On Some Aspects of Ensemble Prediction

O. Talagrand[1], G. Candille[1,2] and L. Descamps[1,3]

1. Laboratoire de Météorologie Dynamique, École Normale Supérieure, Paris, France
2. Université du Québec à Montréal, Montréal, Canada
3. Centre National de Recherches Météorologiques, Météo-France, Toulouse, France

Workshop *Dynamics and Statistics in Weather and Climate*
Max-Planck-Institut für Physik komplexer Systeme, Dresden, Germany
30 July 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*

There is (and, as far as we know, there will always be) uncertainty on the future state of the atmosphere. That uncertainty varies from day to day, or at least from week to week.

That uncertainty is (and, as far as we know, will always be) large enough so that the question of quantifying it *a priori* seems worth investigating (for instance, in anticipation of situations where a user must make a decision that can involve a financial risk in relation with weather).

For some reason, uncertainty is conveniently described by probability distributions (don't know too well why, but it works; see also Jaynes, E. T. (edited by G. L. Bretthorst), 2007, *Probability Theory: The Logic of Science*, Cambridge University Press, Cambridge, United Kingdom, 727 pp., ISBN 978-0-521-59271-0.).

Prediction of uncertainty 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, Inverse Problem Theory and Methods for Model Parameter Estimation, Society for Industrial and Applied Mathematics, Philadelphia, USA, 342 pp., ISBN 0-89871-572-5).

Remark. Problem is the same for evaluation of the present state of the system, and for assimilation of observations.

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 $n \approx 10^{6-8}$ of present Numerical Weather Prediction models.

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

This has led a number of meteorological services to develop ***Ensemble Prediction Systems*** (*EPS*s), which produce an ensemble of estimates of the future state of the flow, which are meant to sample the corresponding conditional probability distribution.

Global Ensemble Prediction Systems have been run operationally at

- National Centers for Environmental Prediction (NCEP) since 1992
- European Centre for Medium-range Weather Forecasts (ECMWF) since 1992
- Meteorological Service of Canada (MSC) since 1998
- …

In addition, *Regional Ensemble Prediction Systems*, intended at more or less local applications *(e. g.*, prediction of meteorological conditions at an airport) are run in a number of meteorological services (UK, France, Italy, …)

More recently, set-up of

THORPEX INTERACTIVE GRAND GLOBAL ENSEMBLE (TIGGE)

About 20 meteorological services contribute their ensemble predictions.

Ensembles are basically produced from runs of a numerical deterministic model of the flow that differ through initial conditions, but also through specific features in the forecast model (‘*stochastic physics*’ at ECMWF). *Multi-model ensembles*, in which runs from different models are merged together in the same ensemble, are more and more frequent (TIGGE).

Initial conditions are defined through various procedures : singular modes (ECMWF), Ensemble Transform Kalman Filter (ETKF) (NCEP), ‘perturbed observation’ assimilation (MSC).

