

TIGGE/GIFS research needs & priorities

Olivier Talagrand

Laboratoire de Météorologie Dynamique, École Normale Supérieure, Paris, France

With contribution from C. Bishop, G. Candille, L. Descamps and C. Reynolds

Eighth meeting of the GIFS-TIGGE Working Group
World Meteorological Organization, Geneva, Switzerland
23 February 2010

THORPEX INTERACTIVE GRAND GLOBAL ENSEMBLE (TIGGE)

At present, ensemble short- and medium- range numerical weather forecasts are routinely available, produced by 10 different meteorological services and institutions, making up a global ensemble comprising about 200 ~ 300 elements.

How to make the best of those forecasts ?

The key objectives of TIGGE

An enhanced collaboration on development of ensemble prediction, internationally and between operational centres and universities,

- * New methods of combining ensembles from different sources and of correcting for systematic errors (biases, spread over-/under-estimation),
- * A deeper understanding of the contribution of observation, initial and model uncertainties to forecast error,
- * A deeper understanding of the feasibility of interactive ensemble system responding dynamically to changing uncertainty (including use for adaptive observing, variable ensemble size, on-demand regional ensembles) and exploiting new technology for grid computing and high-speed data transfer,

Test concepts of a TIGGE Prediction Centre to produce ensemble-based predictions of high-impact weather, wherever it occurs, on all predictable time ranges,

The development of a prototype future Global Interactive Forecasting System.

New methods of combining ensembles from different sources and of correcting for systematic errors (biases, spread over-/under-estimation),

- Calibration of model ensembles on the basis of past performance (typically, correcting ensemble mean for observed systematic bias, or ensemble spread on the basis of observed spread-skill relationship; see, e. g., Gneiting *et al.*, *MWR*, 2005; 'dressing' ensembles)
- Use of reforecasts, performed on past situations, for increasing size of training sample (Hamill *et al.*)
- Combining different ensembles, possibly by assigning them weights on the basis of observed past performance (Weigel and Bowler, *QJRMS*, 2009, Johnson and Swinbank, *QJRMS*, 2009)

Most, if not all, systems come with a *control forecast* which emanates for the operational analysis, and is performed at a higher spatial resolution than the ensemble forecasts.

The control forecast is statistically more accurate than a randomly chosen member of the predicted ensemble.

