Verification of Ensemble Systems

O. Talagrand[1] and G. Candille[1,2]

1. Laboratoire de Météorologie Dynamique, École Normale Supérieure, Paris, France
2. Université du Québec à Montréal, Montréal, Canada

Workshop *Diagnostics of data assimilation system performance*
ECMWF, Reading, England
16 June 2009

With thanks to F. Atger, R. Buizza, T. Palmer, and to participants in Interest Group 5 of THORPEX Working Group on *Predictability and Dynamical Processes*

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 a problem in bayesian estimation.

Determine the conditional probability distribution for the state of the system, knowing everything we know (unambiguously defined if a prior probability distribution is defined; see Tarantola, 2005).

Bayesian estimation is however impossible in its general theoretical form in meteorological or oceanographical practice because

- It is impossible to explicitly describe a probability distribution in a space with dimension even as low as $n \approx 10^3$, not to speak of the dimension $n \approx 10^{6\text{-}8}$ of present Numerical Weather Prediction models.

- Probability distribution of errors on data very poorly known (model errors in particular).

Purpose of ensemble assimilation (and prediction too)

- Produce an ensemble of estimates which are meant to sample the conditional probability distribution (dimension $N \approx O(10\text{-}100)$).

Ensemble Kalman Filter (*EnKF*) and its many variants.

Cannot be exactly bayesian for nonlinear systems (it has been rigorously shown by Le Gland *et al*. that, in the limit of infinite ensemble size, the probability distribution defined by EnKF tends to a limit which is different in the nonlinear case from the bayesian distribution)

Exact bayesian estimation

Particle filters

Predicted ensemble at time t : $\{x^b{}_n, n = 1, \ldots, N\}$, each element with its own weight (probability) $P(x^b{}_n)$

Observation vector at same time : $y = Hx + \varepsilon$

Bayes' formula

$$P(x^b{}_n \mid y) \sim P(y \mid x^b{}_n)\, P(x^b{}_n)$$

Defines updating of weights

Particle filters are bayesian in the limit of infinite number of particles (proven)

Particle filters are now a ‘hot’ research topic, studied in many places (C. Snyder, P. J. van Leeuwen, S. Nakano, C. Baehr, …)

A general problem is that weights tend to concentrate on one, or a small number of particles (⇒ need for regenerating new particles, for instance through *Sequential Importance Resampling*)

The main question is whether it is possible to implement useful (and stable) particle filters with ensemble dimensions that are not prohibitive

Another question, as always with sequential estimation, is the possibility of taking temporal error dependence into account.

Exact bayesian estimation

Acceptation-rejection

Bayes' formula

$$f(x) \equiv \mathrm{P}(x \mid y) = P(y \mid x)\, P(x) \,/\, P(y)$$

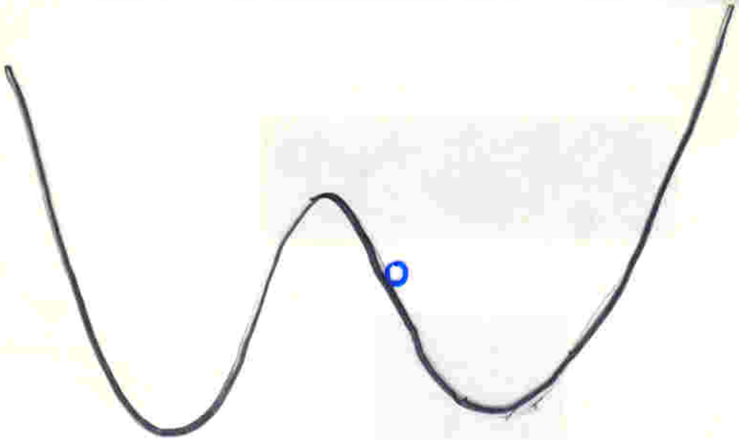
defines probability density function for x.

Construct sample of that pdf as follows.

Draw randomly couple $(\xi, \psi) \in S$ x $[0, \mathrm{sup}f]$.

Keep ξ if $\psi < f(\xi)$. ξ is then distributed according to $f(x)$.