Size of ensembles typically lies in the range ***N ~ 10 - 100***

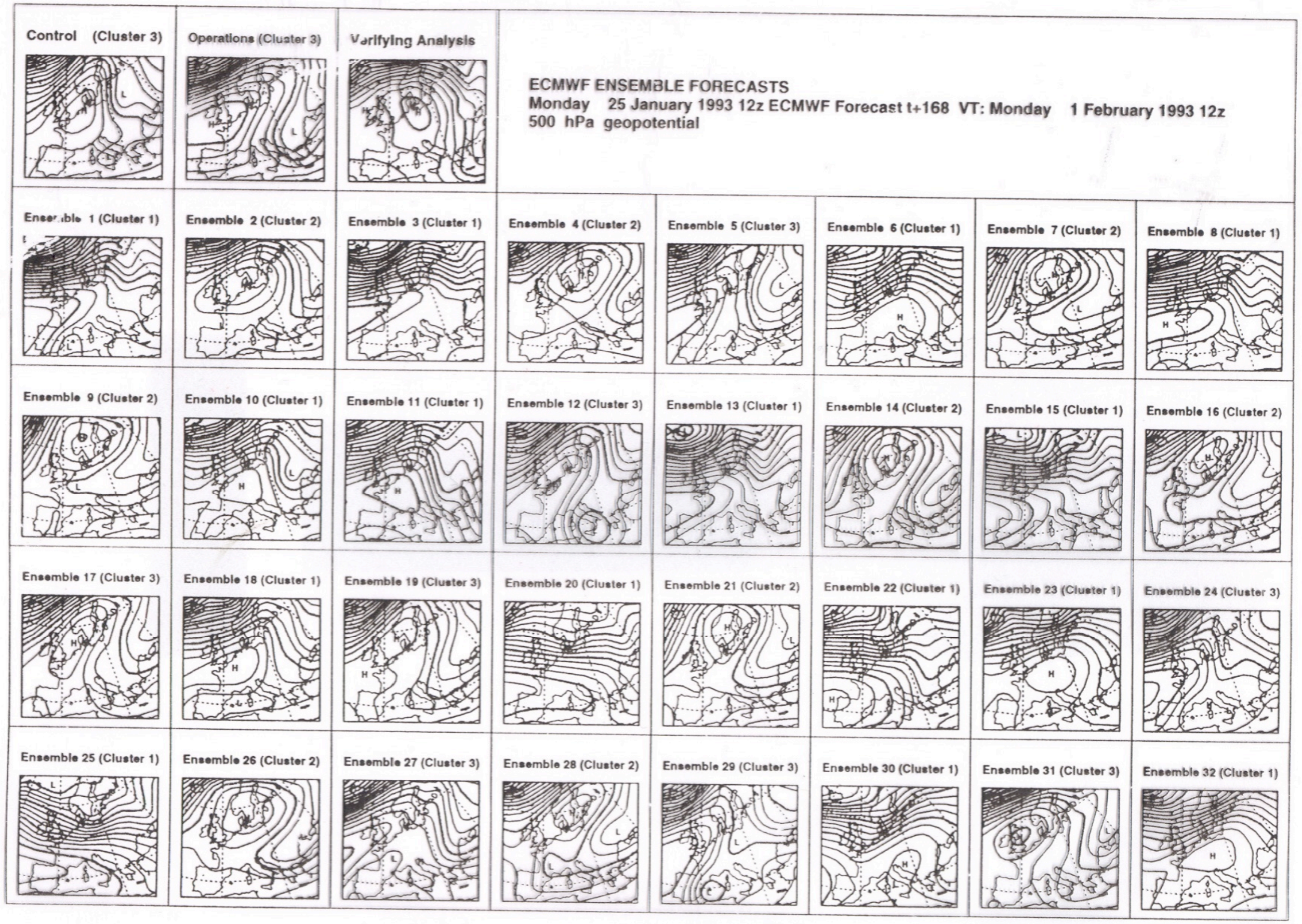


Fig. 1: Members of day 7 forecast of 500 hPa geopotential height for the ensemble originated from 25 January 1993.

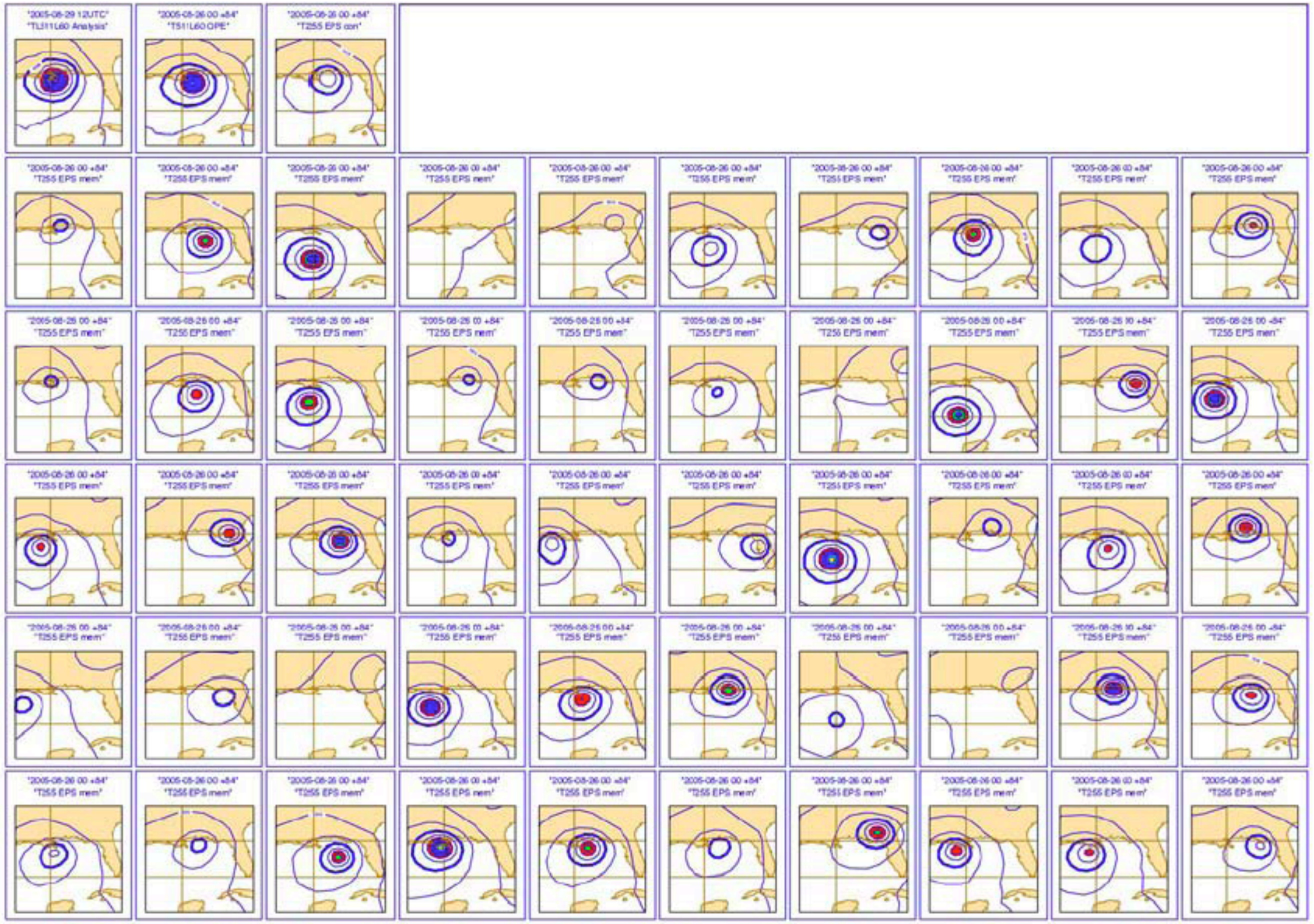

"2005-08-29 12UTC" "TL511L60 Analysis"
"2005-08-26 00 +84" "T511L60 OPE"
"2005-08-26 00 +84" "T255 EPS con"
"2005-08-26 00 +84" "T255 EPS mem"