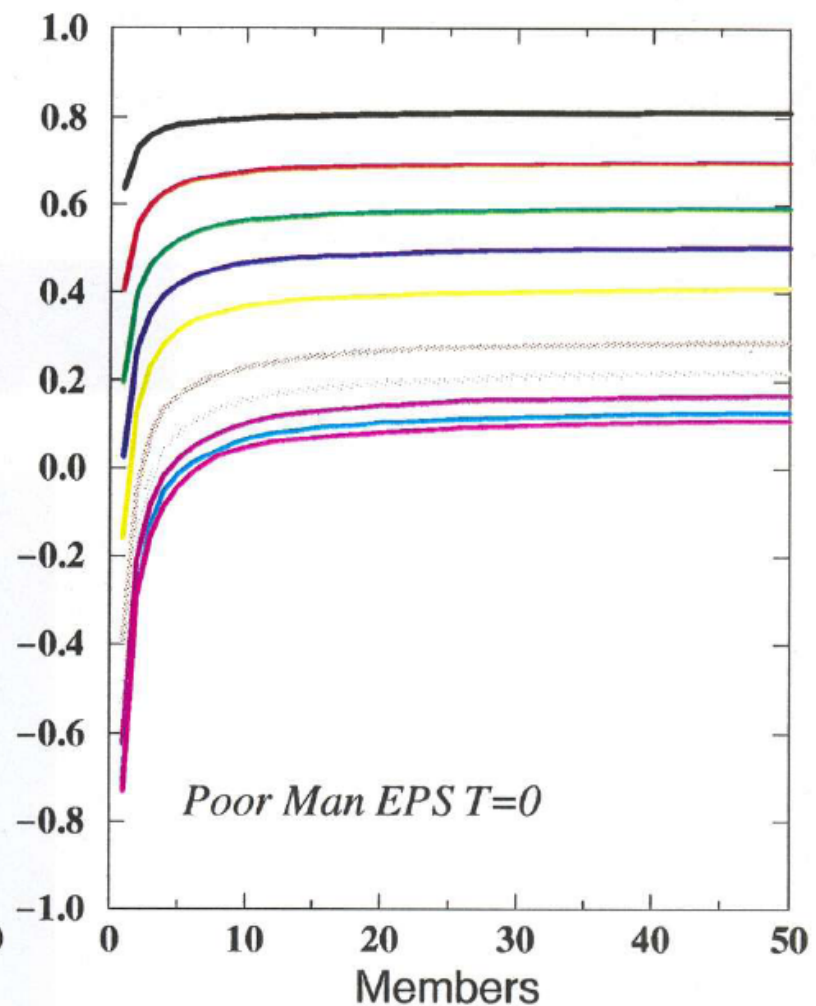
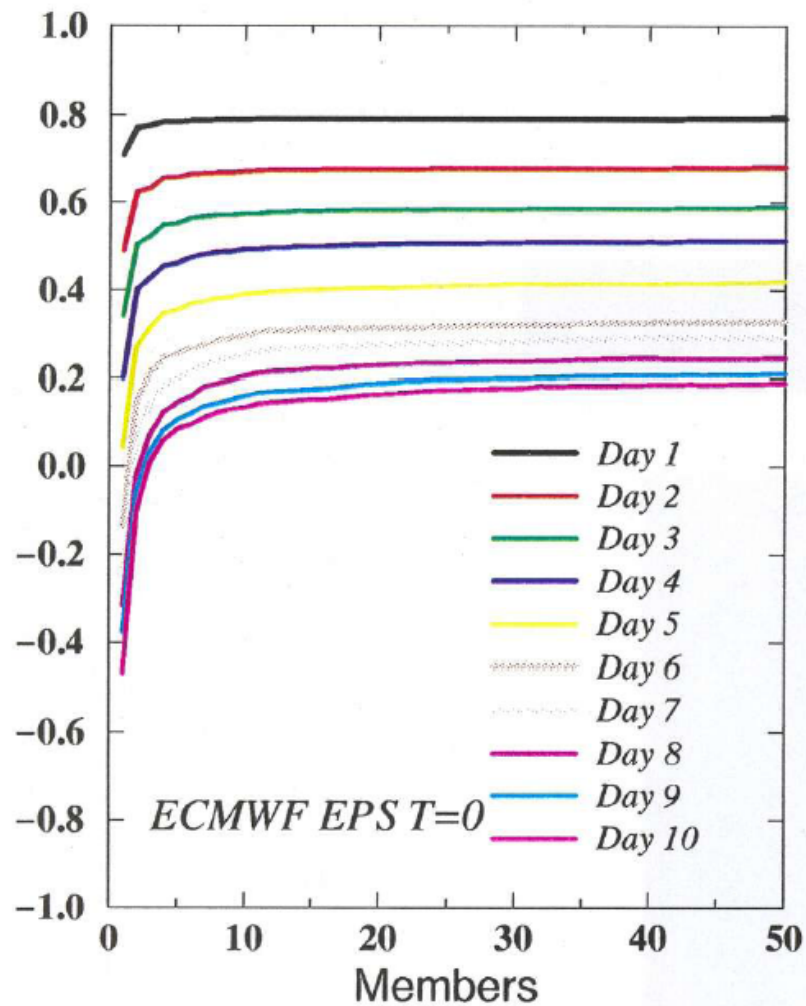
At initial time, the mean of the predicted ensemble coincides with the large scales of the operational analysis. That property is not conserved in the prediction.

The control forecast therefore contains information that is not contained in the predicted ensemble, especially at smaller scales. It contributes to further reduce the uncertainty on the predicted state. How to use that information ?

Evaluate what can, and cannot be achieved by ensemble (or, more generally, probabilistic) prediction.

