On Some Aspects of Validation of Probabilistic Prediction

O. Talagrand, G. Candille and L. Descamps
Laboratoire de Météorologie Dynamique, École Normale Supérieure
Paris, France

Workshop *Ensemble Prediction*
ECMWF, Reading, England
7 November 2007

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*

Questions

- What can one expect from ensemble prediction ? Which goal(s) can be assigned to ensemble prediction ?

- How can one objectively (and quantitatively) evaluate the degree to which the goal(s) assigned to ensemble prediction has (have) been achieved ?

What can one expect from ensemble predictions ?

- Increase confidence in prediction of high impact weather ?

- Put bounds on future state of the flow ?

- Predict ‘scenarii’ ?

- Produce more accurate (deterministic) forecasts, for instance by taking the average of the ensemble ?

All those possible goals are actually included in the broader goal of predicting probabilities of occurrence (for events), or more generally probability distributions (for variables such as temperature or rainfall, or even for whole meteorological fields).

How can one objectively evaluate ensemble prediction ?

Point has been strongly made (in THORPEX discussion group) that

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

- As a consequence, validation of ensemble prediction can only be statistical, and it is meaningless (except in limit cases, as when the predicted 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 predicted probability distribution) to speak of the quality of ensemble predictions on a case-to-case basis

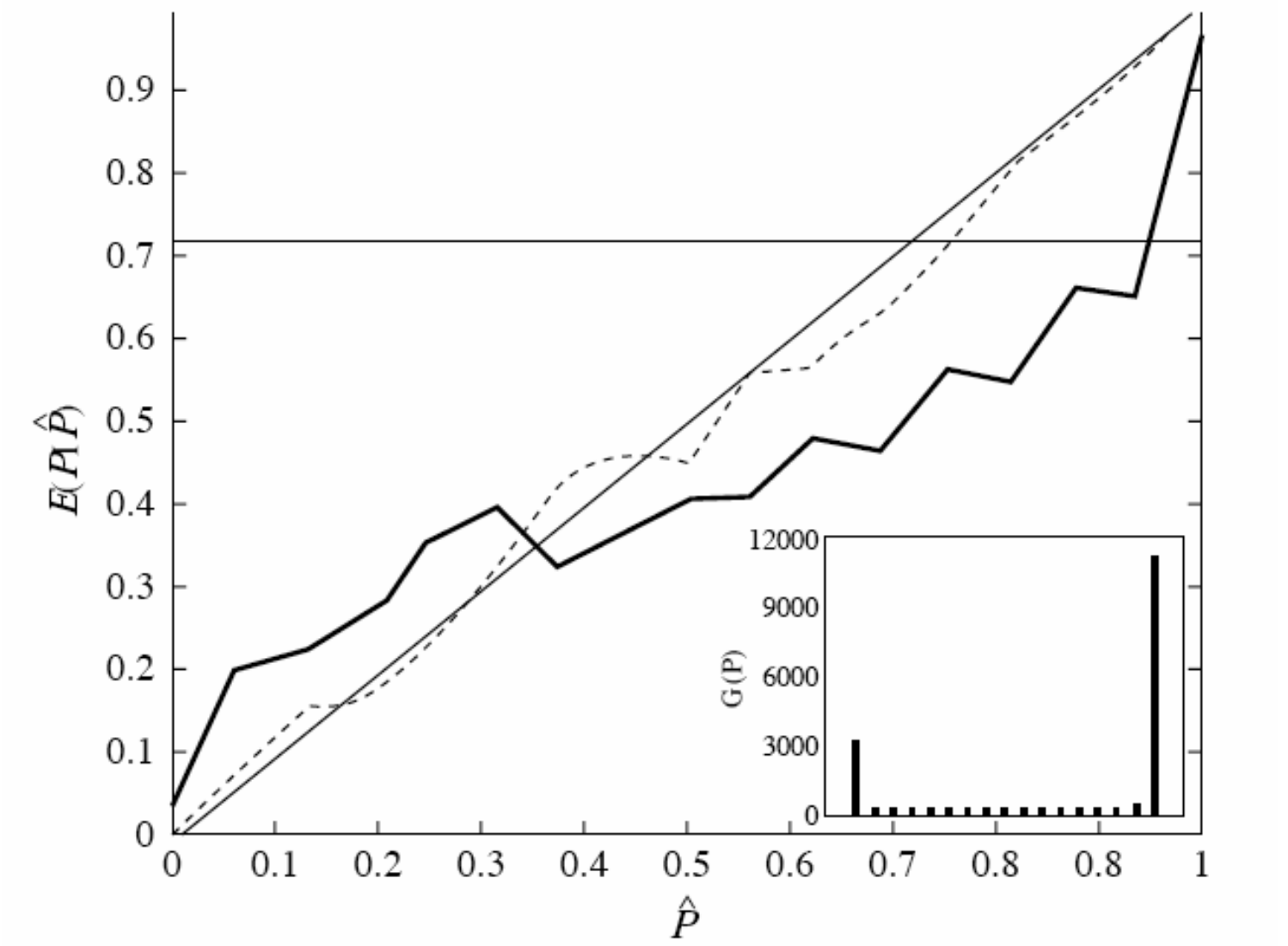
That point of view has not however gained complete agreement, and discussion has been going on how to evaluate the quality of individual ensemble predictions.