Miller, Carter and Blue, Tellus, 1999



$$\frac{d^2x}{dt^2} = -\frac{d\Phi}{dx} - \alpha\frac{dx}{dt} + \text{Noise}$$

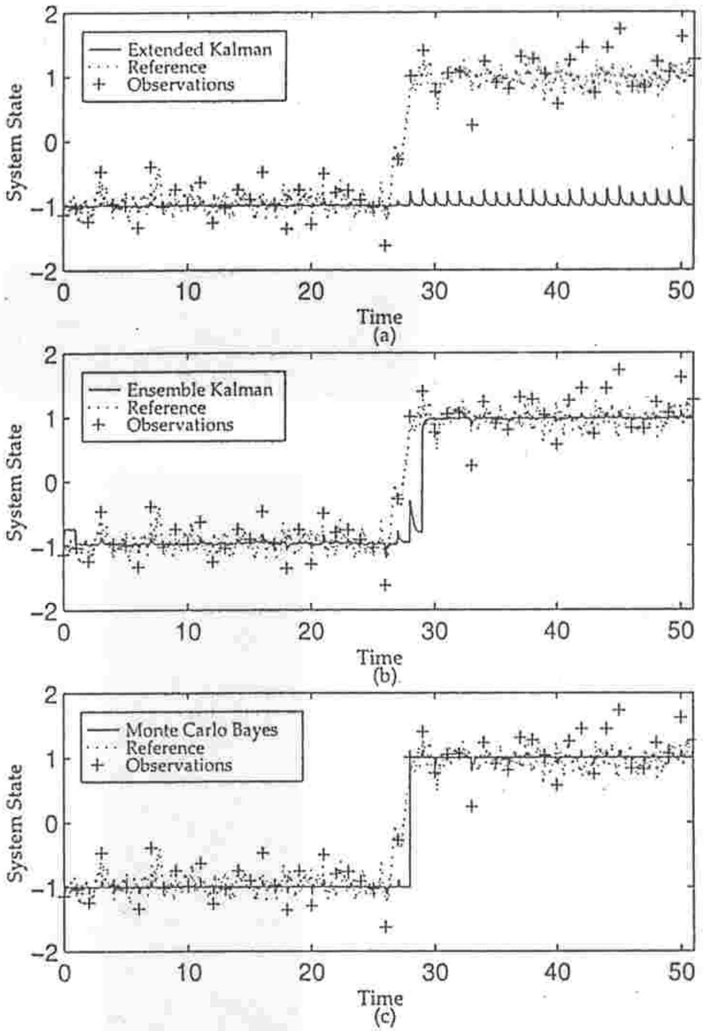


Fig. 4. Comparison of the EKF, the ensemble method and nonlinear filtering by Bayes' theorem for the double-well problem.

Miller, Carter and Blue, 1999, *Tellus*, **51A**, 167-194

Acceptation-rejection

Seems costly.

Requires capability of permanently interpolating probability distribution defined by finite sample to whole state space.

Question

- Accepting that the purpose of ensemble assimilation is to obtain a sample of the underlying conditional probability distribution for the state of the flow, how can one objectively (and quantitatively) evaluate the degree to which that purpose has been achieved ?

Remark. The same question arises for ensemble prediction

My point of view

- Ensemble estimation (either assimilation or prediction) is of a different essence than deterministic estimation in that the object to be estimated (basically a probability or a probability distribution) is not better known *a posteriori* than it was *a priori* (in fact, that object has no objective existence and cannot be possibly observed at all)

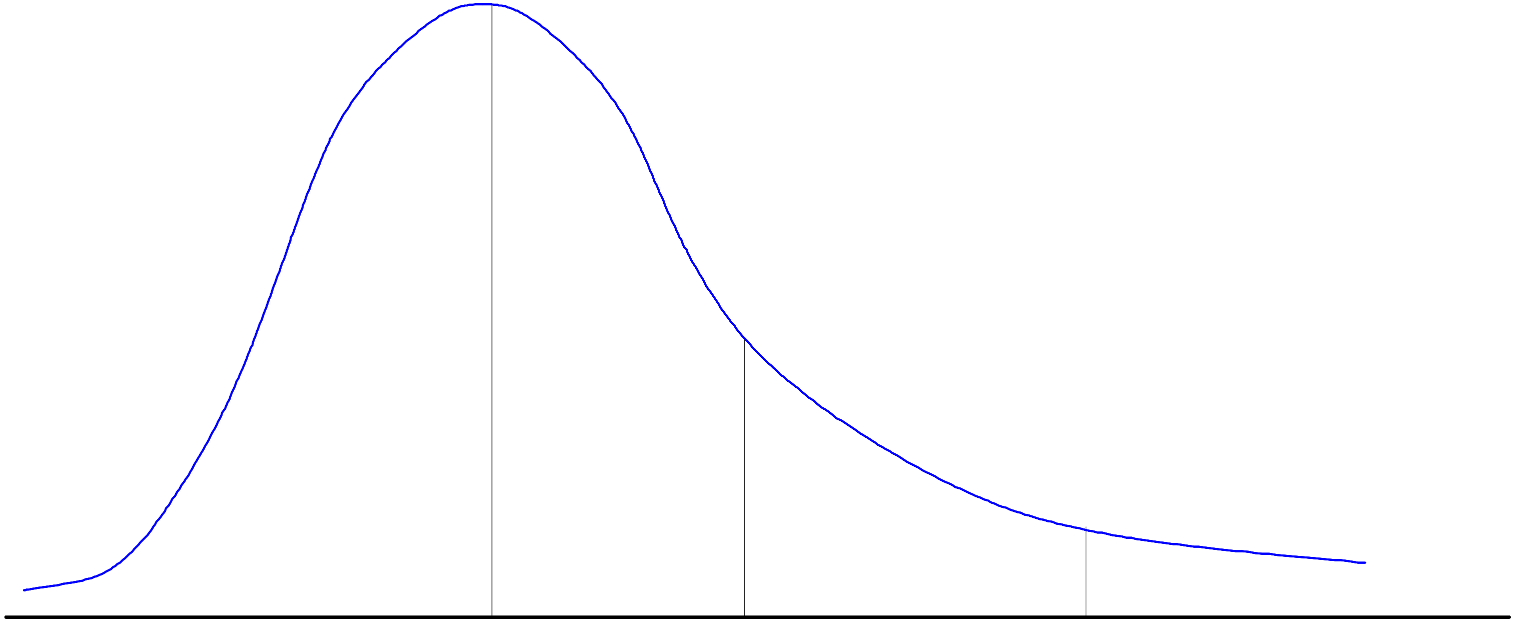
- As a consequence, validation of ensemble estimation can only be statistical, and it is meaningless (except in limit cases, as when the estimated probability distribution has a very narrow spread, and the verifying observation falls within the predicted spread, or on the contrary when the verifying observation falls well outside the spread of the estimated probability distribution) to speak of the quality of ensemble estimations on a case-to-case basis