Diagnostics to be performed on present (particularly TIGGE) systems

- Is non-gaussianity in ensembles meaningful (replace ensemble with gaussian pdf with same expectation and variance; see, *e. g.*, Atger)
- Compare performance with that of ‘poor man’s ensembles’ (obtained for instance from analogues in the archives of deterministic forecasts, or even analyses only)



Impact of ensemble size on Brier Skill Score
ECMWF, event $T_{850} > T_c$ Northern Hemisphere
(Talagrand *et al.*, ECMWF, 1999)

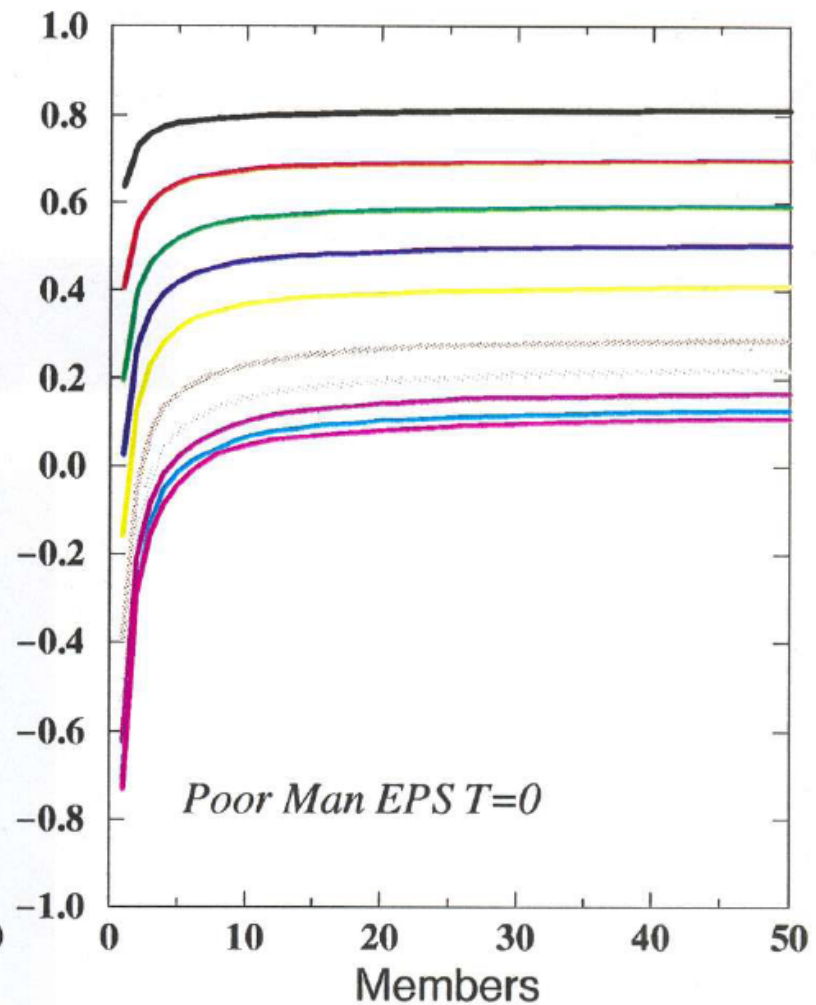
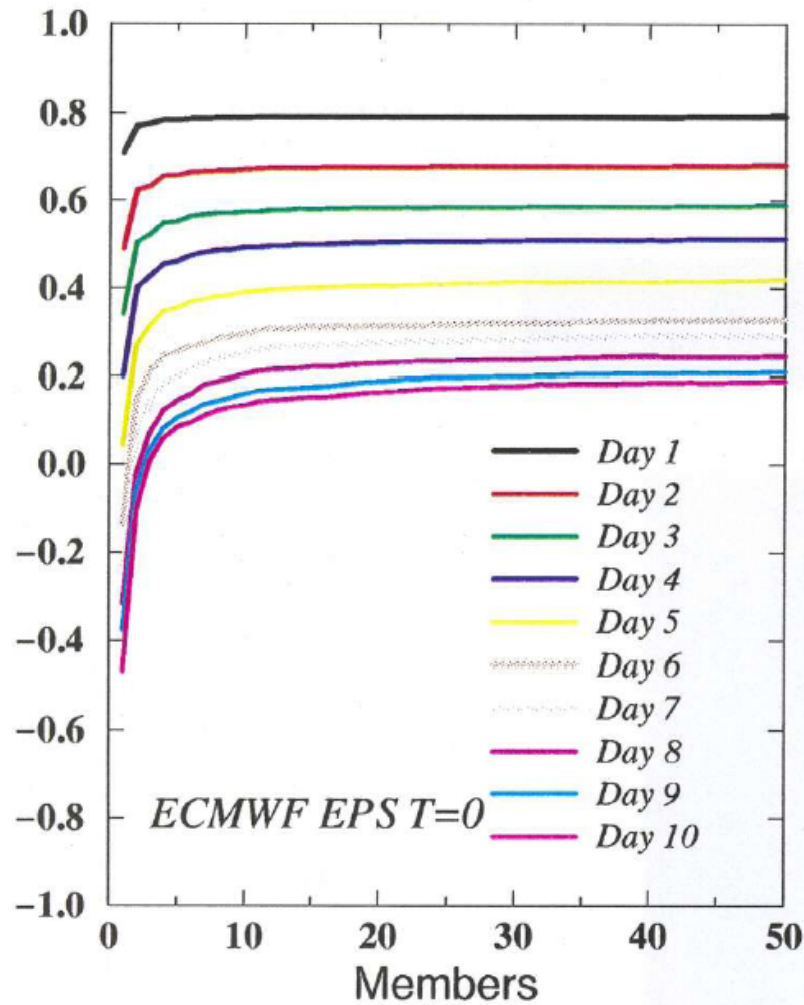
Theoretical estimate (raw Brier score)