What are the attributes which make a good Ensemble Prediction System ?

- ***Reliability***

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

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



Reliability diagramme, NCEP, event $T_{850} > T_c$ - 4C, 2-day range, Northern Atlantic Ocean, December 1998 - February 1999

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

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

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

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

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

- '***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 prediction system. Provided a large enough valaidation sample is available, each of these qualities can be objectively and quantitatively measured by a number of different, not exactly equivalent, scores.

Three causes of ‘noise’ in diagnostics

- Finiteness of ensembles
- Finiteness of validation sample
- Noise on validating observations

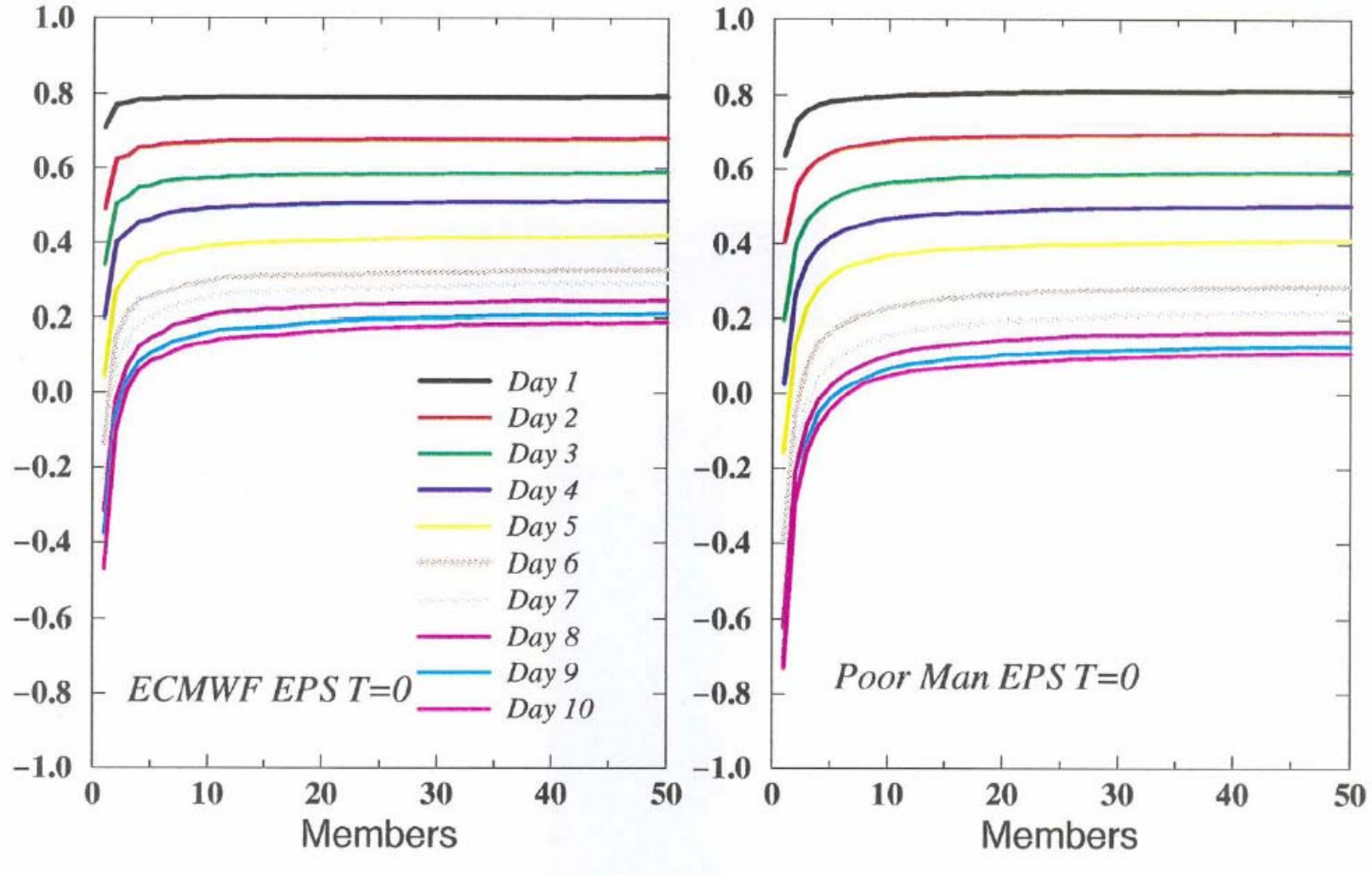
(impact of all three studied by Candille, 2003)