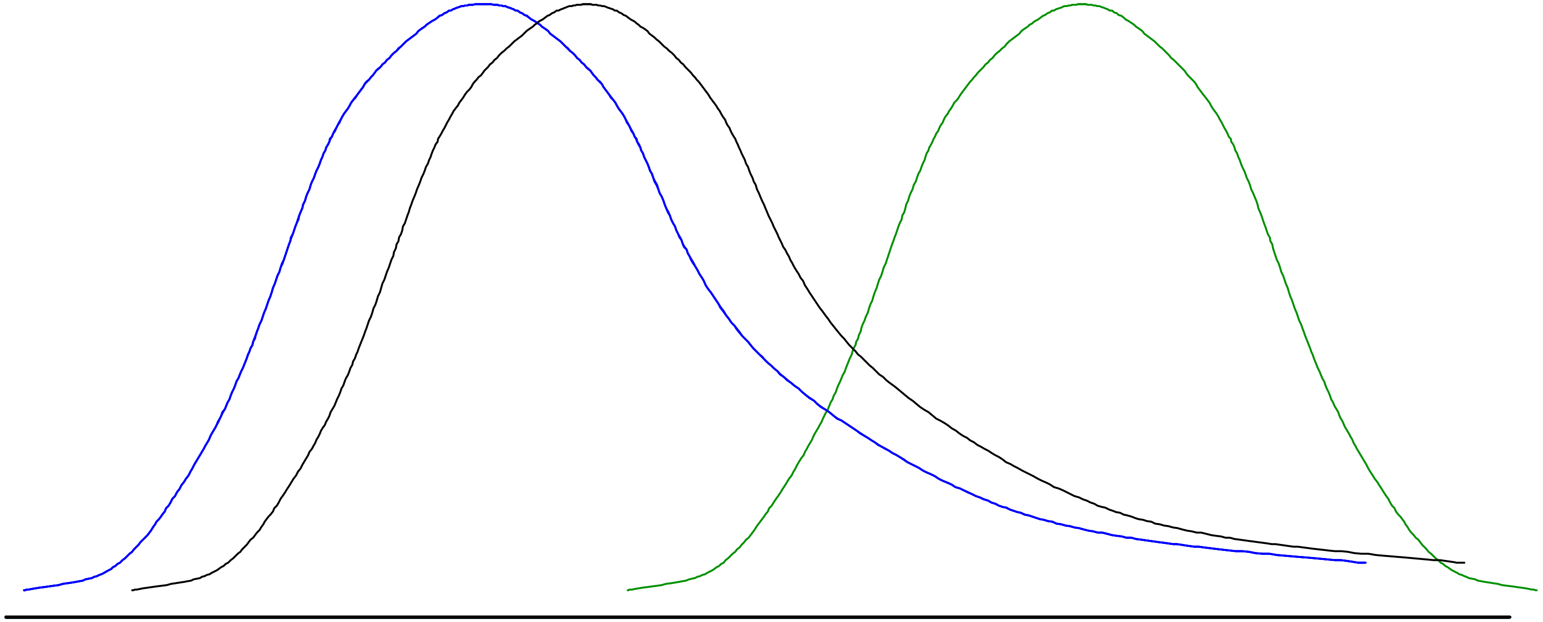
What are the attributes which make a good ensemble estimation system ?

