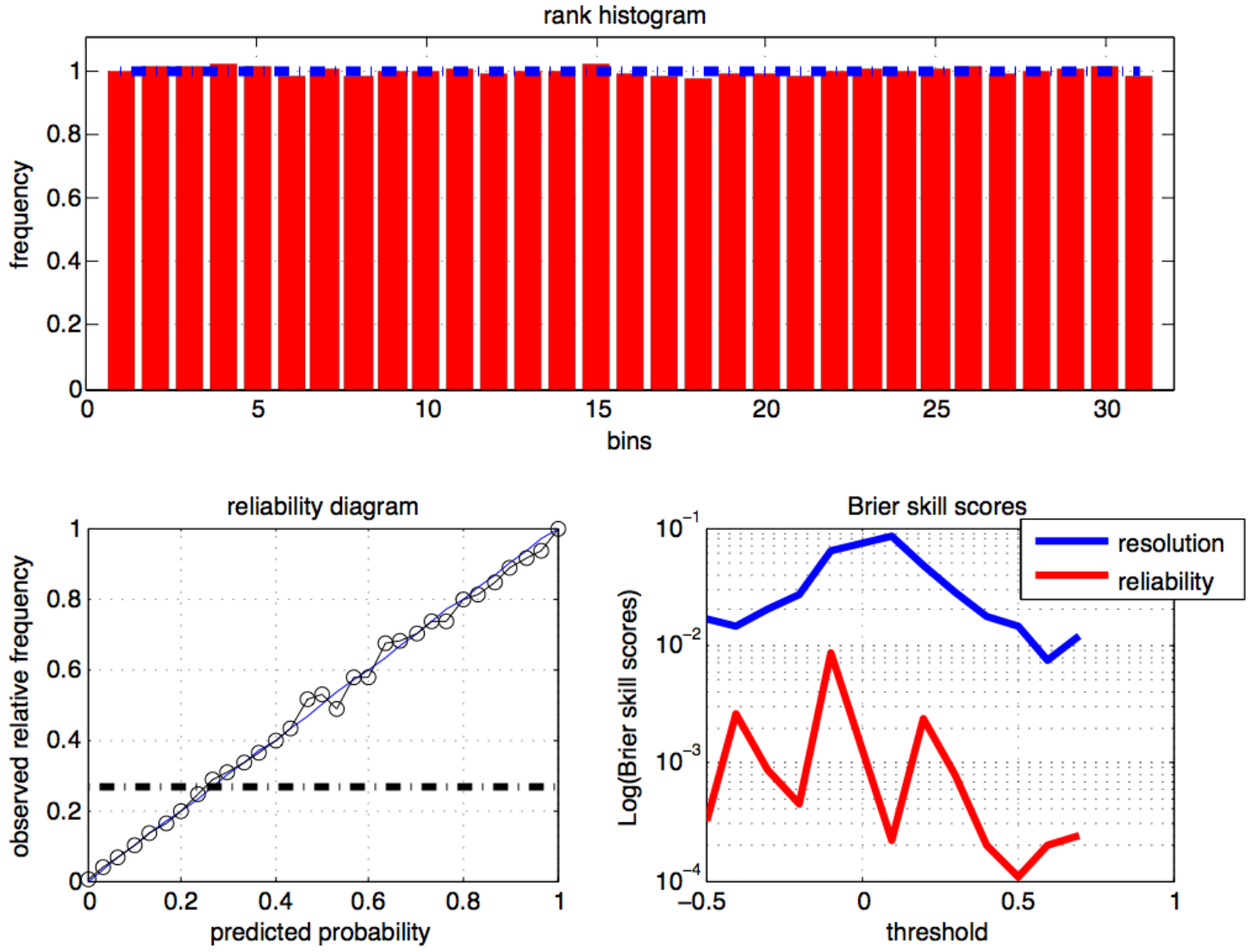
How to objectively evaluate the performance of an ensemble (or more generally probabilistic) estimation system ?

- There is no general objective criterion for Bayesianity

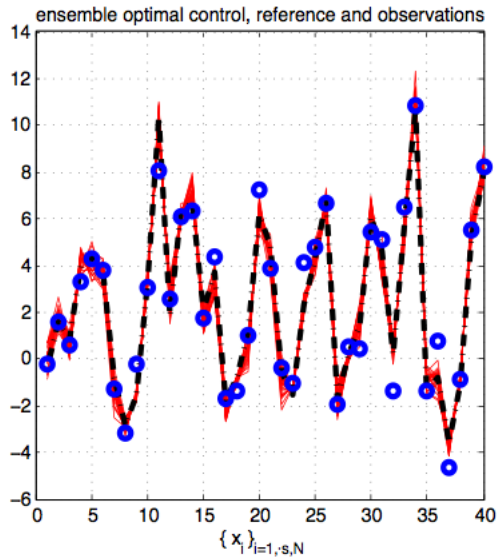
- We use instead the weaker property of *reliability*, *i. e.* statistical consistency between predicted probabilities and observed frequencies of occurrence (it rains with frequency 40% in the circonstances where I have predicted 40% probability for rain).

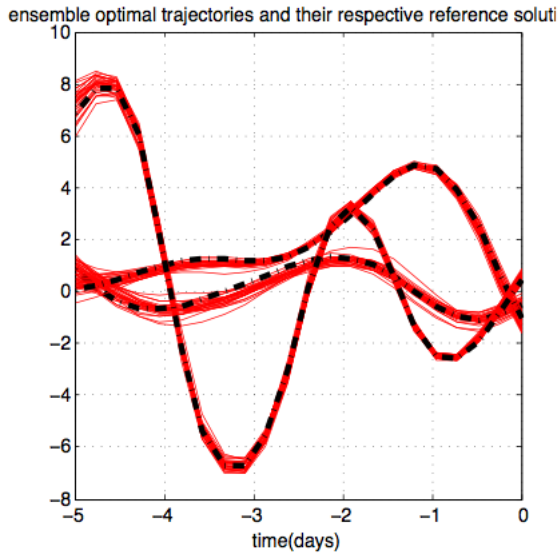
Reliability can be objectively validated, provided a large enough sample of realizations of the estimation system is available.

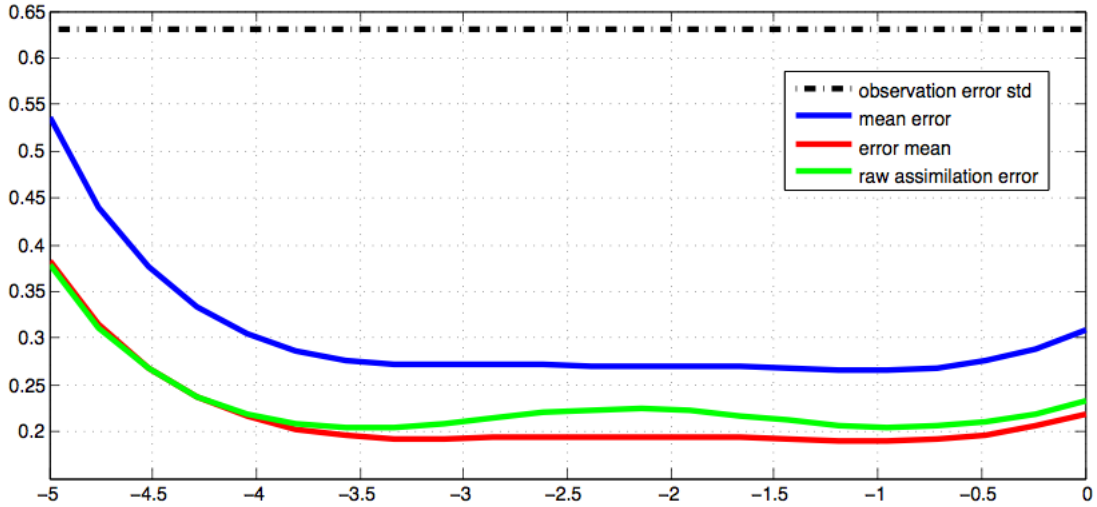
Bayesianity implies reliability, the converse not being true.