Figure 6 Hurricane Katrina mean-sea-level-pressure (MSLP) analysis for 12 UTC of 29 August 2005 and t+84h high-resolution and EPS forecasts started at 00 UTC of 26 August:

1st row: 1st panel: MSLP analysis for 12 UTC of 29 Aug
2nd panel: MSLP t+84h T_L511L60 forecast started at 00 UTC of 26 Aug
3rd panel: MSLP t+84h EPS-control T_L255L40 forecast started at 00 UTC of 26 Aug
Other rows: 50 EPS-perturbed T_L255L40 forecast started at 00 UTC of 26 Aug.

The contour interval is 5 hPa, with shading patters for MSLP values lower than 990 hPa.

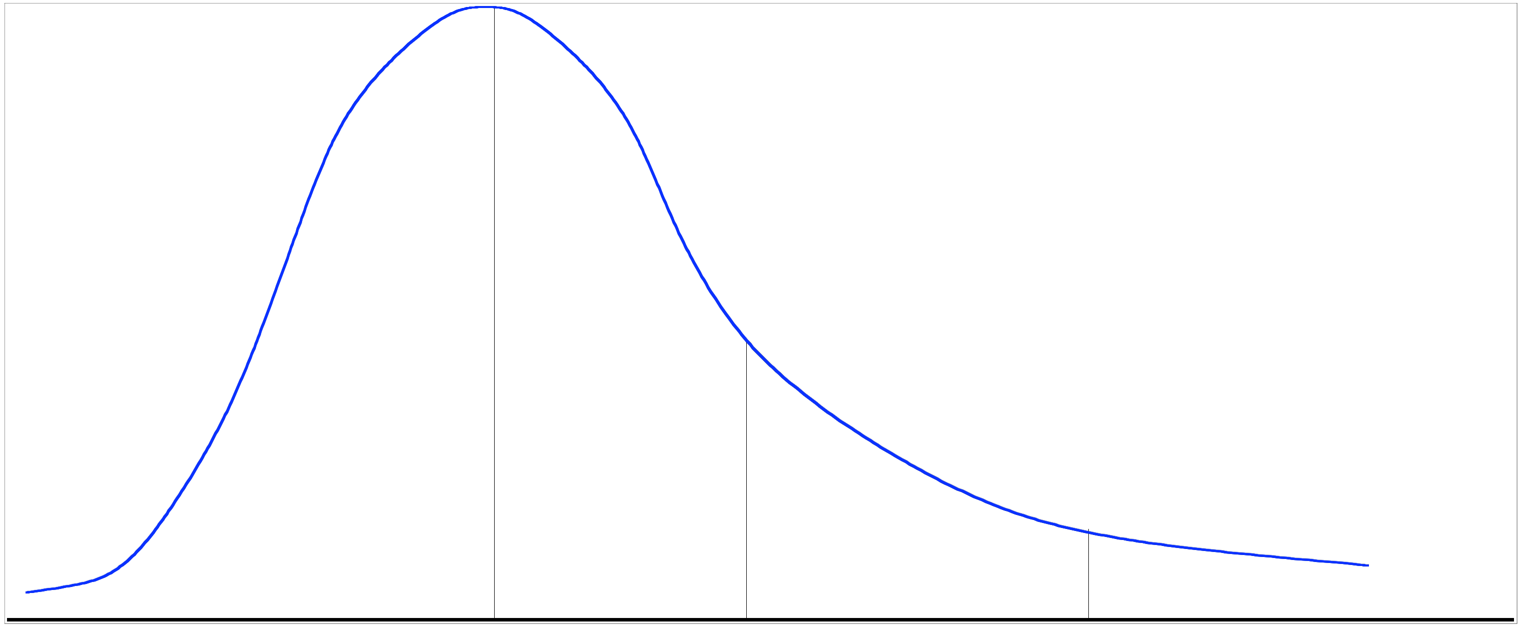
My point of view

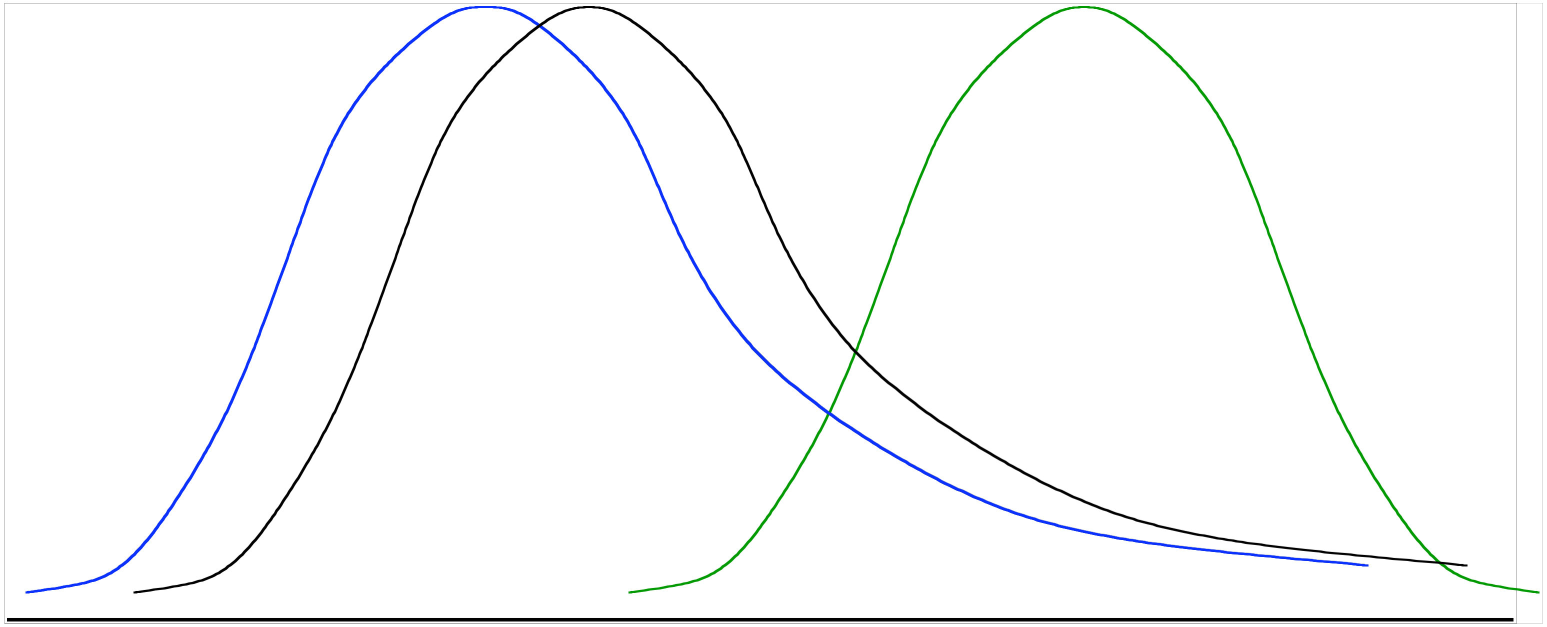