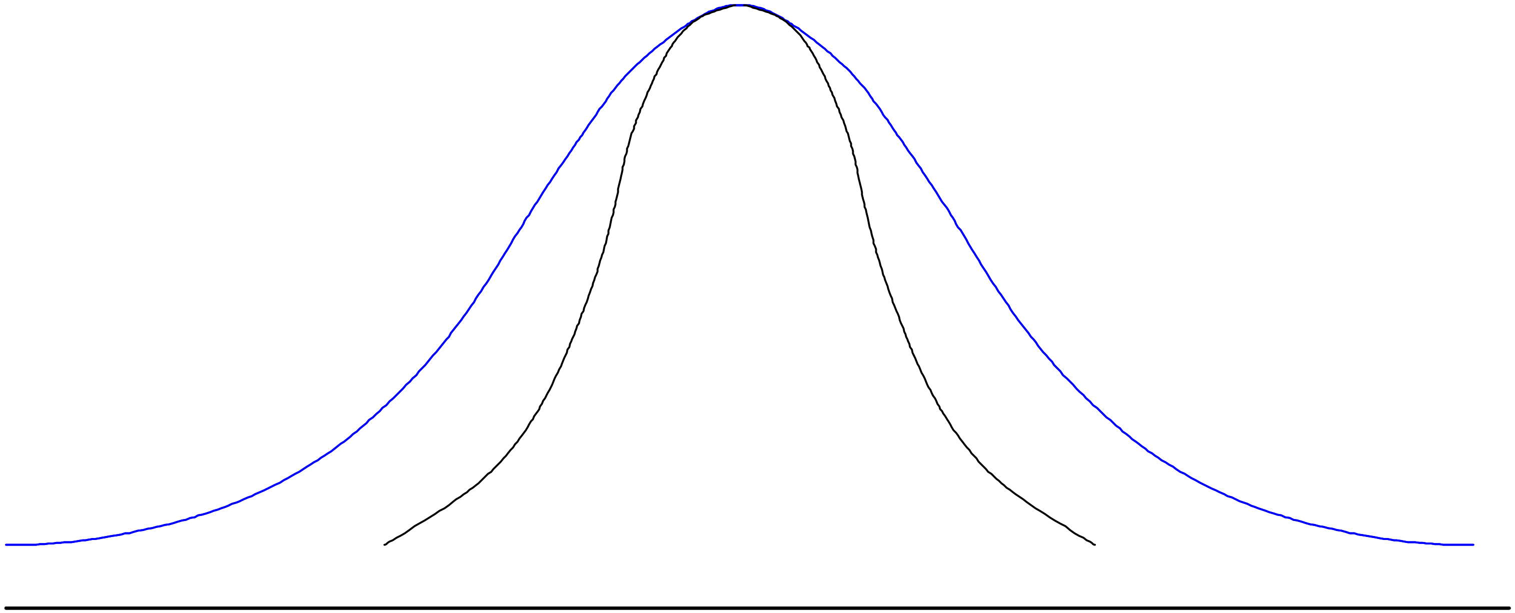
o ***Reliability***

(*it rains 40% of the times I predict 40% probability for rain*)

- Statistical agreement between estimated probability and observed frequency for all events and all probabilities







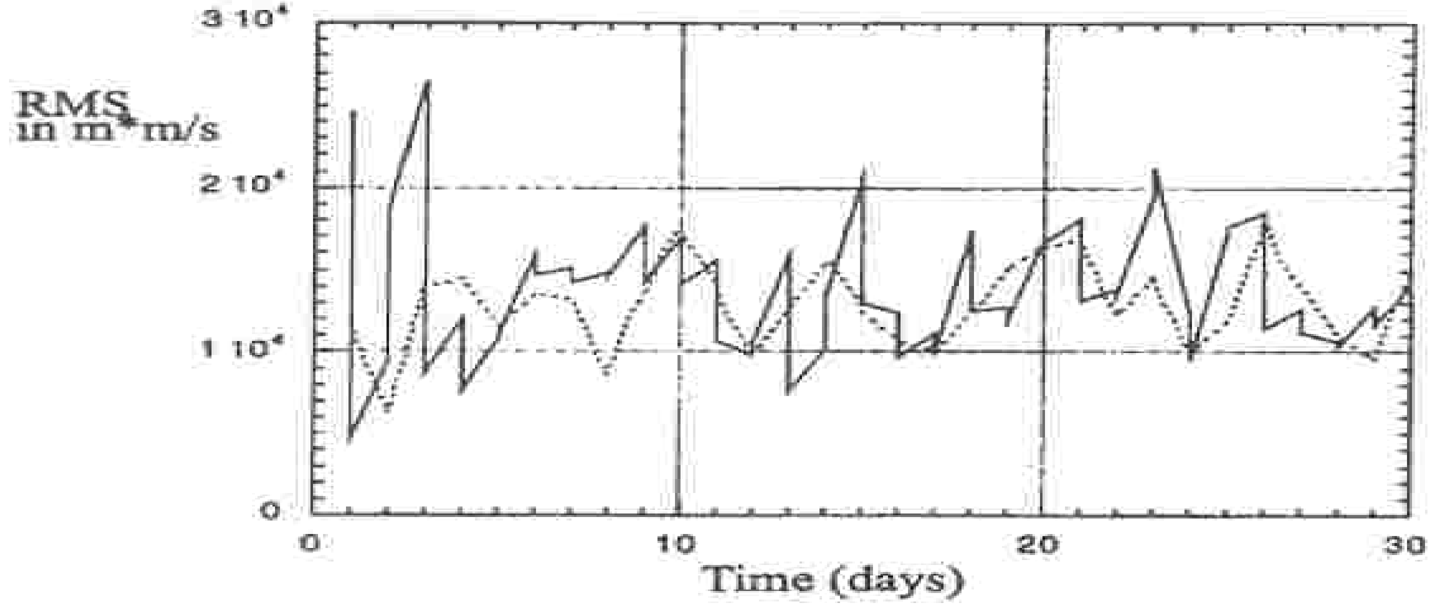


FIG. 12. Comparison of rms error ($m^2\ s^{-1}$) between ensemble mean and independent observations (dotted line) and the std dev in the ensemble (solid line). The excellent agreement shows that the SIRF is working correctly.