Size of Prediction Ensembles ?

Given the choice, is it better to improve the quality of the forecast model, or to increase the size of the predicted ensembles ?

Actually, the really significant parameter is not the size of the ensembles, but the numerical resolution with which probabilities are forecast.

- Observed fact : 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)

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

Size of Prediction Ensembles (continuation 1) ?

This observed fact raises two questions

- Why do the scores saturate so rapidly ?

- Is it worth increasing N beyond values 30-50 ?

- If we take, say, $N = 200$, which user will ever care whether the probability for rain for to-morrow is 123/200 rather 124/200 ?

- What is the size of the verifying sample that is necessary for checking the reliability of a probability forecast of, say, $1/N$ for a given event E?

Answer. Assume one 10-day forecast every day, so that 10 forecasts are available for any given day. E must have occurred at least $\alpha N/10$ times, where α is of the order of a few units, before reliability can be reliably assessed.

If event occurs ~ 4 times a year, you must wait 10 years for $N = 100$, and 50 years for $N = 500$ ($\alpha = 4$).

This leads to question. Is reliable large-N probabilistic prediction of (even moderately) rare events possible at all ? Use ‘reforecasts’ (see T. Hamill’s presentation) ?

Size of Prediction Ensembles (continuation 2) ?

Theoretical fact: According to Chi-square statistics, with $N=30$ and a true variance of 1, the sample variance has a 95% chance of lying between 0.56 and 1.57; i.e. variance estimates are very inaccurate. With $N=100$, the corresponding 95% confidence interval (0.74,1.29) is significantly smaller.

Conclusion. If we want to accurately predict variances, large values of N are necessary.

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.

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 $\Rightarrow$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.

Size of Prediction Ensembles ? (conclusion)

More work is necessary to identify useful size of prediction ensembles, and practically possible size for verification sample.

Conclusions on this part

Reliability and *resolution (sharpness)* are the attributes that make the quality of a probabilistic prediction system. These are routinely measured in weather forecasting by a number of scores, each of which has its own particular significance. Other scores may be useful.