- Ensemble estimation (either prediction or assimilation) is of a different essence than deterministic estimation in that the object to be estimated (basically a probability or a probability distribution) is not better known *a posteriori* than it was *a priori* (in fact, that object has no objective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

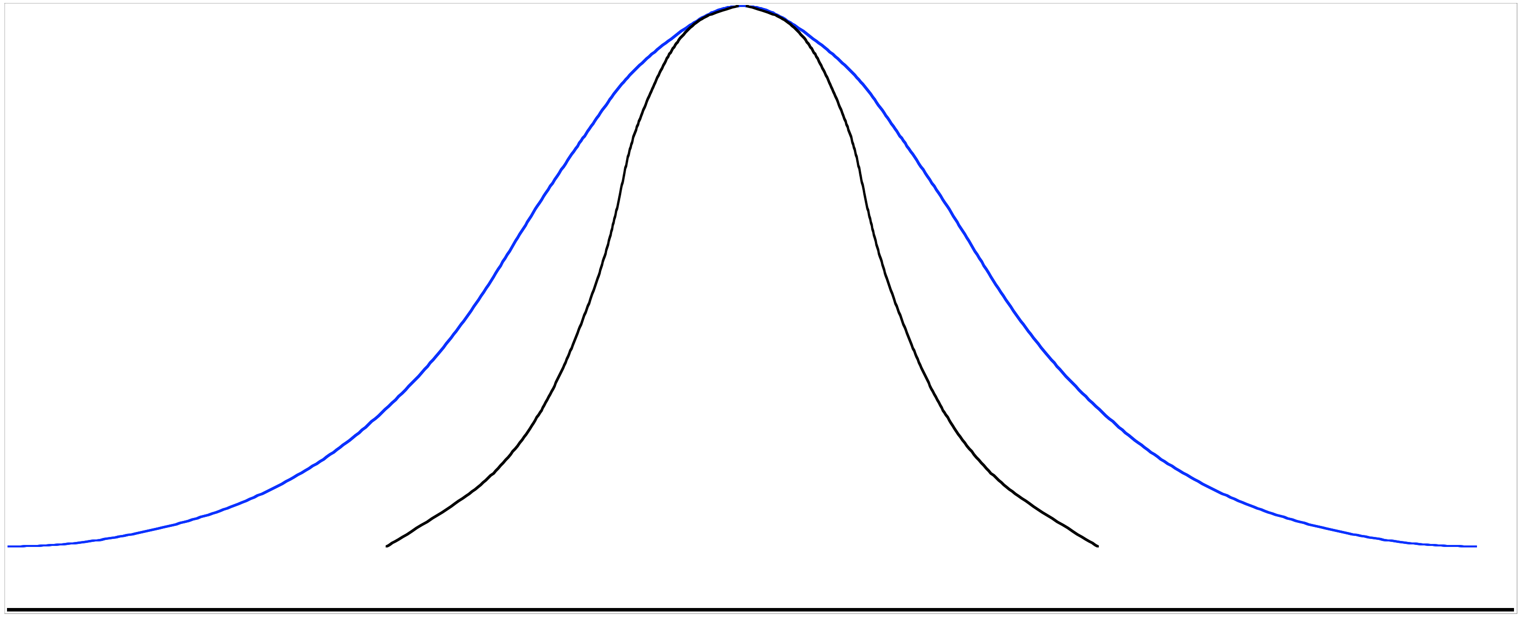
- As a consequence, validation of ensemble estimation can only be statistical, and it is meaningless (except in limit cases, as when the estimated probability distribution has a very narrow spread, and the verifying observation falls within the predicted spread, or on the contrary when the verifying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

Question

- The purpose of ensemble estimation being 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 ?