van Leeuwen, 2003, *Mon. Wea. Rev.*, **131**, 2071-2084

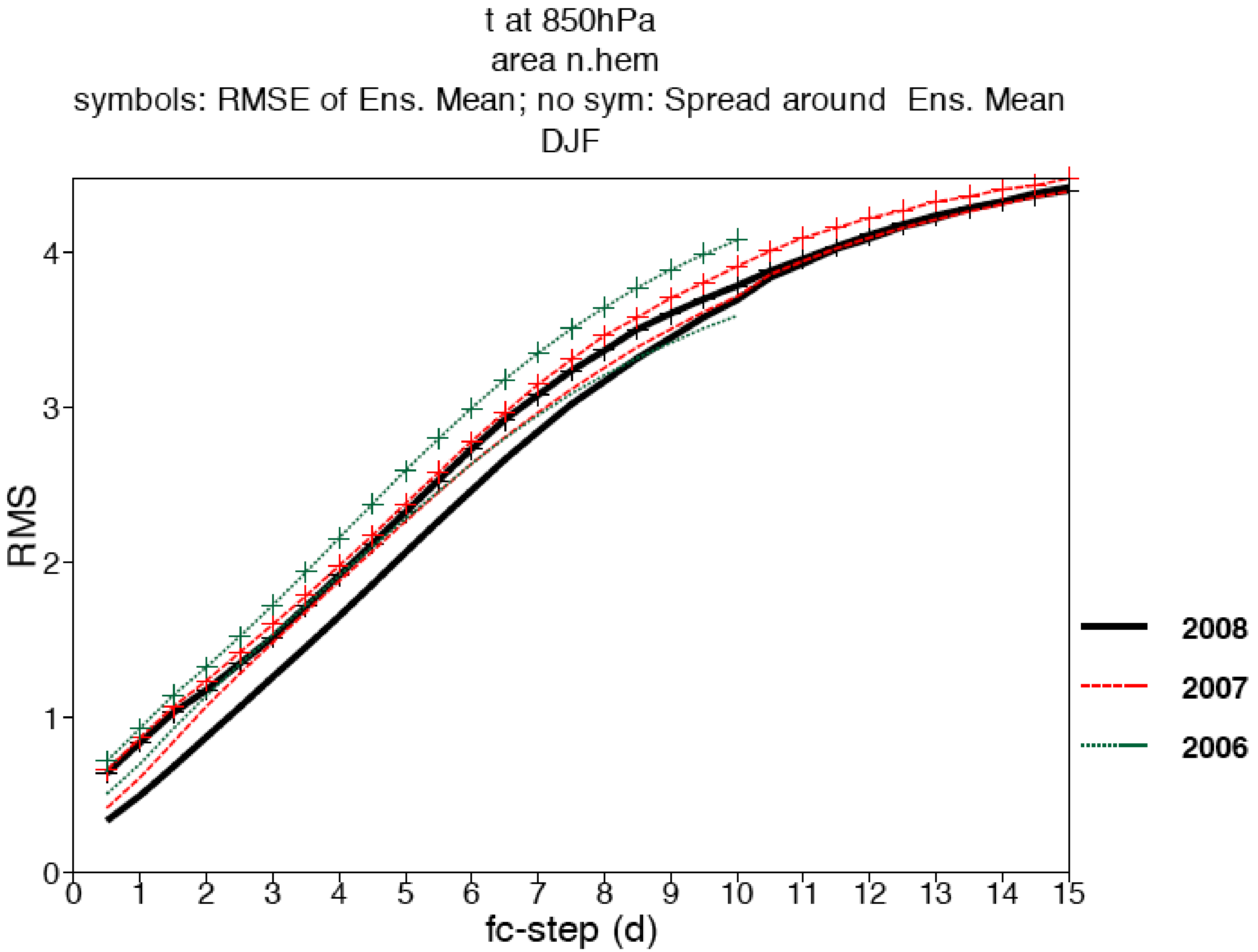


Figure 8: Ensemble spread (standard deviation) and root mean square error of ensemble-mean (lines with crosses) for 500 hPa height (top) and 850 hPa temperature (bottom) for winter 2007-08 (black), 2006-07 (red) and 2005-06 (green) over the extra-tropical northern hemisphere.

Richardson *et al*., 2008, ECMWF Technical Memorandum 578

More generally

- Consider a probability distribution F. Let $F‘(F)$ be the conditional frequency distribution of the observed reality, given that F has been predicted. Reliability is the condition that

$$F‘(F) = F \qquad\qquad \text{for any } F$$

Measured by reliability component of Brier and Brier-like scores, rank histograms, Reduced Centred Variable, …

More generally, for a given scalar variable, *Reduced Centred Random Variable* (RCRV, Candille *et al*., 2006)

$$s = \frac{\xi - \mu}{\sigma}$$

where ξ is verifying observation, and μ and σ are respectively the expectation and the standard deviation of the predicted probability distribution.

Over a large number of realizations of a reliable probabilistic prediction system

$$E(s) = 0 \qquad , \qquad E(s^2) = 1$$

If observations show that $F'(F) \neq F$ for some F, then *a posteriori* calibration

$$F \Rightarrow F'(F)$$

renders system reliable. Lack of reliability, under the hypothesis of stationarity of statistics, can be corrected to the same degree it can be diagnosed.

Second attribute

o ‘***Resolution***’ (also called ‘***sharpness***’)

Reliably predicted probabilities $F'(F)$ are distinctly different from climatology

Measured by resolution component of Brier and Brier-like scores, ROC curve area, information content, …

It is the conjunction of reliability and resolution that makes the value of a probabilistic estimation system. Provided a large enough validation sample is available, each of these qualities can be objectively and quantitatively measured by a number of different, not exactly equivalent, scores.