Strong limitations exist as to what can be achieved in practice by ensemble weather prediction. It is not clear whether there can be any gain in using ensemble sizes beyond $N \approx$ 30-50. And, even if there is, the unavoidably (relatively) small size of the verifying sample will often make it impossible to objectively evaluate the gain.

Much work remains to be done as to the optimal use of available resources for probabilistic weather prediction.

Definition of initial ensembles

Three basic approaches

- Singular modes (ECMWF)

  Singular modes are perturbations that amplify most rapidly in the tangent linear approximation over a given period of time. ECMWF uses a combination of ‘evolved’ singular vectors defined over the last 48 hours period before forecast, and of ‘future’ singular vectors determined over the first 48 hours of the forecast period. Mixture of past and future.

- ‘Bred’ modes (NCEP)

  Bred modes are modes that result from integrations performed in parallel with the assimilation process. Come entirely from the past.

- ‘Perturbed observation’ method (MSC)

  A form of ensemble assimilation. Comes entirely from the past.

L. Descamps (LMD)

Systematic comparison of different approaches, on simulated data, in as clean conditions as possible.

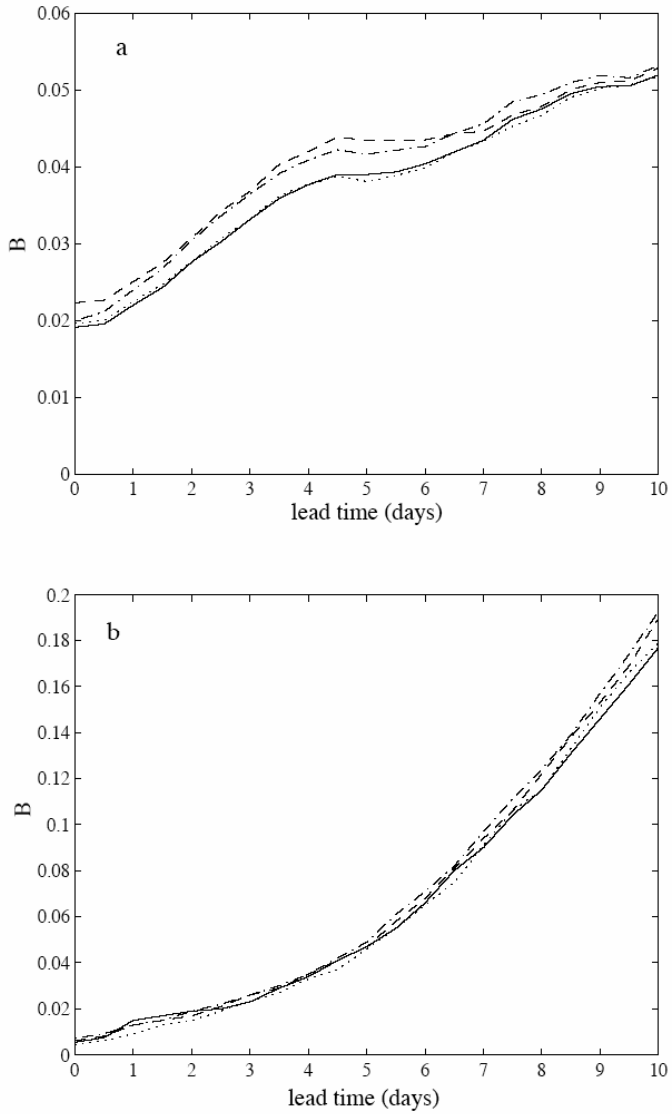
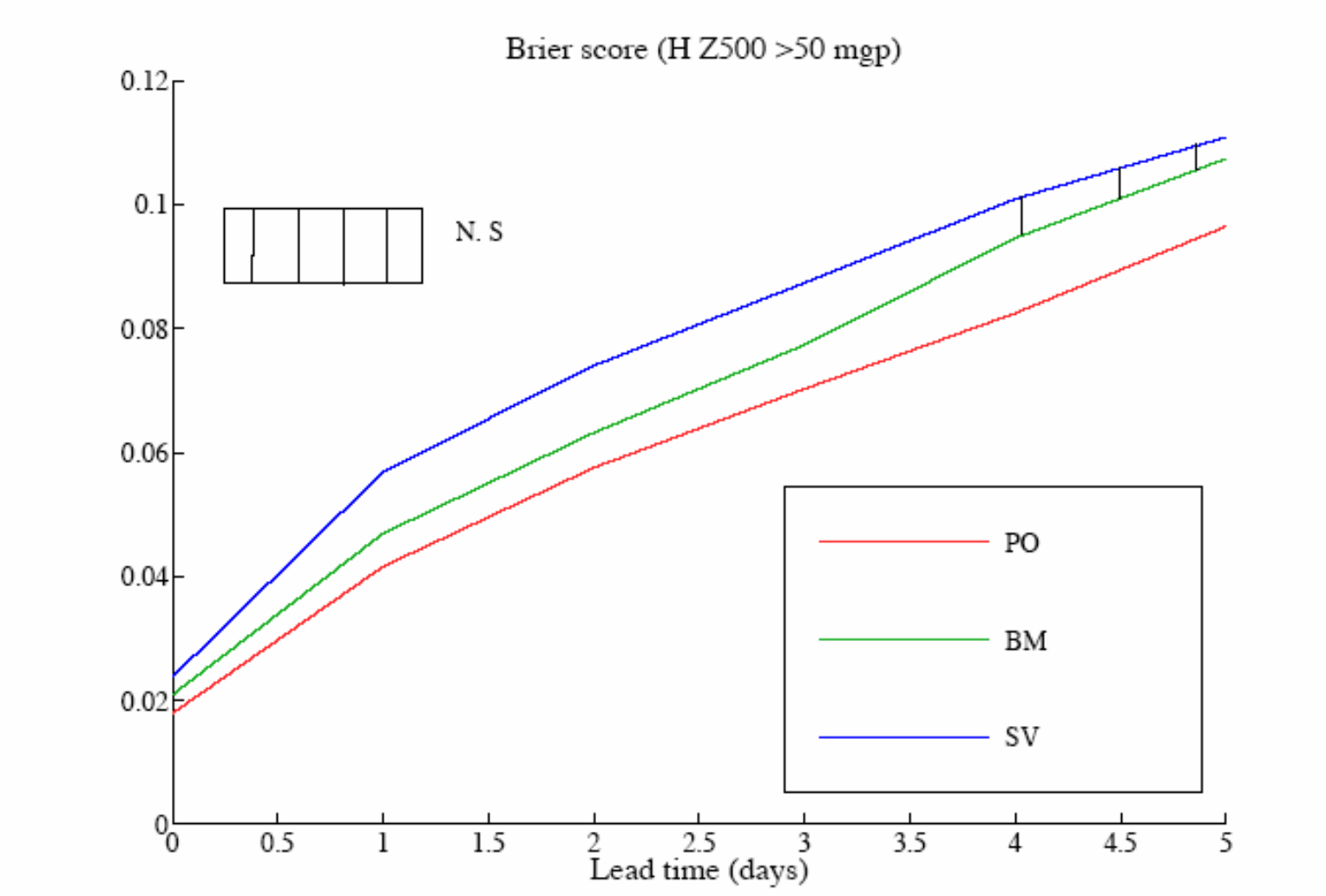


Figure 6: Evolution of the Brier score, as a function of lead time, for the four different methods: EnKF (solid line), ETKF (dotted line), BM (dashed line), SV (dash-dotted line) ; panel a: QG model; panel b: Lorenz model.

Descamps and Talagrand, *Mon. Wea. Rev.*, 2007



Arpège model (Météo-France)

Conclusion. If ensemble predictions are assessed by the accuracy with which they sample the future uncertainty on the state of the atmosphere, then the best initial conditions are those that best sample the initial uncertainty. Any anticipation on the future evolution of the flow is useless for the definition of the initial conditions.