Statistical consistency between prediction and observation

Rain must occur with frequency 40% in the circumstances when it has been predicted to occur with probability 40%.

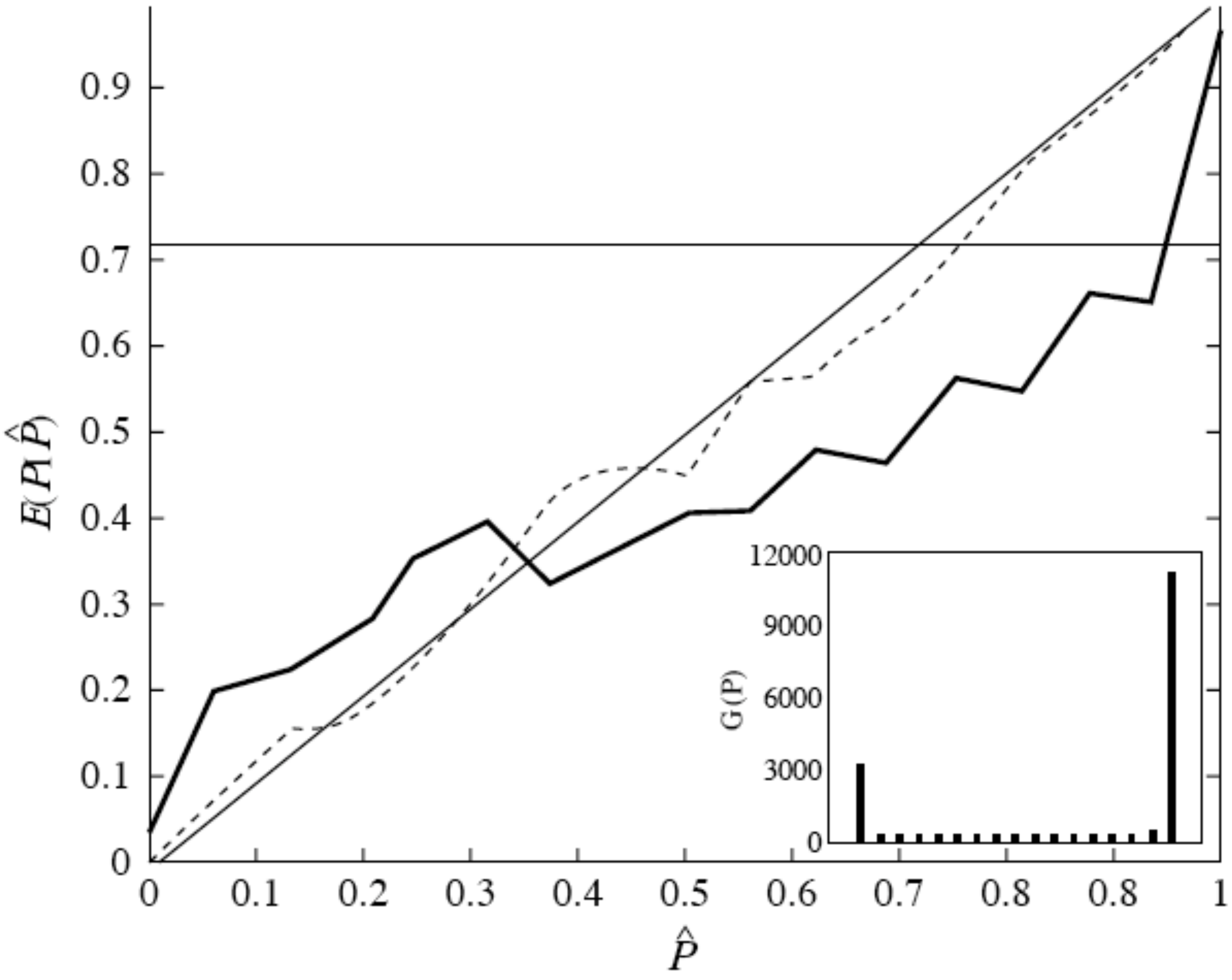
Observed frequency of occurrence $p‘(p)$ of event, given that it has been predicted to occur with probability p, must be equal to p.