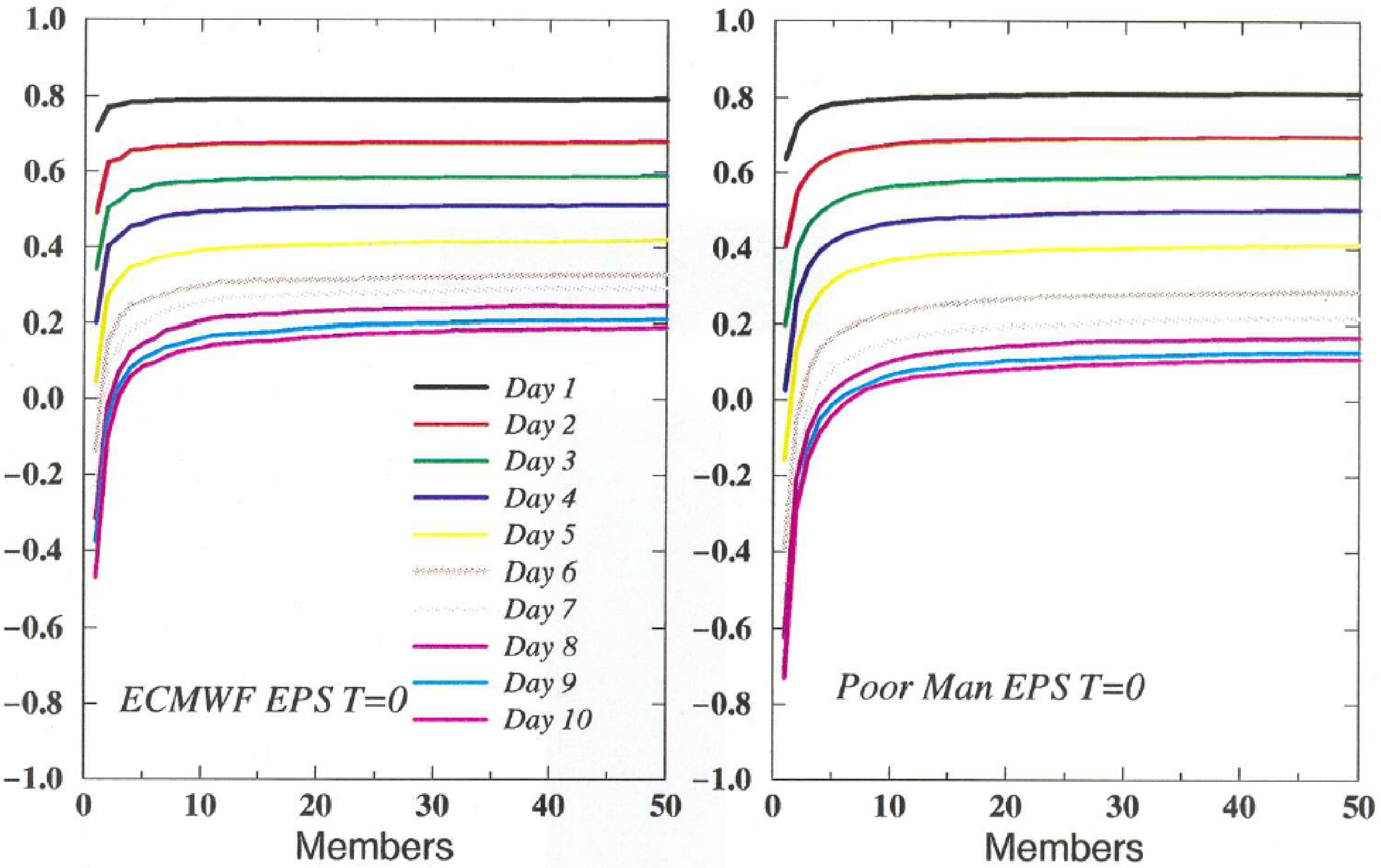
In the specific context of ensemble assimilation, checking reliability is checking statistical consistency between *a priori* estimated and *a posteriori* observed innovation (as done by Desroziers for 'deterministic' assimilation).

Checking resolution is checking magnitude of innovation.

Size of Assimilation Ensembles ?

Two aspects at least can have an impact on the size of assimilation ensembles : the numerical stability of the assimilation process, and the quality of the results.

o Observed fact : in ensemble prediction, present scores saturate for value of ensemble size N in the range 30-50, independently of quality of score.