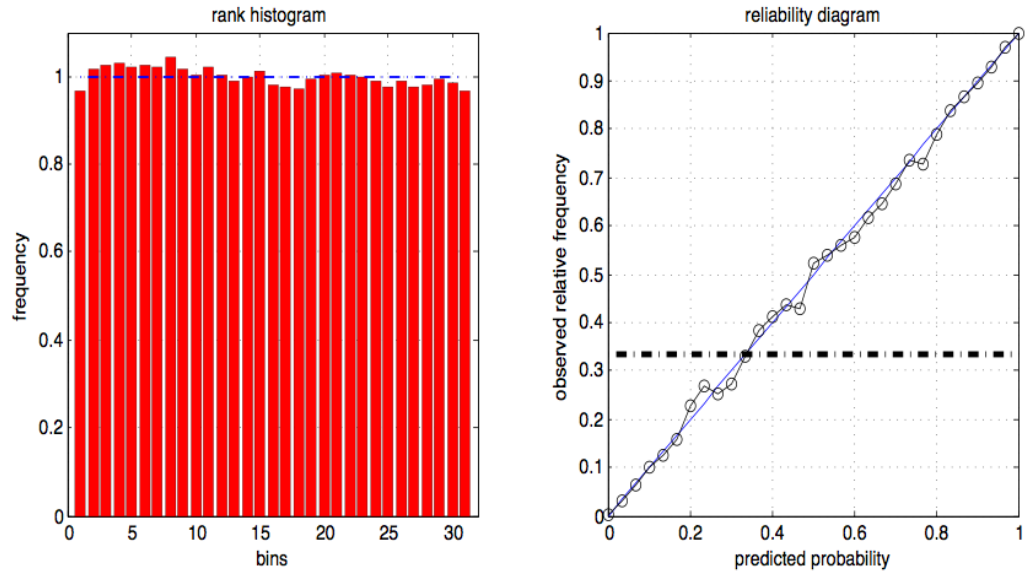
Linearized Lorenz'96. 5 days

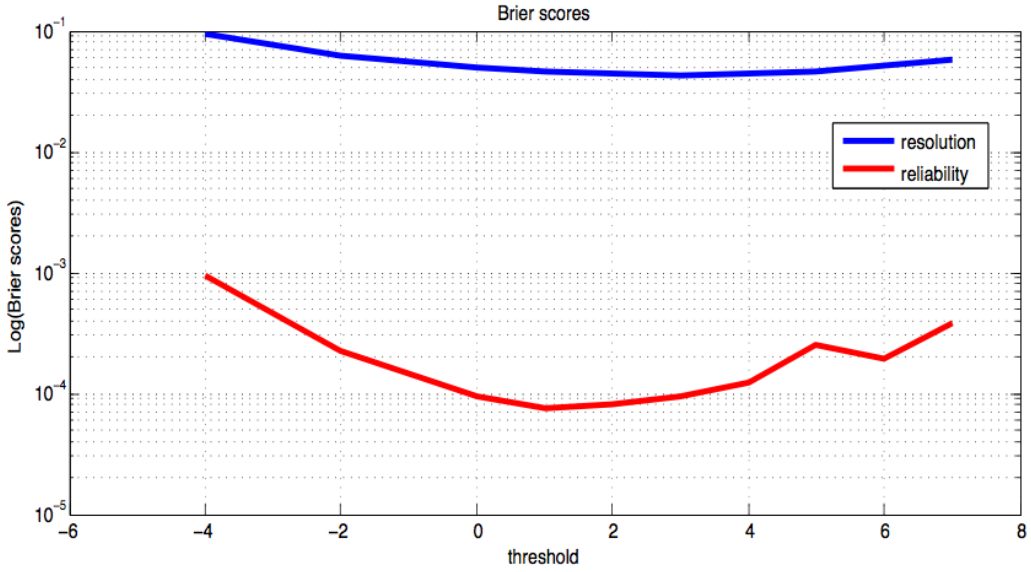




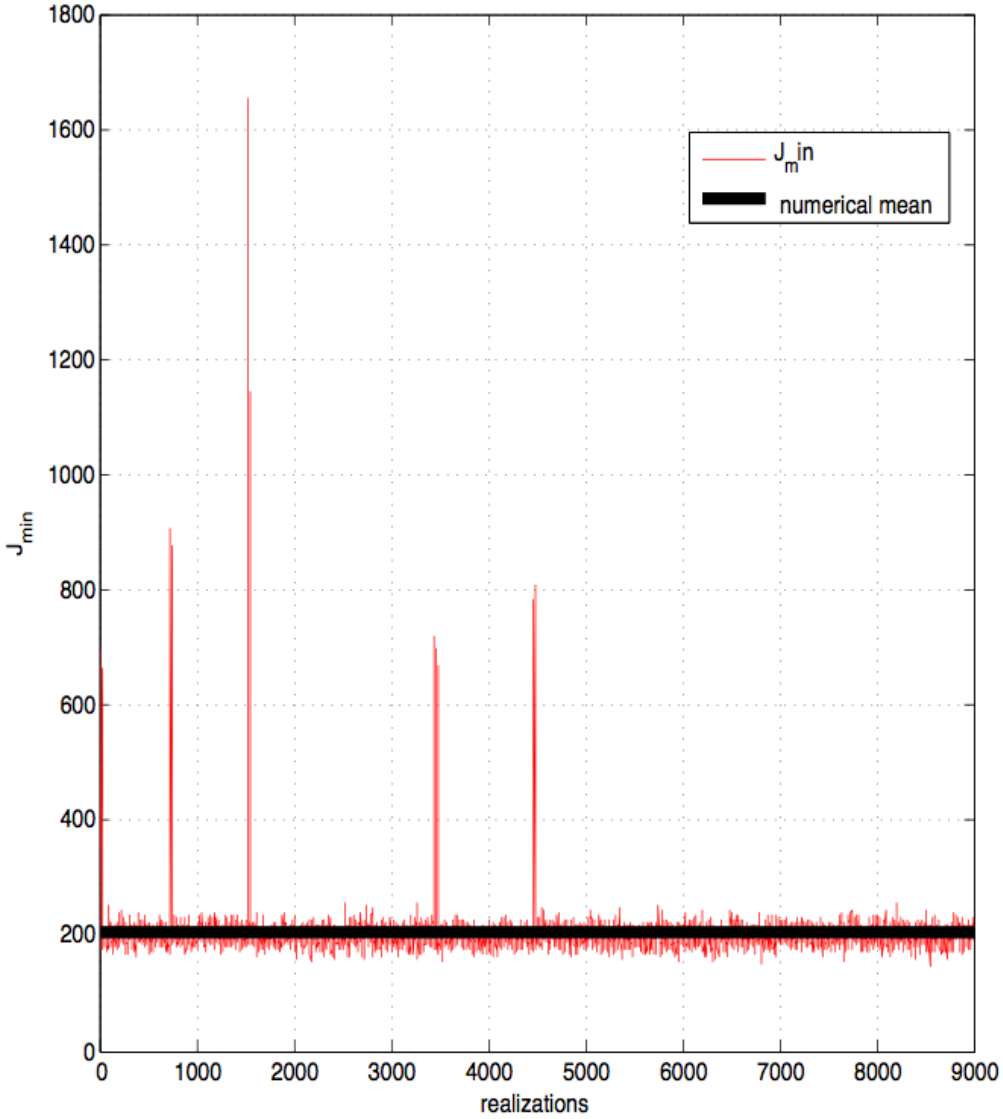


Nonlinear Lorenz'96. 5 days



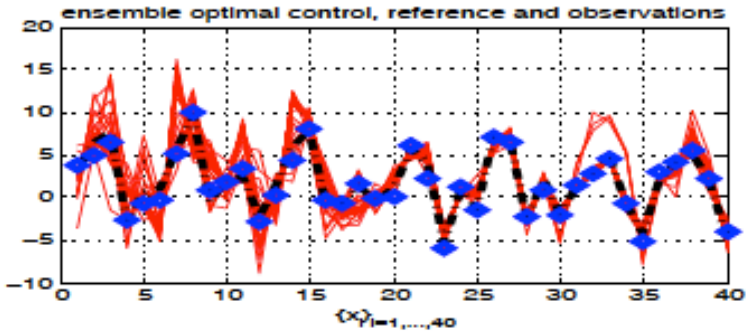


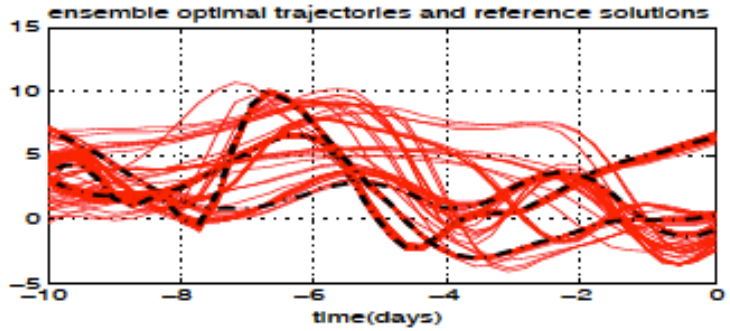
Nonlinear Lorenz'96. 5 days

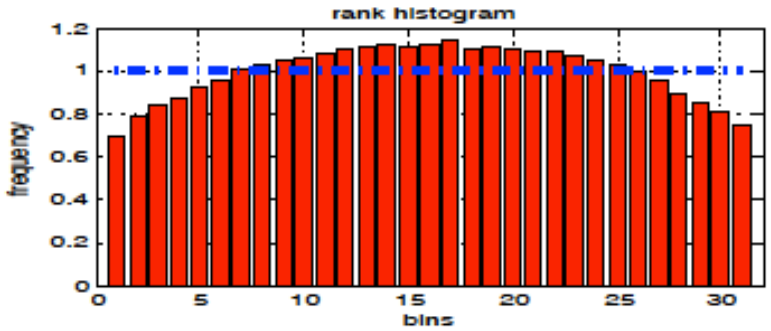


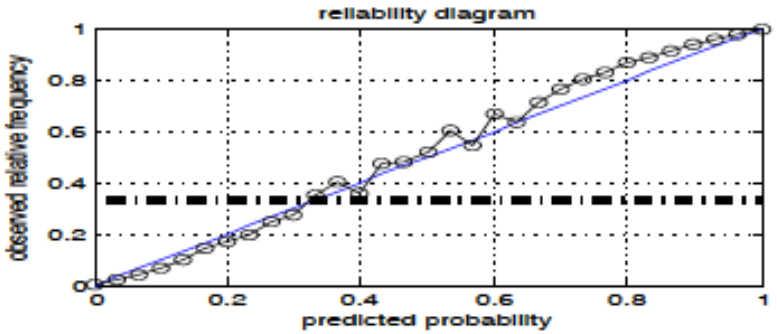
Nonlinear Lorenz'96. 5 days. Histogram of $\mathcal{J}_{min}$

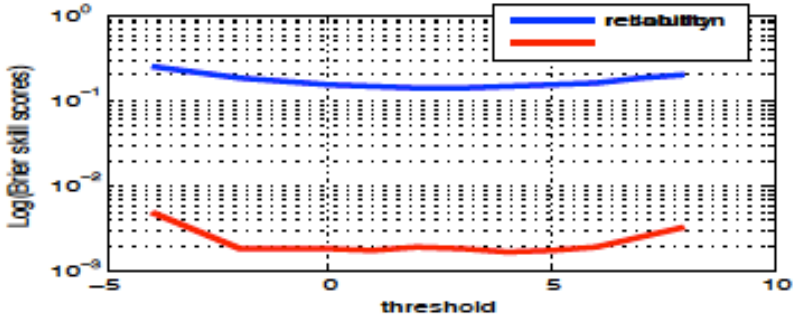
EnsVar : the non-linear Lorenz96 model (10 days $\simeq$ 2 TU)