For any p, $p‘(p) = p$

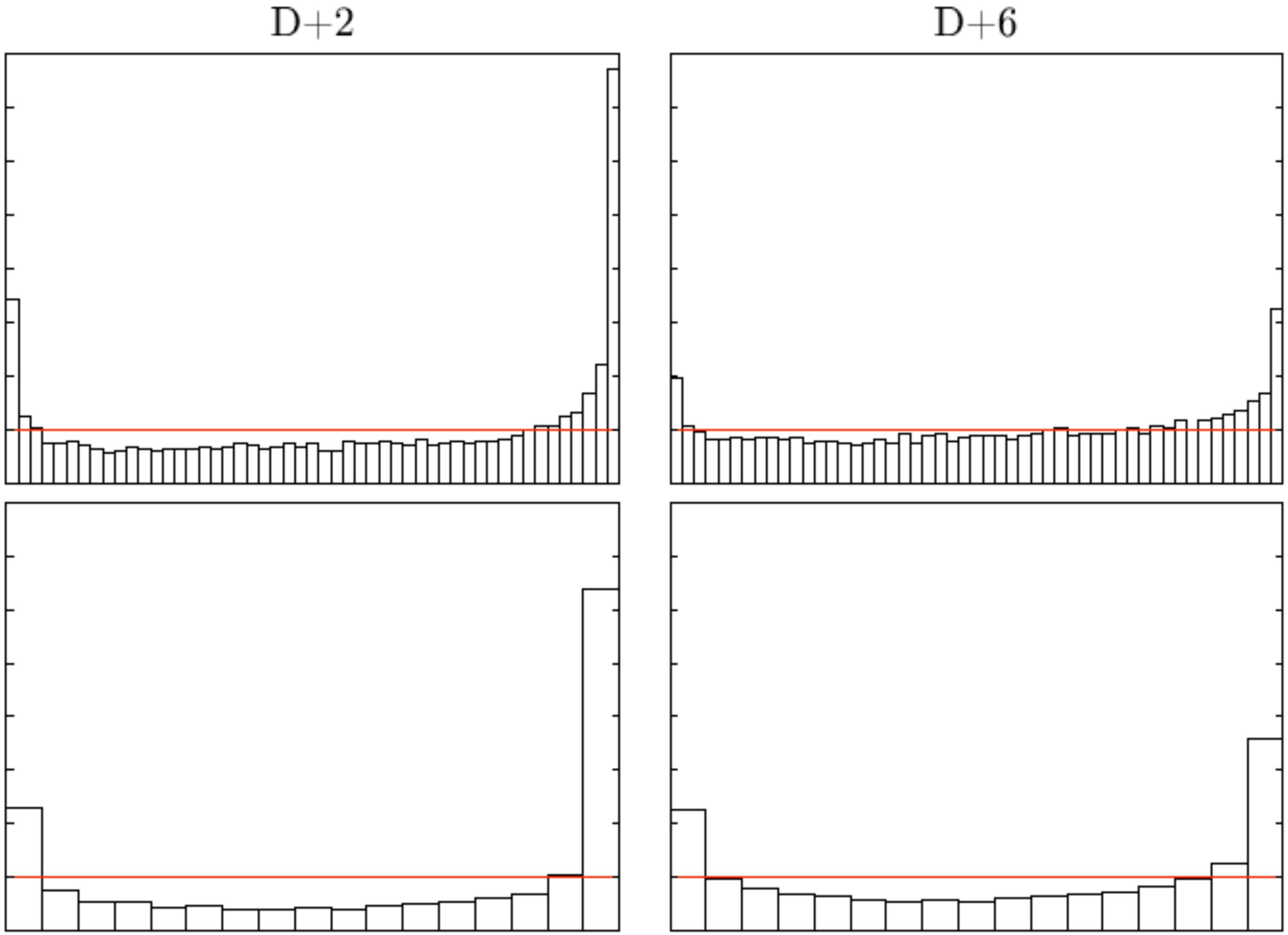
Reliability

More generally, frequency distribution of observation $F‘(F)$, given that probability distribution F has been predicted, must be equal to F.

For any F, $F‘(F) = F$



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



Rank histograms, T_{850}, Northern Atlantic, winter 1998-99

Top panels: ECMWF, bottom panels: NMC (from Candille, Doctoral Dissertation, 2003)