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

Size of Ensembles ?

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

- Observed fact : in ensemble prediction, present scores saturate for ensemble size N in the range 30-50.



Impact of ensemble size on Brier Skill Score
ECMWF, event $T_{850} > T_c$ Northern Hemisphere
(Talagrand *et al.*, ECMWF, 1999)

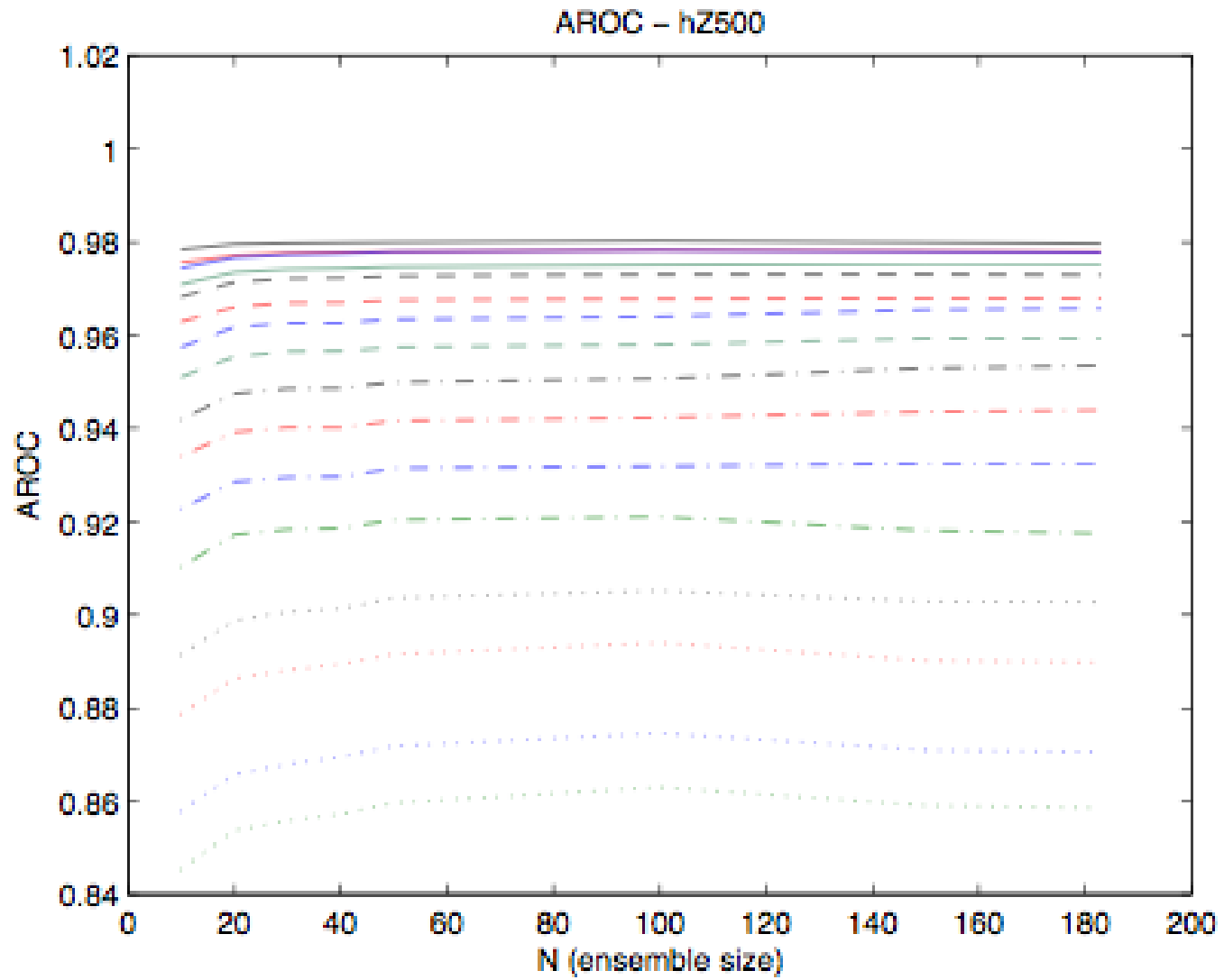
Theoretical estimate (raw Brier score)