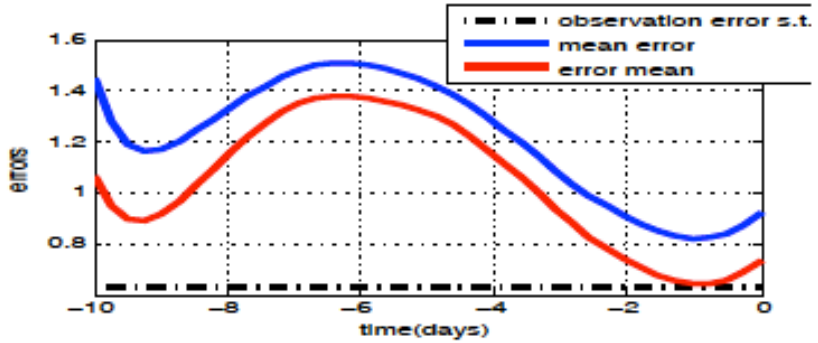




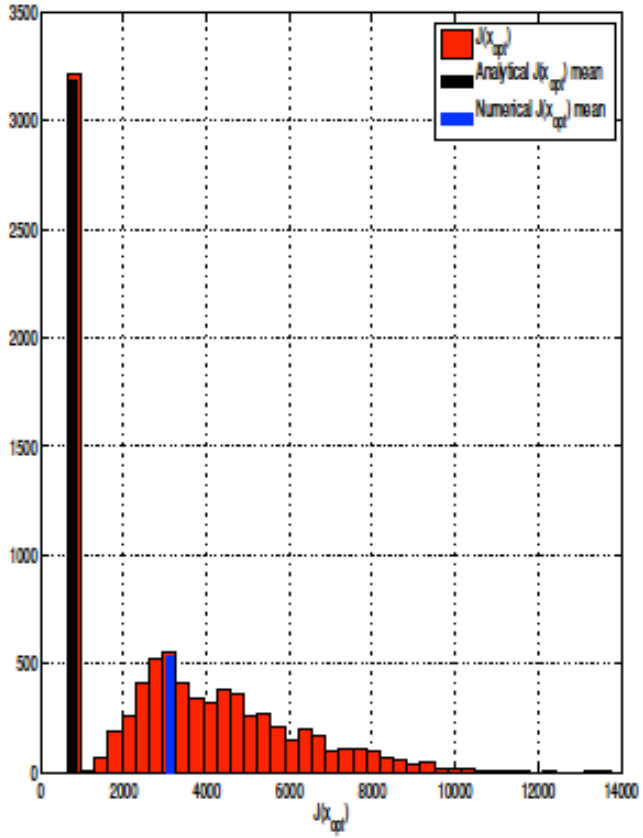






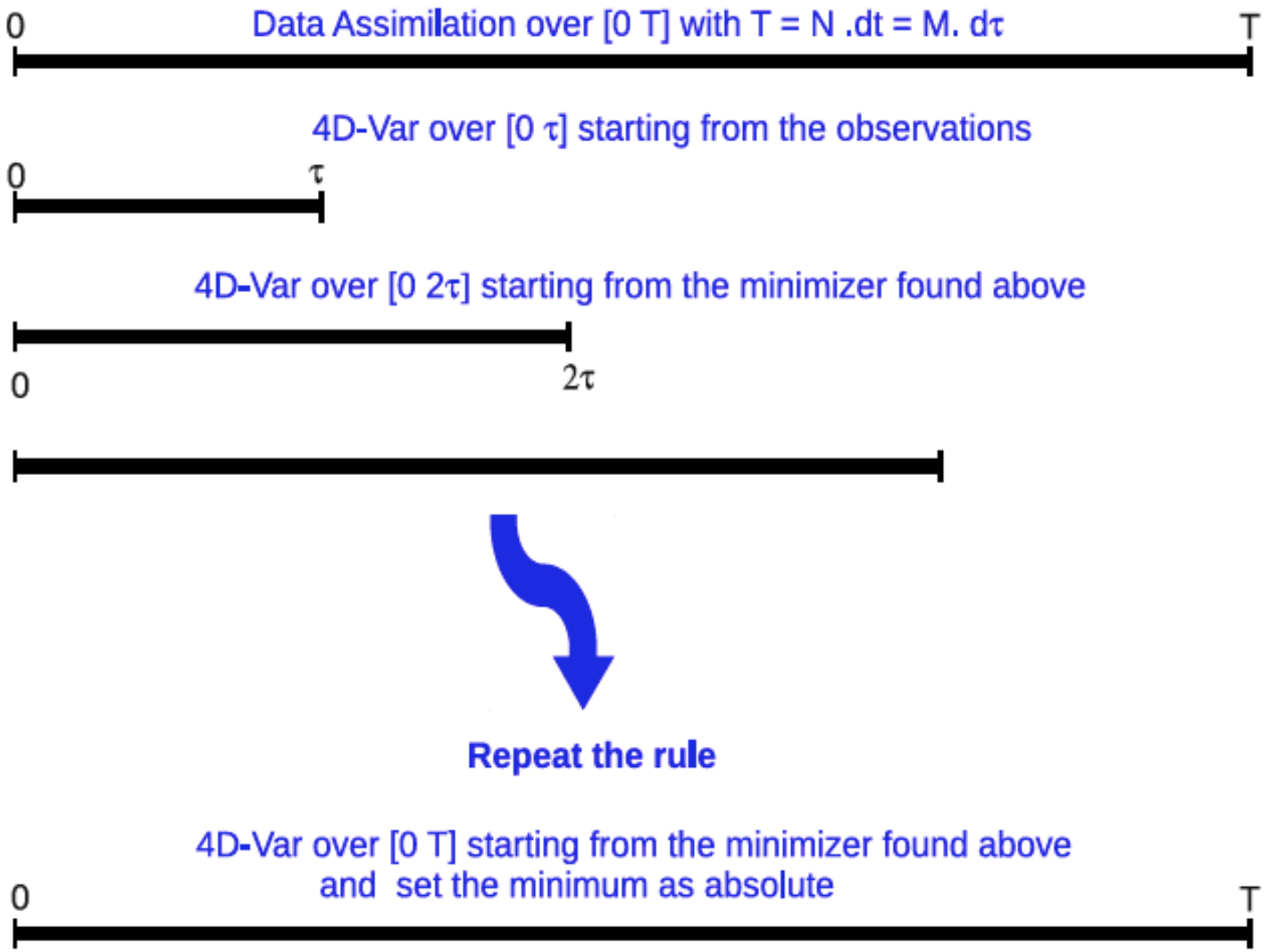


EnsVar : consistency

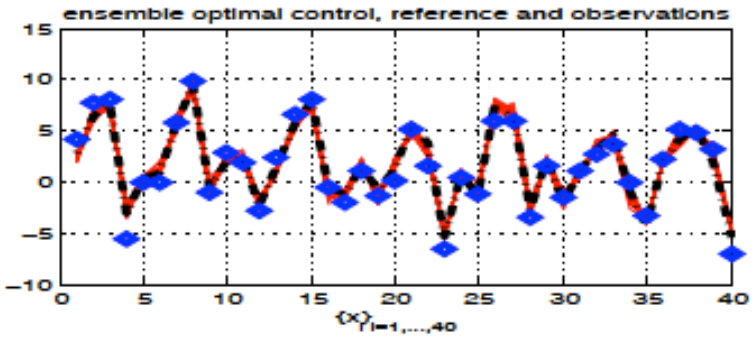


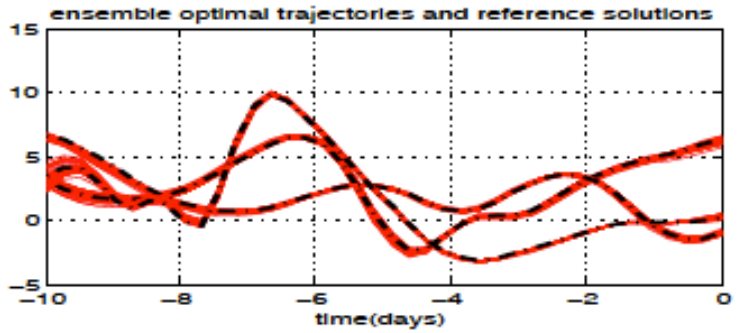
Nonlinear Lorenz'96. 10 days. Histogram of $\mathcal{J}_{min}$

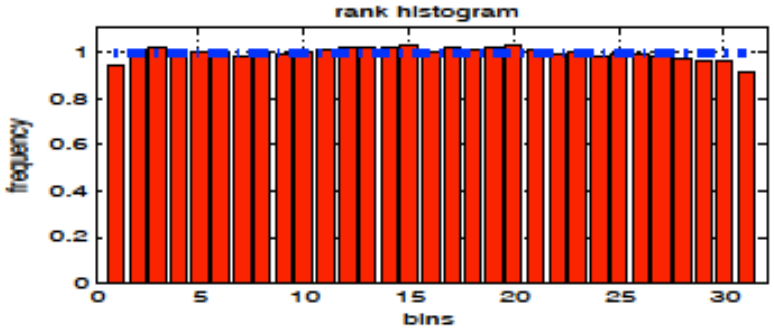
Quasi-Static Variational Assimilation (QSVA)

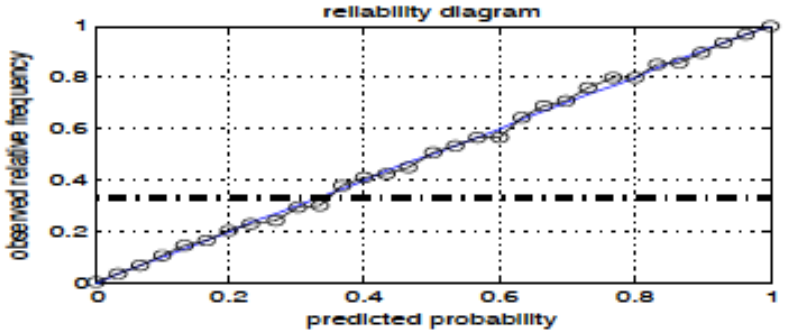