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

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

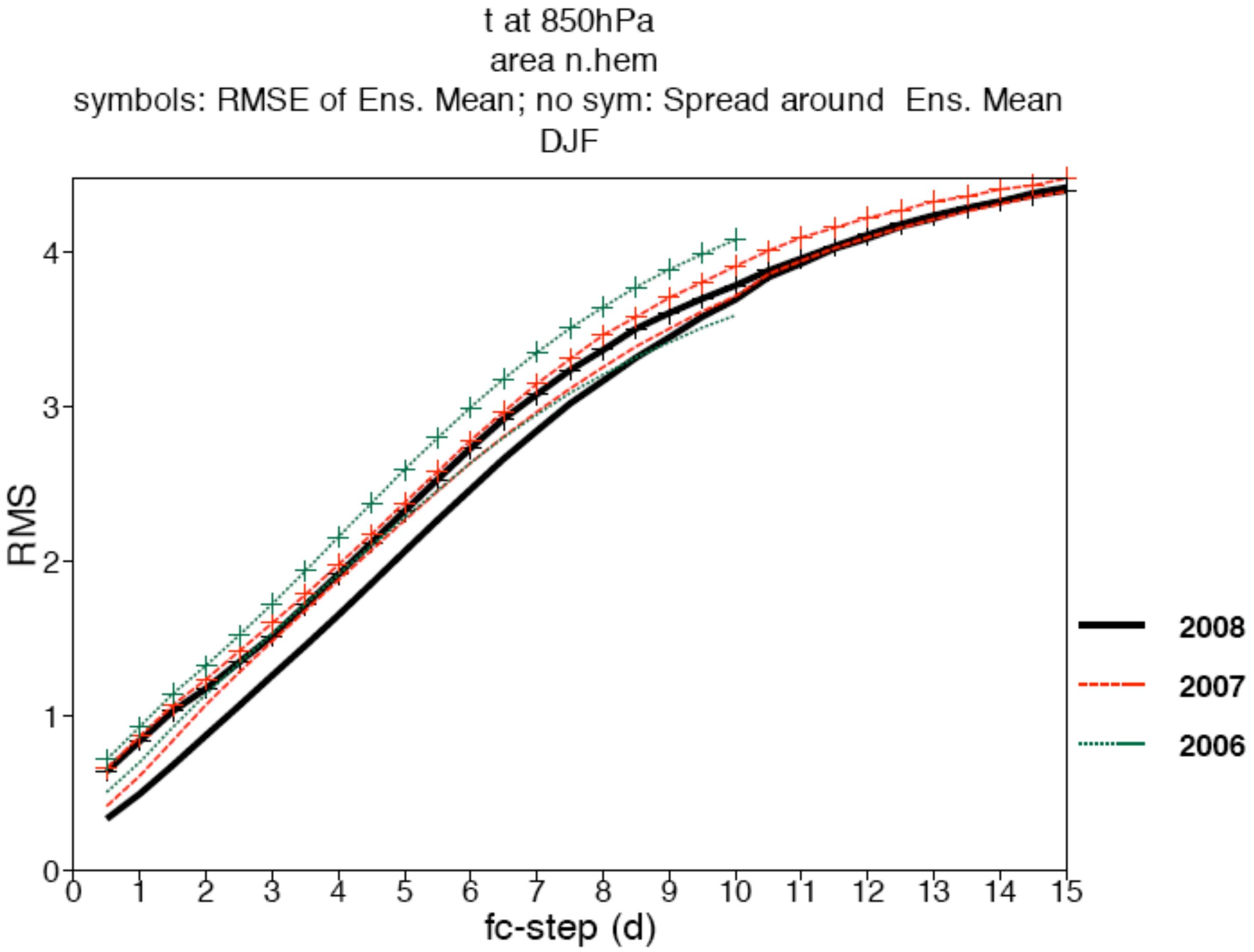


Figure 8: Ensemble spread (standard deviation) and root mean square error of ensemble-mean (lines with crosses) for 500 hPa height (top) and 850 hPa temperature (bottom) for winter 2007-08 (black), 2006-07 (red) and 2005-06 (green) over the extra-tropical northern hemisphere.

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

The degree of reliability of an Ensemble Prediction System is measured by a number of non equivalent objective scores : reliability component of Brier and Brier-like scores, rank histograms, Reduced Centred Variable, …

If sample of realizations of the system is large enough, *a posteriori* calibration is in principle possible

$$F \Rightarrow F`(F)$$

‘*Experience shows that, when you predict F, reality is distributed according to F’. So, next time you predict F, I will predict F’*’

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

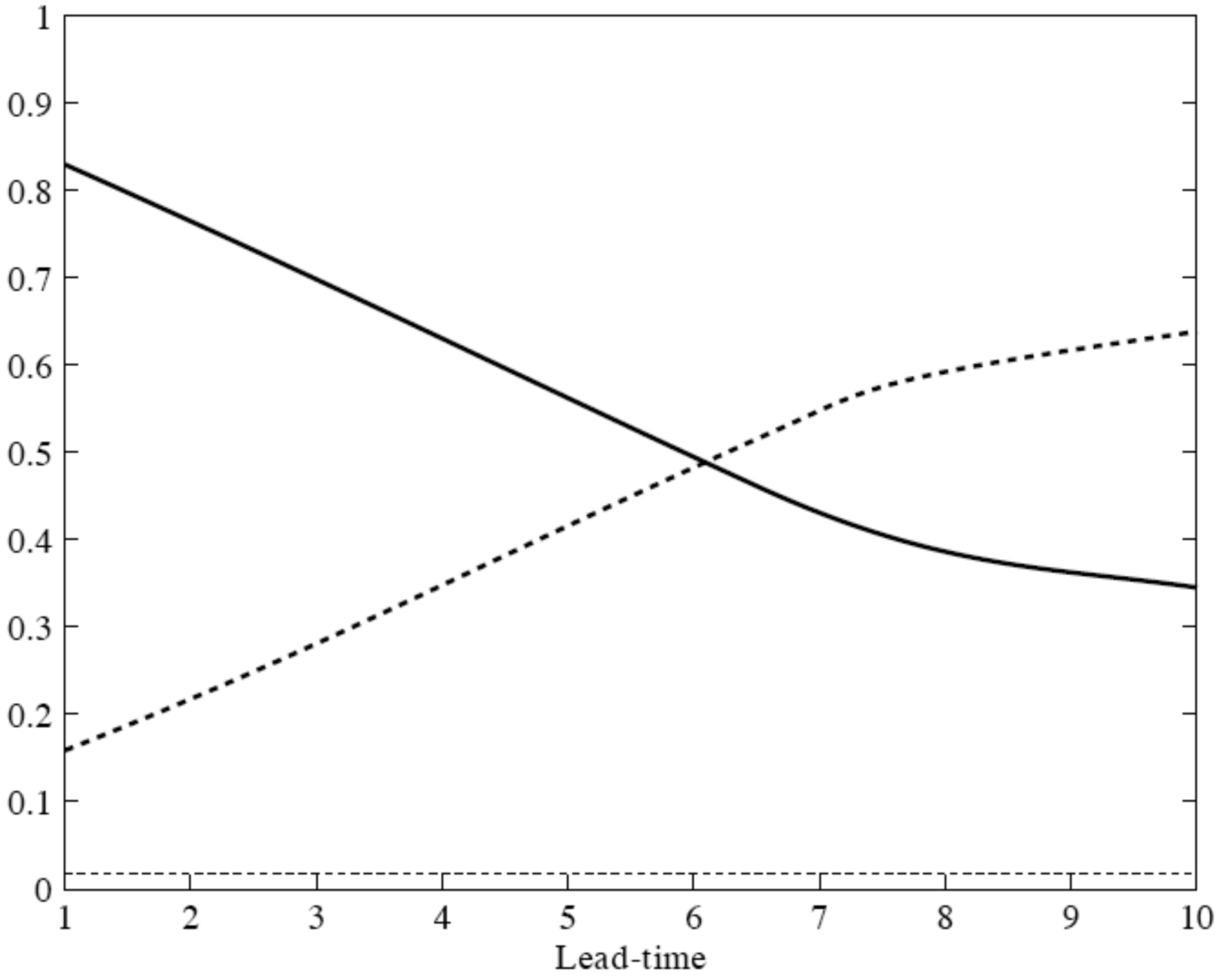
But that is obviously not sufficient to ensure system is practically useful. A system which always predicts climatological frequency of occurrence (or climatological frequency distribution) is reliable in the sense that has just been defined, but nevertheless totally useless.