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

Brier score for ensembles of size N (Talagrand *et al.*, 1999)

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

where $g(p)$ is the relative frequency with which the system predicts probability p . The sharper the distribution of raw predicted probabilities, the more rapid the saturation of the score.



TIGGE, ROC curve area, courtesy L. Descamps

Simulations $M = 100000 : E=(X < X_{\text{clim}} - \sigma)$

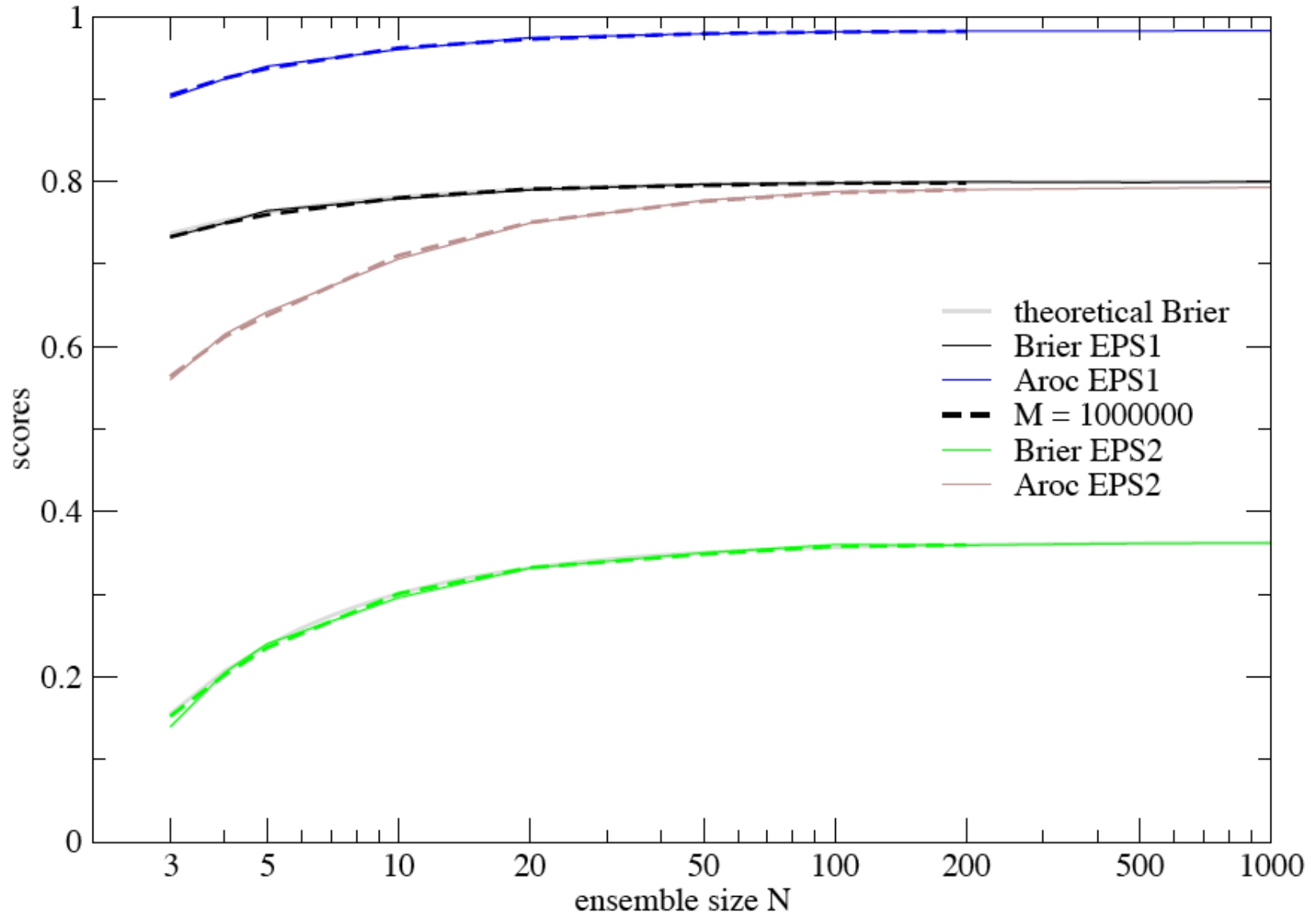


Figure 1: Impact of N on Brier Skill Score and ROC area