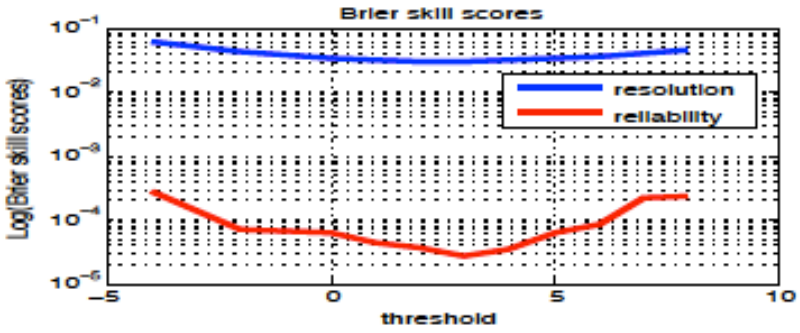
EnsVar : the non-linear Lorenz96 model 10 days with QSVA

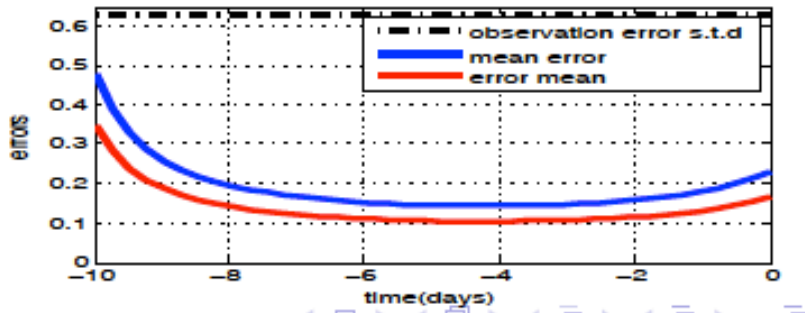




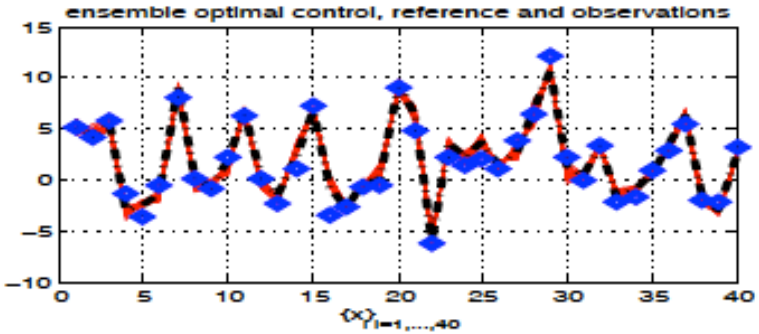


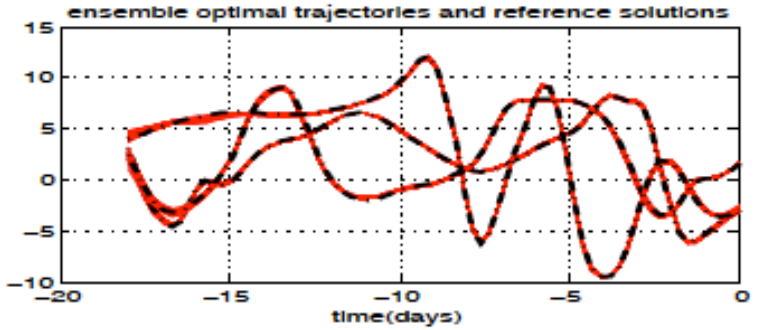


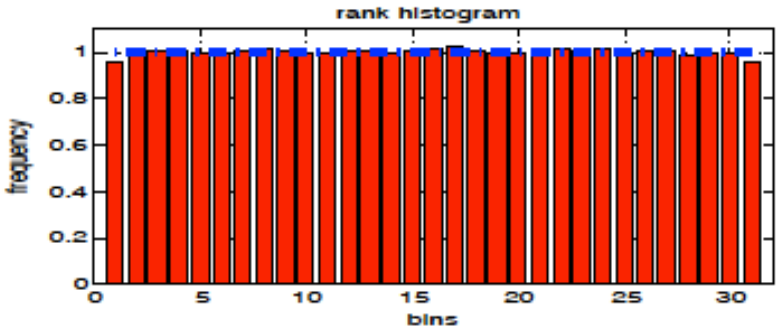


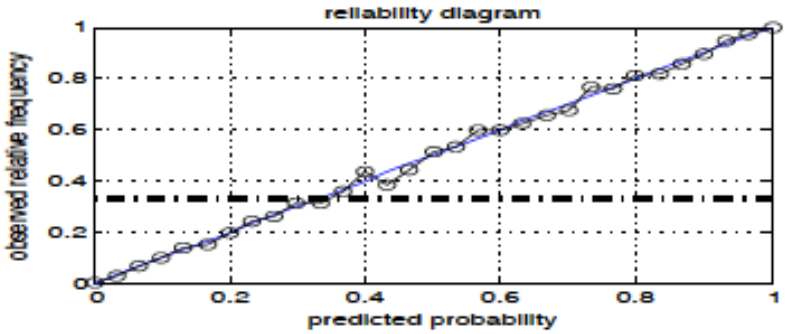


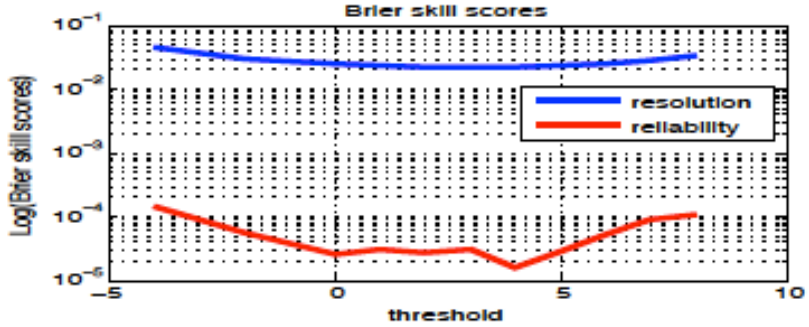
EnsVar : the non-linear Lorenz96 model 18 days with QSVA