Impact of ensemble size on Brier Skill Score
ECMWF, event $T_{850} > T_c$ Northern Hemisphere

(Talagrand *et al*., ECMWF, 1999)

Theoretical estimate (raw Brier score)

$$B_N = B_\infty + \frac{1}{N}\int_0^1 p(1-p)g(p)dp$$

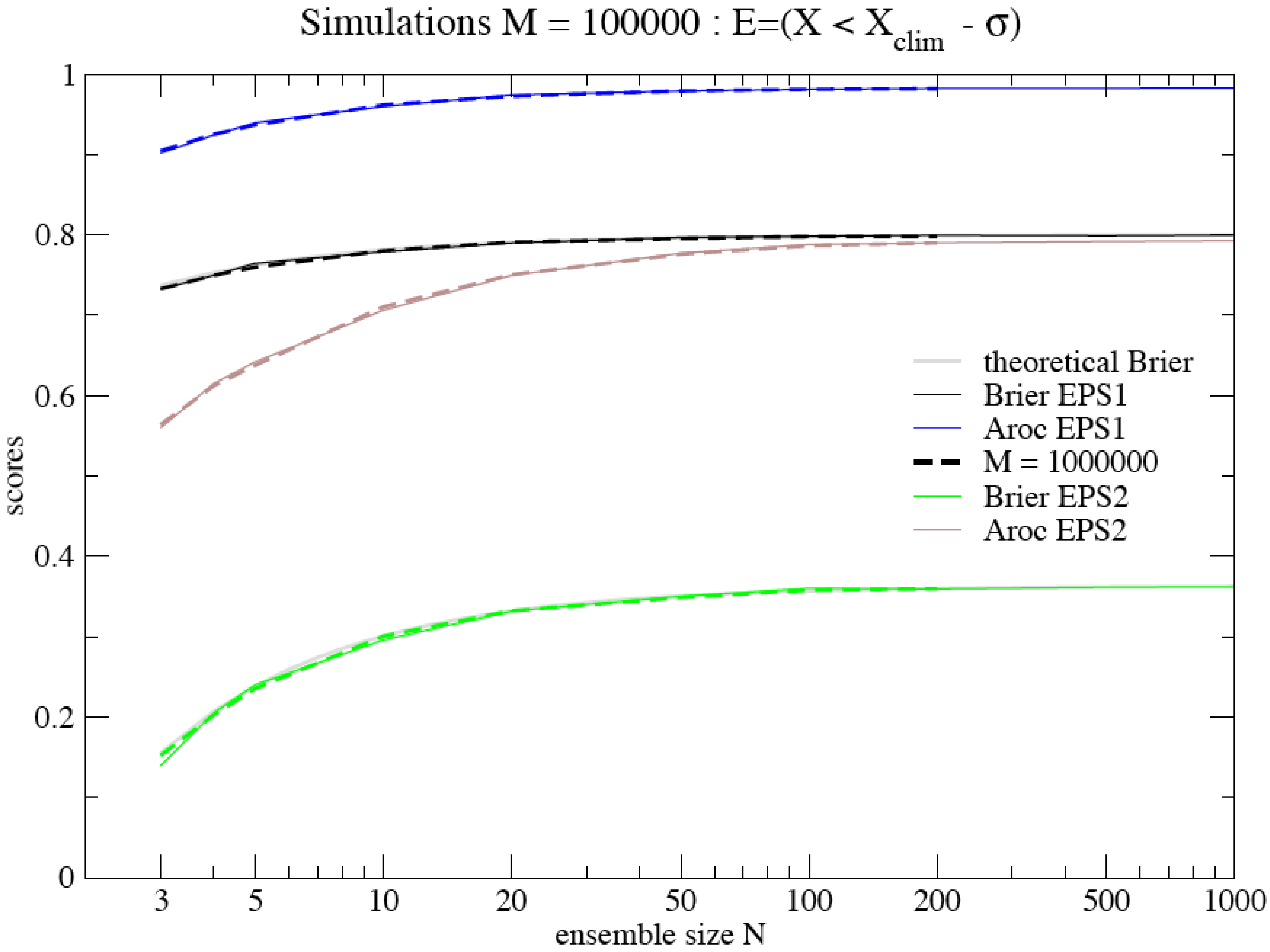


Figure 1: Impact of N on Brier Skill Score and ROC area

G. Candille, 2009

Question

Why do scores saturate for $N \approx$ 30-50 ? Explanations that have been suggested