Scores saturate for ensemble sizes N of the order of a few tens. The higher the sharpness of the predicted probabilities, the more rapid the saturation.

Question

Is there any point in taking larger values of N ?

Questions

- o If we take, say, $N = 200$, which user will ever care whether the probability for rain for to-morrow is $123/200$ rather $124/200$?
- o And even if a user cares, 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. E must have occurred $\alpha N/10$ 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 $N = 500$ ($\alpha = 4$).

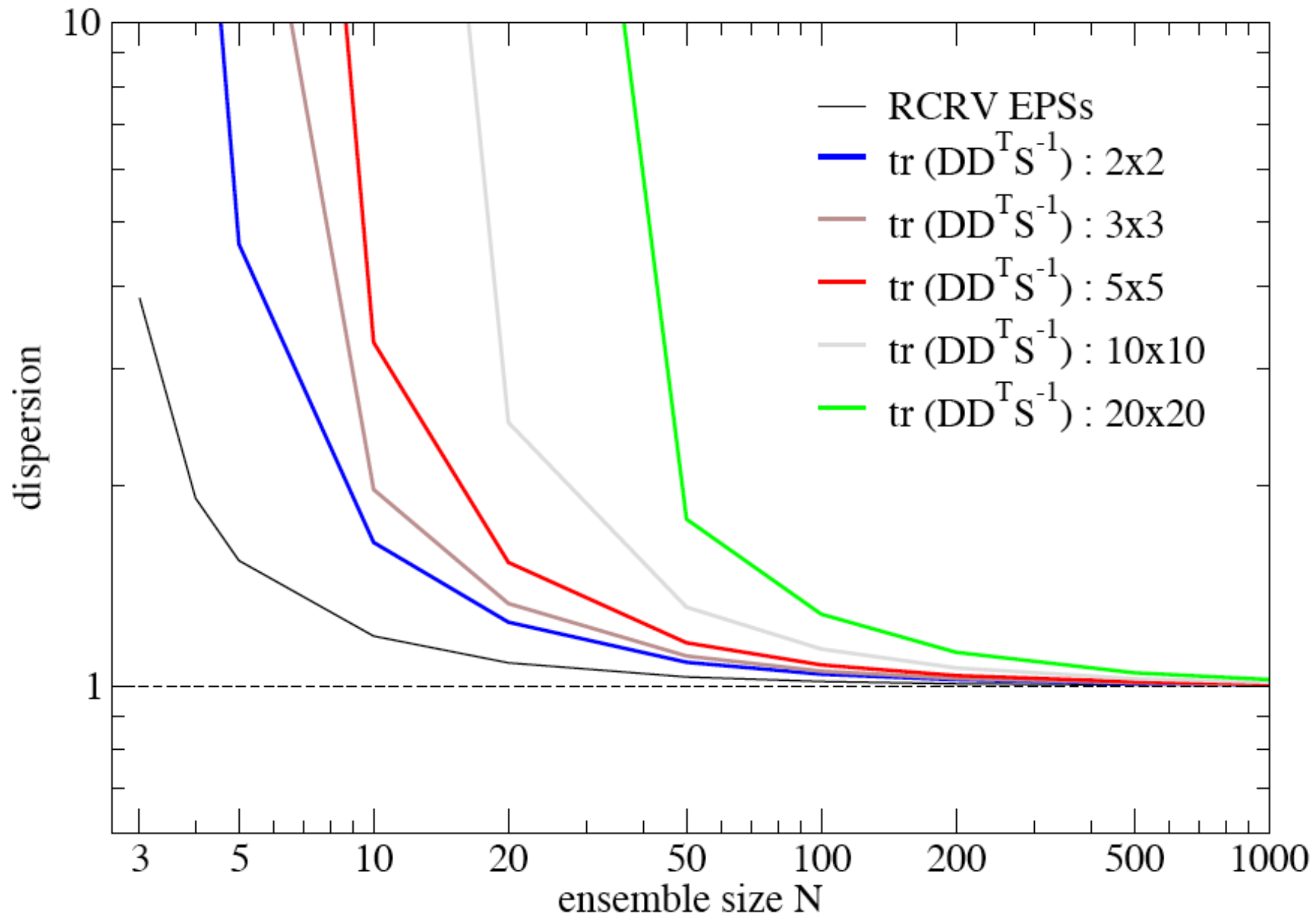
Conclusion. Reliable large- N probabilistic prediction of (even moderately) rare events is simply impossible.