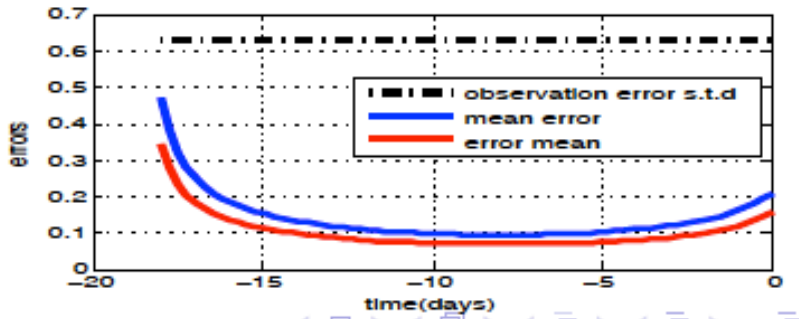






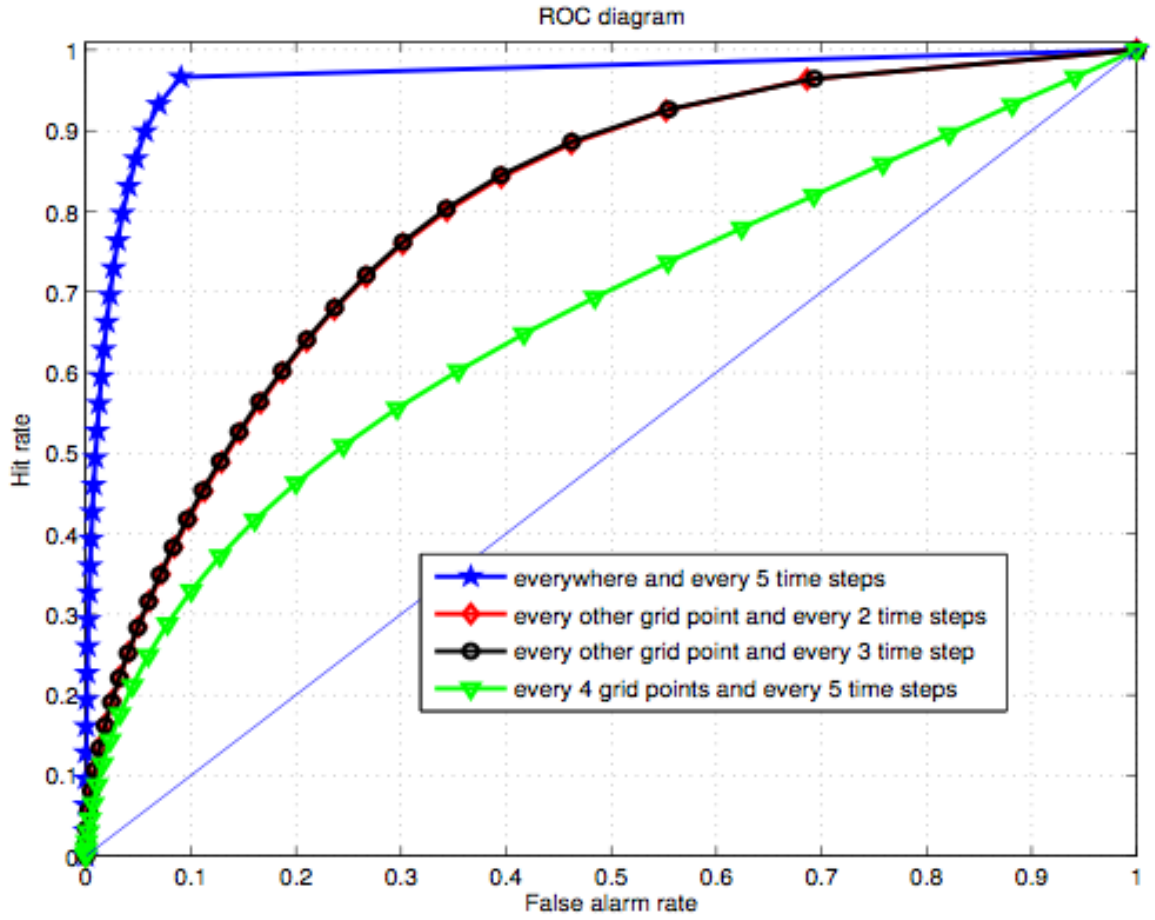






EnsVar : observation frequency impact

Impact of the resolution



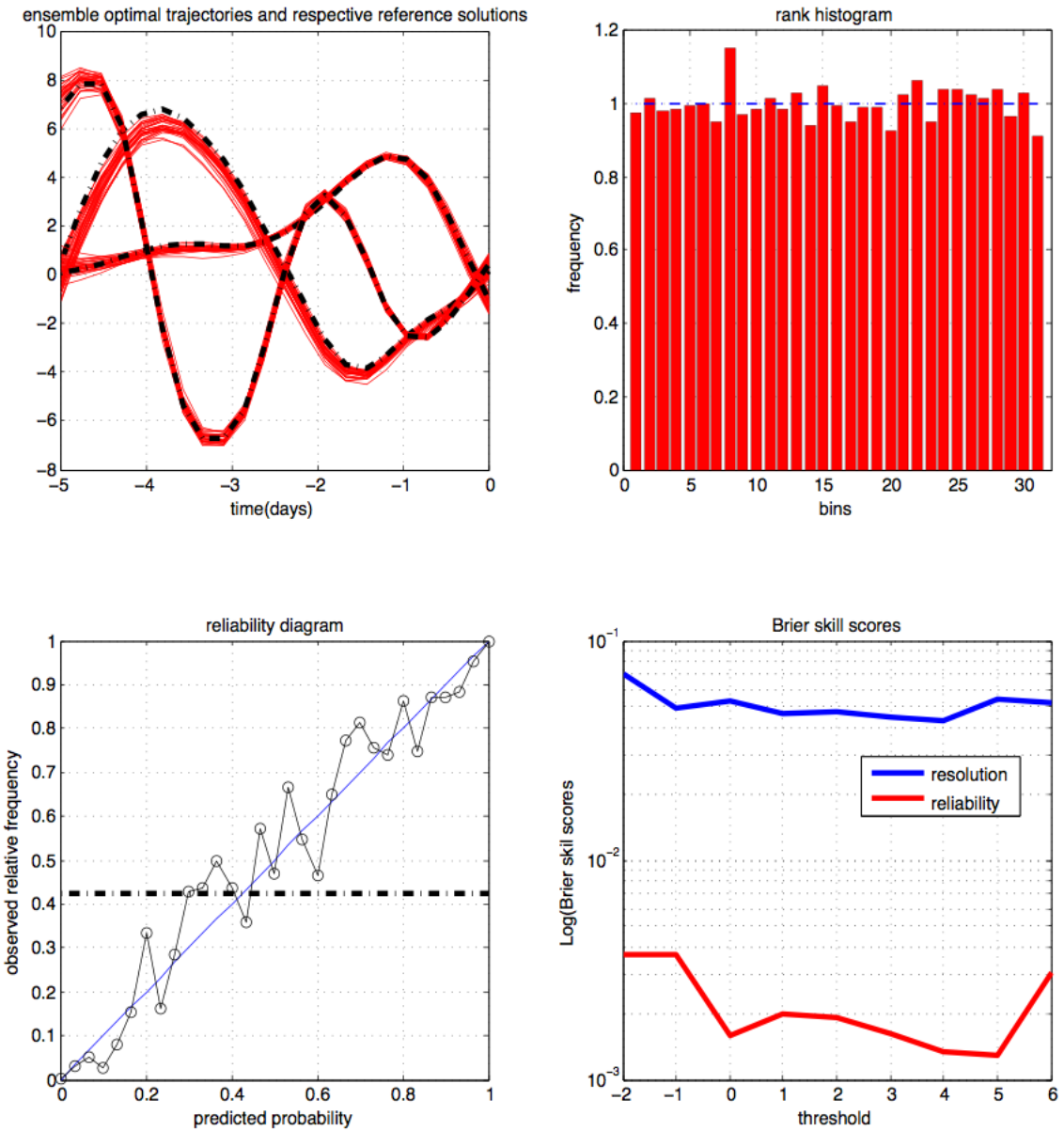
Relative Operating Curve (area below the curve is measure of resolution)