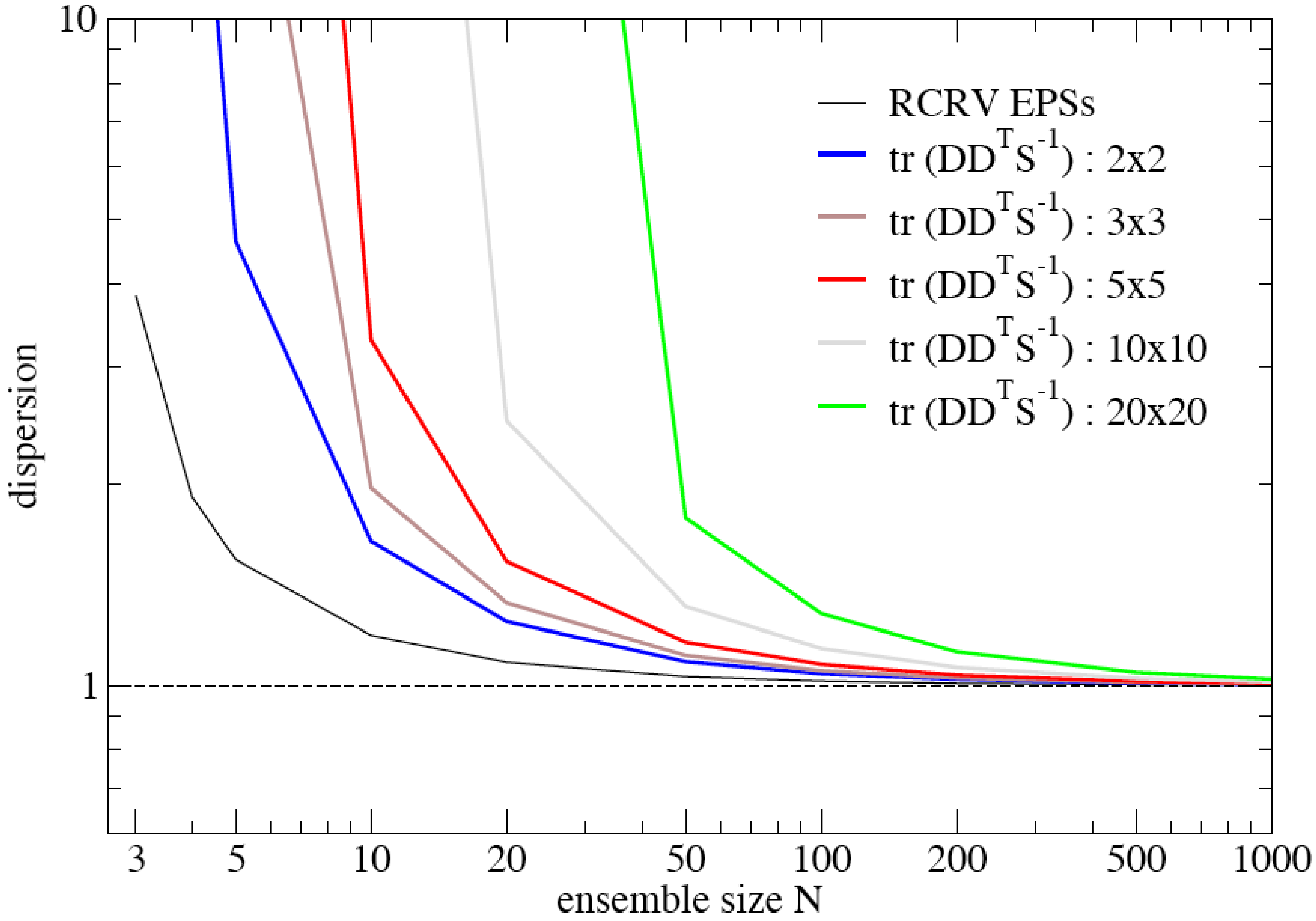
(i) Saturation is determined by the number of unstable modes in the system. Situation might be different with mesoscale ensemble prediction.

(ii) Validation sample is simply not large enough.

(iii) Scores have been implemented so far on probabilisic predictions of events or one-dimensional variables (*e*. *g*., temperature at a given point). Situation might be different for multivariate probability distributions (but then, problem with size of verification sample).

(iv) Probability distributions (in the case of one-dimensional variables) are most often unimodal. Situation might be different for multimodal probability distributions (as produced for instance by multi-model ensembles).

In any case, problem of size of verifying sample will remain, even if it can be mitigated to some extent by using reanalyses or reforecasts for validation.



G. Candille, 2008

Is it possible to objectively validate multi-dimensional probabilistic predictions ?

Consider the case of prediction of 500-hPa winter geopotential over the Northern Atlantic Ocean, (10-80W, 20-70N) over a 5x5-degree^2 grid ⇒165 gridpoints.

In order to validate probabilistic prediction, it is in principle necessary to partition predicted probability distributions into classes, and to check reliability for each class.

Assume $N = 5$, and partitioning is done for each gridpoint on the basis of $L = 2$ thresholds. Number of ways of positioning N values with respect to L thresholds. Binomial coefficient

$$\binom{N+L}{L}$$

This is equal to 21 for $N = 5$ and $L = 2$, which leads to

$$21^{165} \approx 10^{218}$$

possible probability distributions.

Is it possible to objectively validate multi-dimensional probabilistic predictions (continuation) ?

$21^{165} \approx 10^{218}$ possible probability distributions.

To be put in balance with number of available realizations of the prediction system. Let us assume 150 realizations can be obtained every winter. After 3 years (by which time system will have started evolving), this gives the ridiculously small number of 450 realizations.

Is it possible to objectively validate multi-dimensional probabilistic predictions (continuation) ?

For a more moderate example, consider long-range *(e. g*., monthly or seasonal) probabilistic prediction of weather regimes (still for the winter Northern Atlantic). Vautard (1990) has identified four different weather regimes, with lifetimes of between one and two weeks. The probabilistic prediction is then for a four-outcome event. With $N = 5$-sized ensembles, this gives 56 possible distributions of probabilities.

In view of the lifetimes of the regimes, there is no point in making more than one forecast per week. That would make 60 forecasts over a 3-year period. Hardly sufficient for accurate validation.

Conclusions

Bayesian (low-order) ensemble assimilation does not exist at present. May exist in some not-too-distant future. Size of ensembles remains a problem.

Ensemble assimilation must be evaluated, not only in terms of 'deterministic' accuracy (*e. g*., accuracy of mean of ensemble), but also in terms of how well it estimates the spread of uncertainty. That is objectively measured by *reliability* and *resolution (sharpness)*.

But there are limits to what can be achieved in that respect. It is necessary to identify clearly those limits.