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 probabilistic 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-degree² 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.

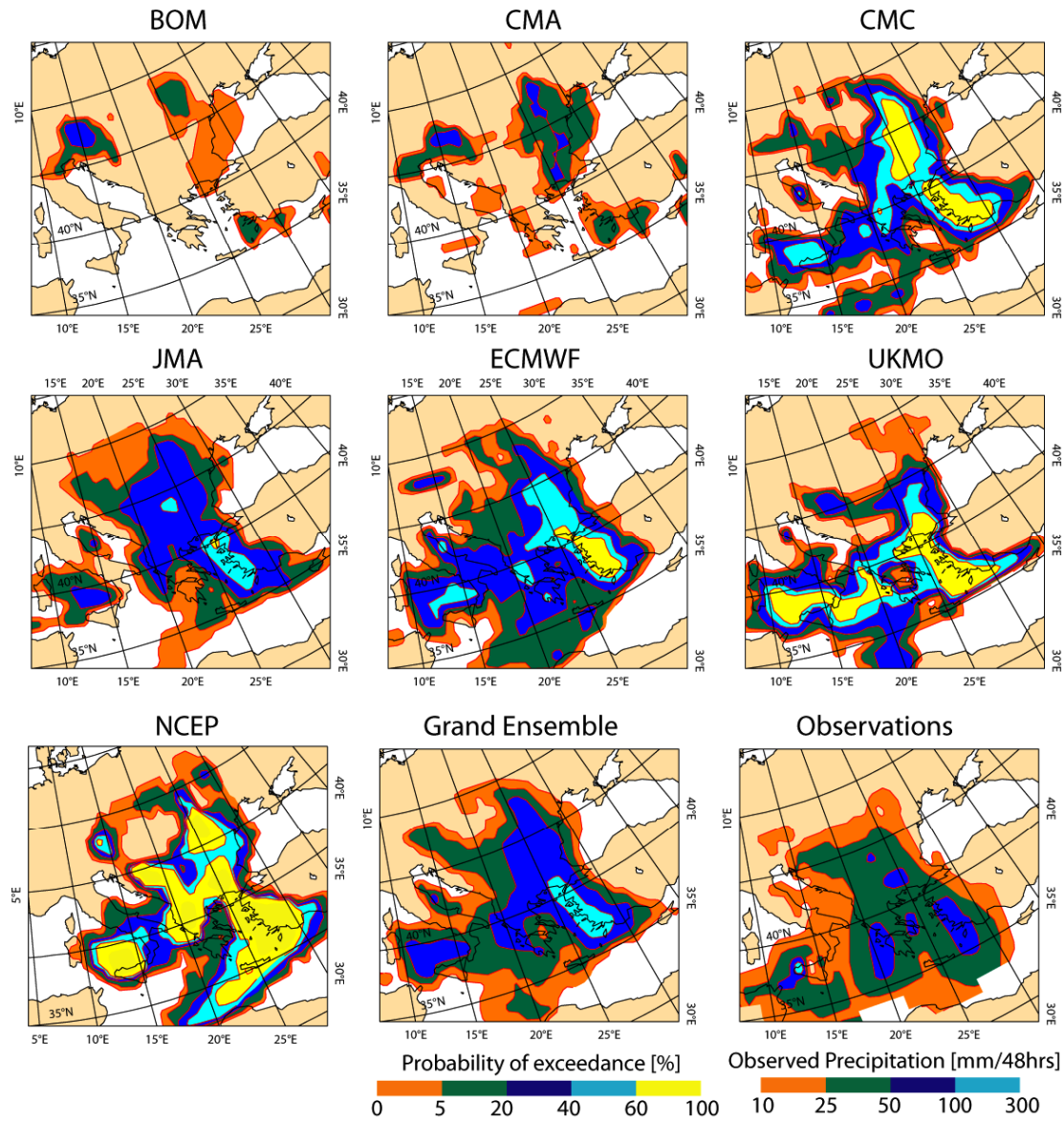
In view of the lifetimes of the regimes, there is no point in making more than one forecast per week. That would make 60 forecasts over a 3-year period. Hardly sufficient for accurate validation.