- Results are independent of the Gaussian character of the observation errors (trials have been made with various parobability distributions)

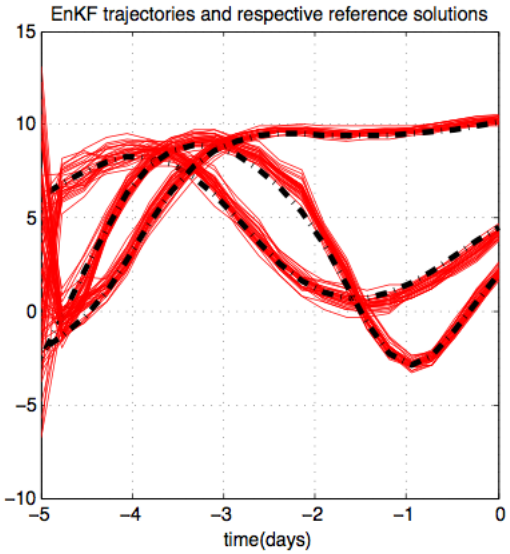
- Ensembles produced by EnsVar are very close to Gaussian, even in strongly nonlinear cases.

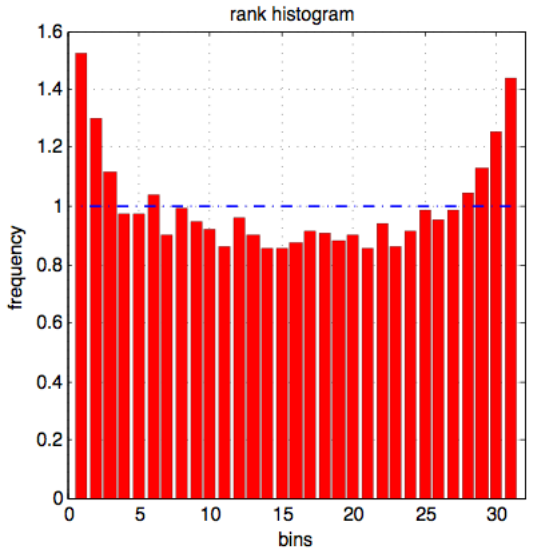
- Comparison *Ensemble Kalman Filter* (*EnKF*) and *Particle Filters* (*PF*)

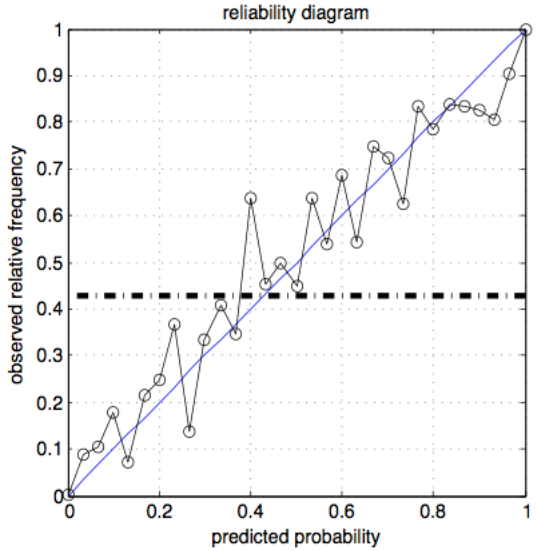
  Both of these algorithms being sequential, comparison is fair only at end of assimilation window