Conclusion on ensemble size

Objective scores saturate in the range $N \approx 30-50$ because it is possible in practice to evaluate only probabilistic predictions of events or one-dimensional variables. Evaluating probabilistic predictions of multi-dimensional variables would require validating samples of inaccessible size.

Is there any point in taking larger values of N ?

Probability of exceeding 25mm/48hrs, Forecast date: 18.10.2007, lead time: 3-5days



There is often a significant correlation between the predicted probability for intense precipitation and the observed amount of precipitation, so that the former can be used as a deterministic predictor of the latter.

Why is it so (it need not be) ? And how to exploit that fact in practice ?

Other questions

- A deeper understanding of the contribution of observation, initial and model uncertainties to forecast error (one of the stated key objectives of TIGGE). In particular, how to objectively quantify the model errors ?

Dispersion of forecasts performed with (exactly) the same model from different initial conditions is a statistical integrated measure of sensitivity to initial conditions. Comparison with deviation from observed reality then provides by difference a statistical integrated measure of model error (idea behind 'Lorenz curves').

Other questions (continued)

- How to define initial ensemble ?

My answer (cf work of L. Descamps). Do ensemble assimilation that is meant to sample the uncertainty on the state of the flow at the start of the forecast, and go on with the ensemble into the forecast. No need to specifically identify unstable modes, either in the assimilation or the forecast phase (except if it reduces cost, as in the case of bred modes).

But in disagreement with results obtained by R. Buizza at ECMWF.

For LAMs, how to define lateral boundary conditions ?

- Develop probabilistic forecasts for occurrence of specific phenomena (tropical cyclones, storms, polar lows,)

Other questions (continued)

- What do we want from ensemble, or more generally, probabilistic prediction (prediction of probabilities and probability distributions, more accurate deterministic forecast, bounds on meteorological variables, scenarii, ...) ?
- Must we tend to a situation where the output of prediction will be a probability distribution (will require appropriate use of control forecasts) ?

Research priorities (personal selection)

- *A posteriori* calibration of all forms (debiasing, ‘dressing’ of ensembles, ...). Mutually combine ensembles produced by different models. Already done actively.
- Determine how to use information contained in control forecasts.
- Determine limits of what can reasonably be expected from ensemble prediction, and use resources in a way that is appropriate to those limits.
- Determine most cost-efficient way of defining initial (and lateral) conditions ?
- Quantify model errors