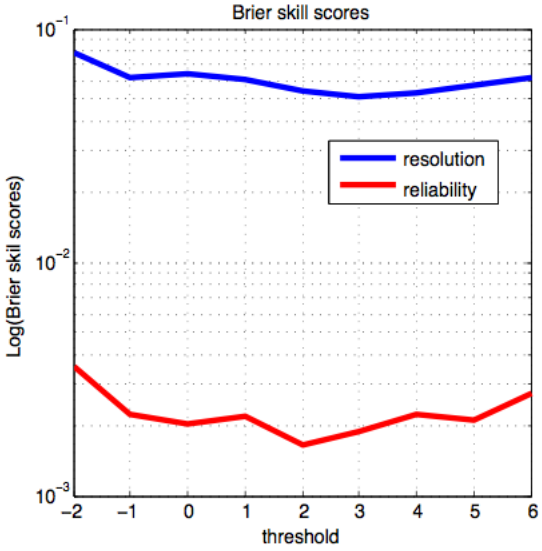


Nonlinear Lorenz'96. 5 days. Diagnostics at end of assimilation window

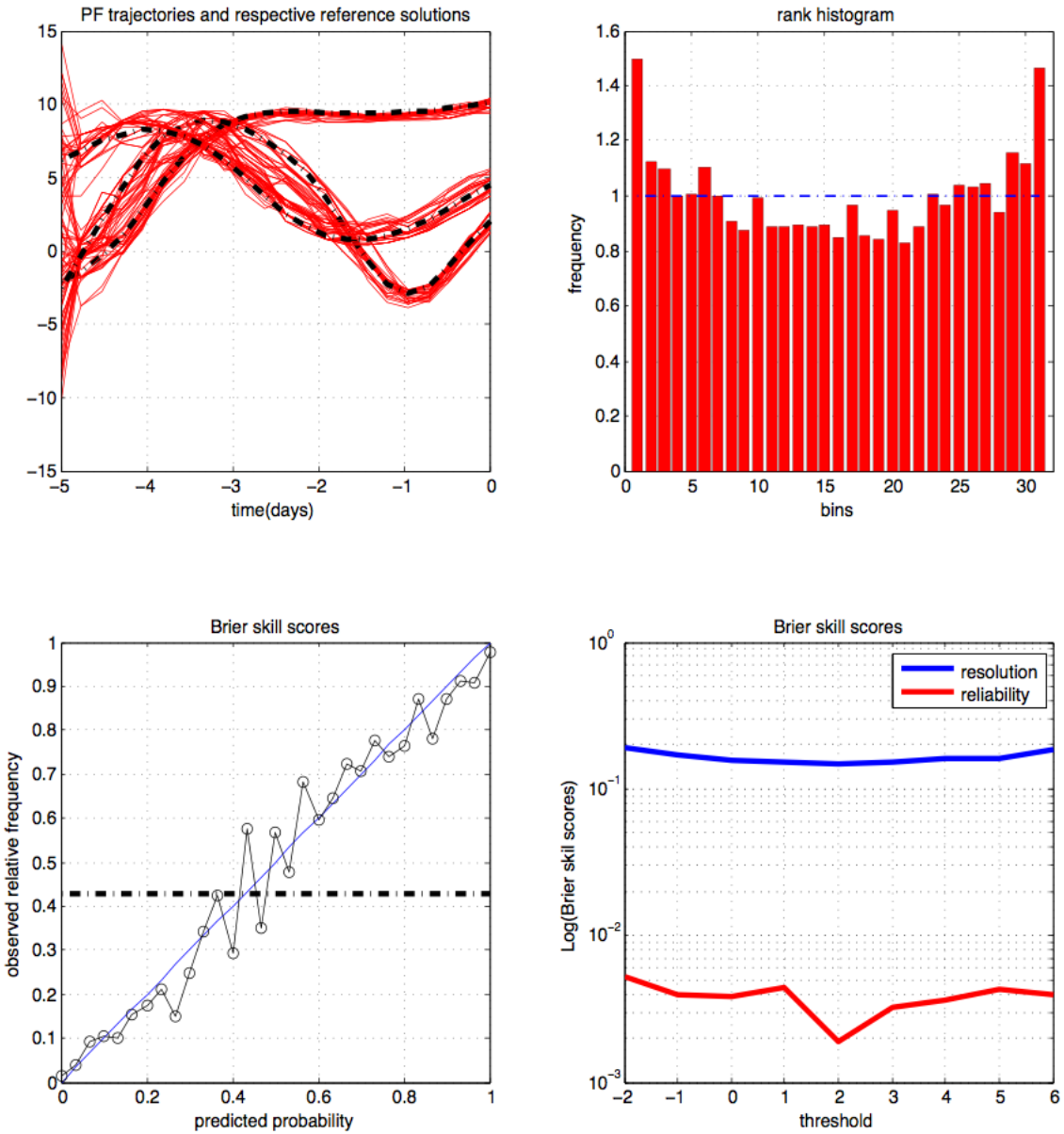




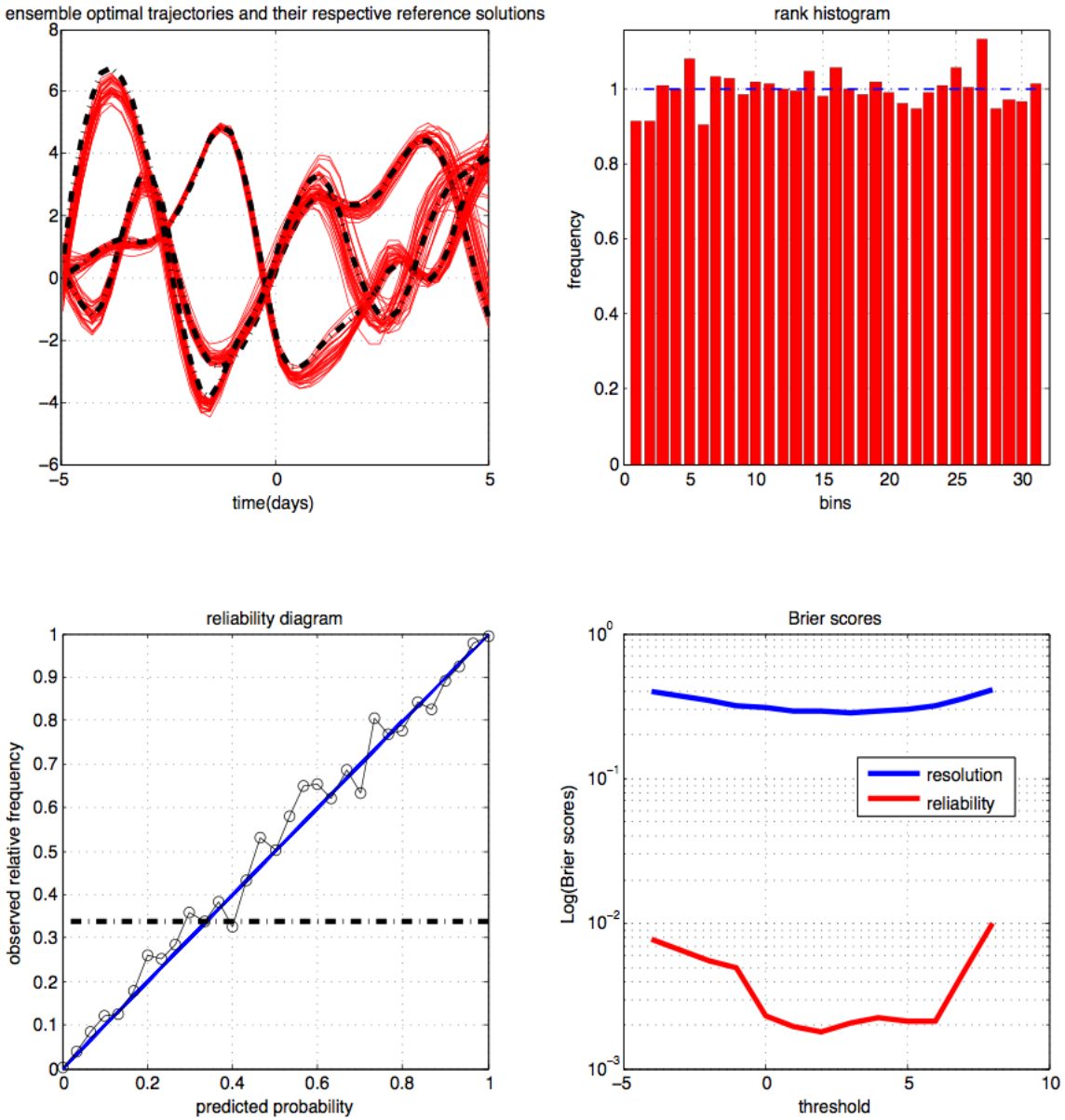




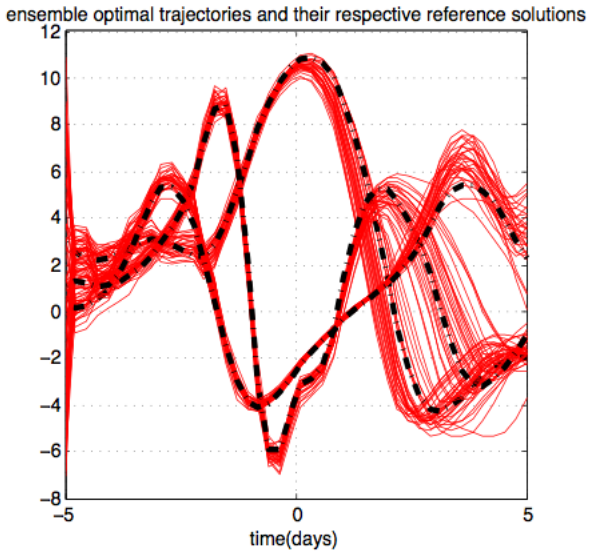
Nonlinear Lorenz'96. EnKF. Diagnostics after 5 days of assimilation

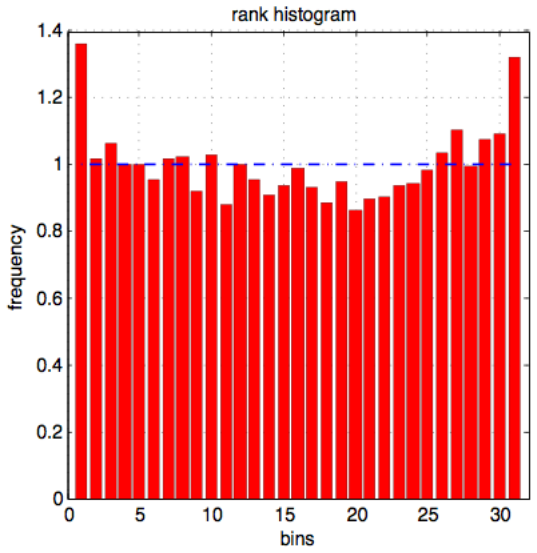


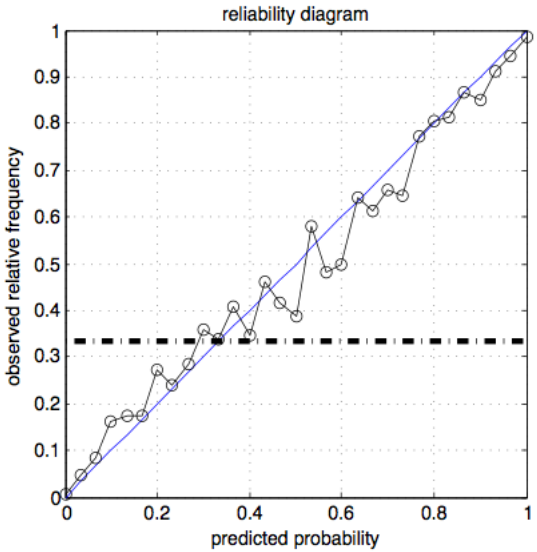
Nonlinear Lorenz'96. PF. Diagnostics after 5 days of assimilation

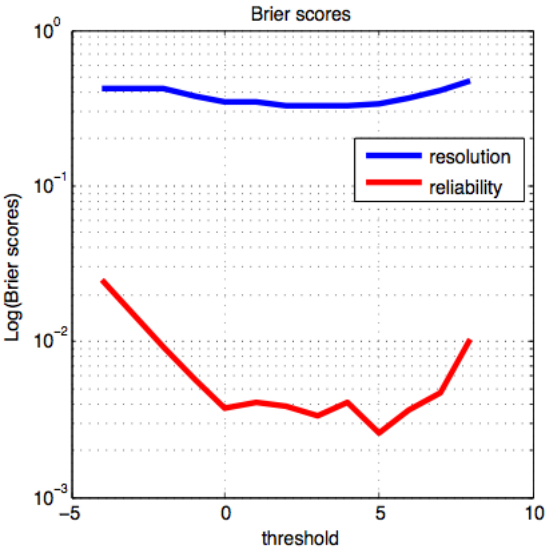


EnsVAR. Diagnostics for 5-day forecasts

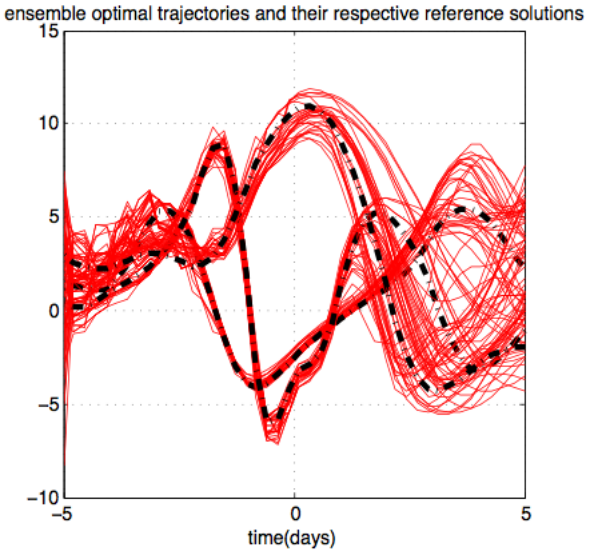


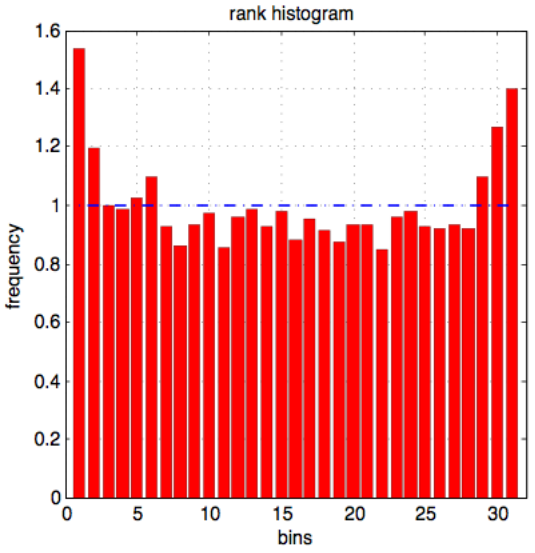


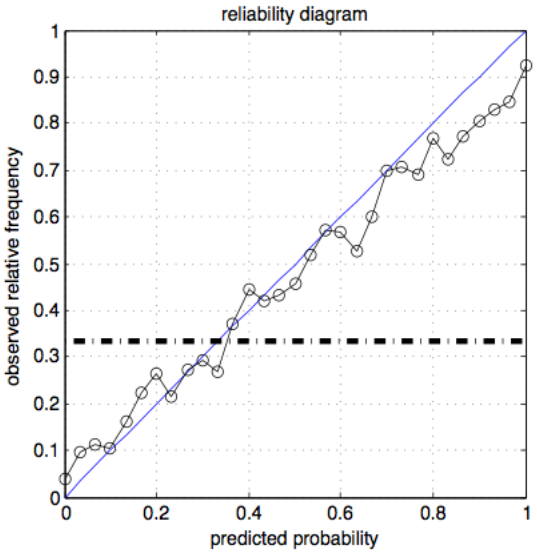


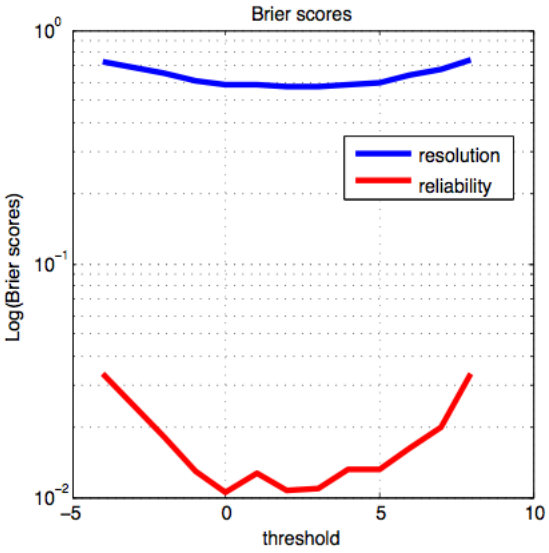


EnKF. Diagnostics for 5-day forecasts









PF. Diagnostics for 5-day forecasts

DA procedure / *method*	*Assimilation*	Forecasting
EnsVAR	0.2193510	1.49403506
EnKF	0.2449690	1.67176110
PF	0.7579790	2.62461295

RMS errors at the end of 5-day assimilations and 5-day forecasts

Weak constraint EnsVar

- define the objective function.

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

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

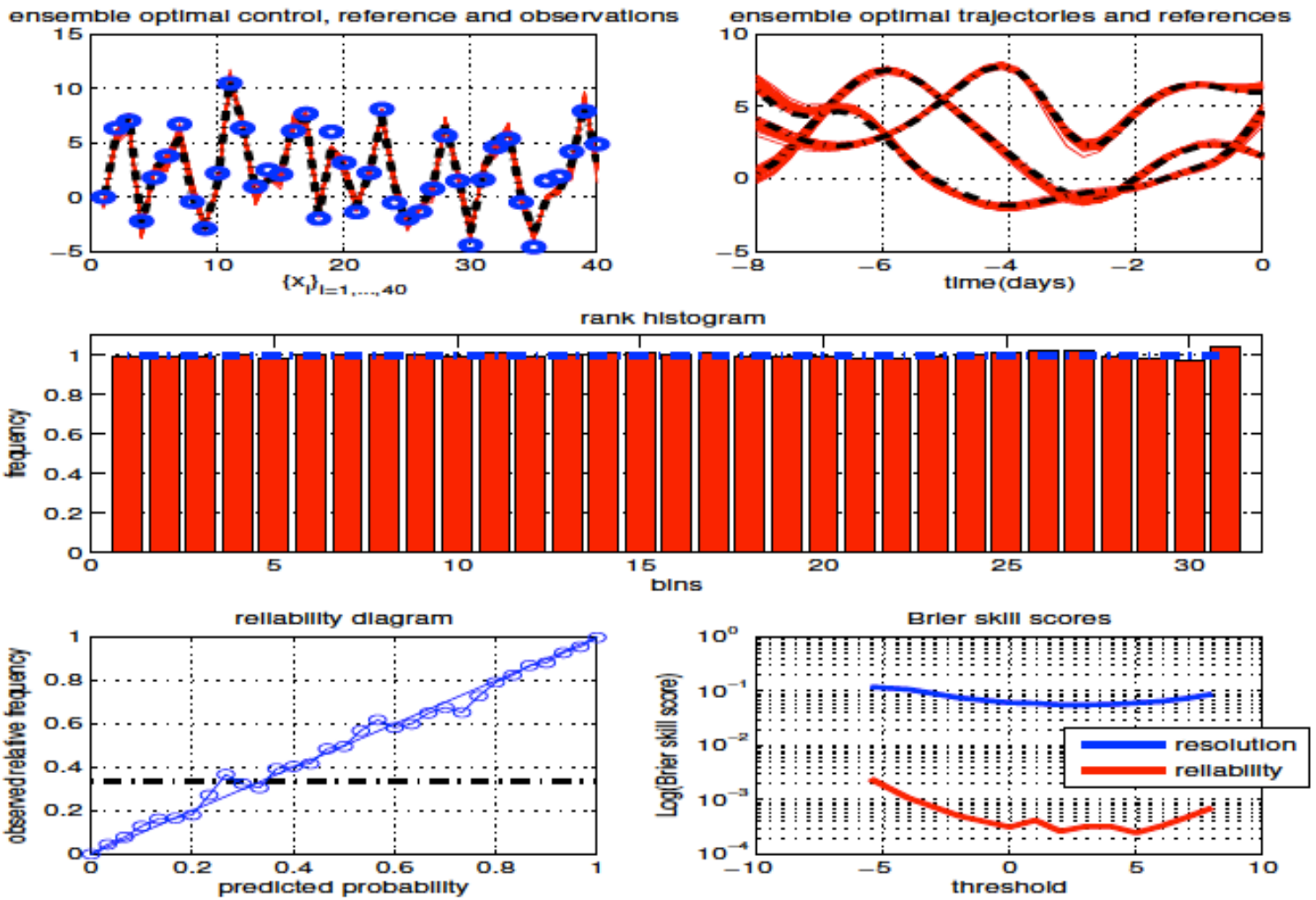
1. B background error covariance matrix and R observation error covariance matrix.
2. Q model error covariance matrix.
3. $H : \mathbb{R}^{state} \longrightarrow \mathbb{R}^{obs}$ observation operator.
4. x_b background state vector and y_i observation vector at time $t = t_i$.
5. η_i model error vector 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}.

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

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

Weak EnsVar : the Lorenz96 model 8 days



Summary

Summary

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

Ensembles obtained are Gaussian, even if errors in data are not

Produces Monte-Carlo sample of (probably not) bayesian pdf

EnsVar : Pros and cons

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).
- Empirical.
- Cycling of the process (work in progress).

La suite ?

- Mettre en œuvre sur modèle physiquement plus réaliste (QG, Saint-Venant, …) et/ou contenant plus de nonlinéarités ?

- Comparaison avec d'autres algorithmes (IEnKS)

- Cyclage et/ou chevauchement

- Minimisation dans l'espace instable

- Améliorations algorithmiques

The End