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, …



Brier Skill Score and components, ECMWF, event $T_{850} > T_c$ - 2C, Northern Atlantic Ocean, December 1998 - February 1999

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.

A highly desirable property of scores is that they are ***proper***. A proper score is a score that cannot be cheated, *i. e.*, a score that, for given probability distribution G of the verifying observation, assumes its optimum value when the predicted probability distribution is equal to G.

A proper score that is of the form (J. Bröcker, L. Smith)

$$E[S(F, \xi)]$$

where S is a function of F and of the verifying observation ξ, can be decomposed into a reliability and a resolution component (like the Brier score).

Definition of initial ensembles

Different 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 before forecast, and of ‘future’ singular vectors determined over the first 48 hours of the forecast period. Mixture of past and future.

- ‘Bred’ modes, then Ensemble Transform Kalman Filter (NCEP)

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

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

- ‘Perturbed observation’ method (formerly at 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.

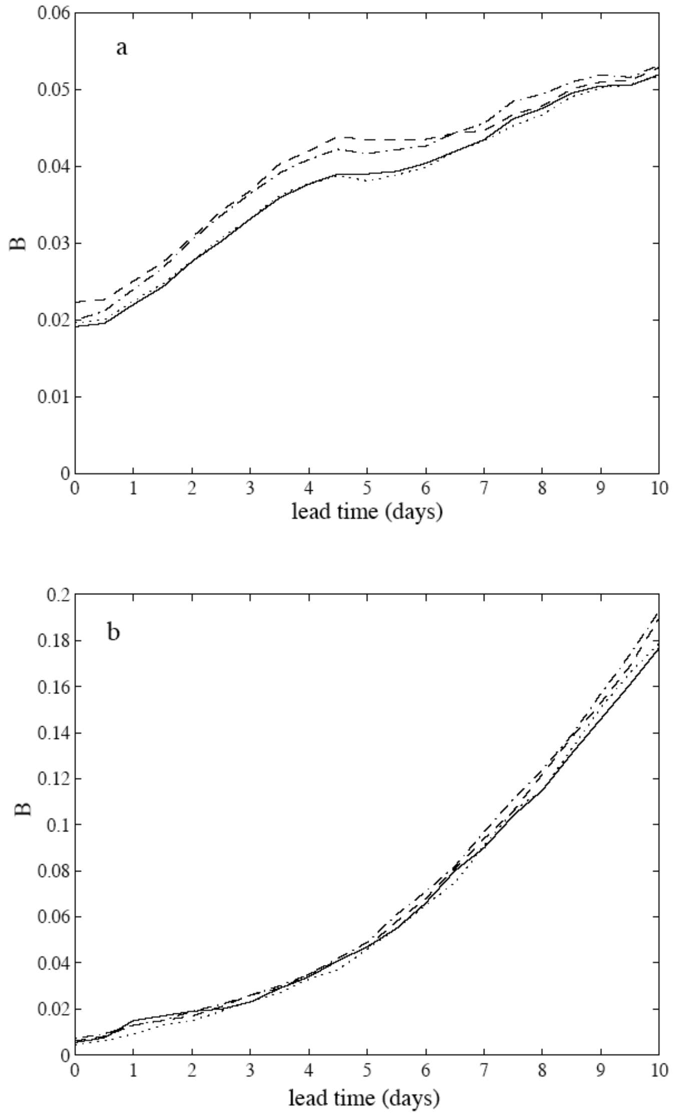
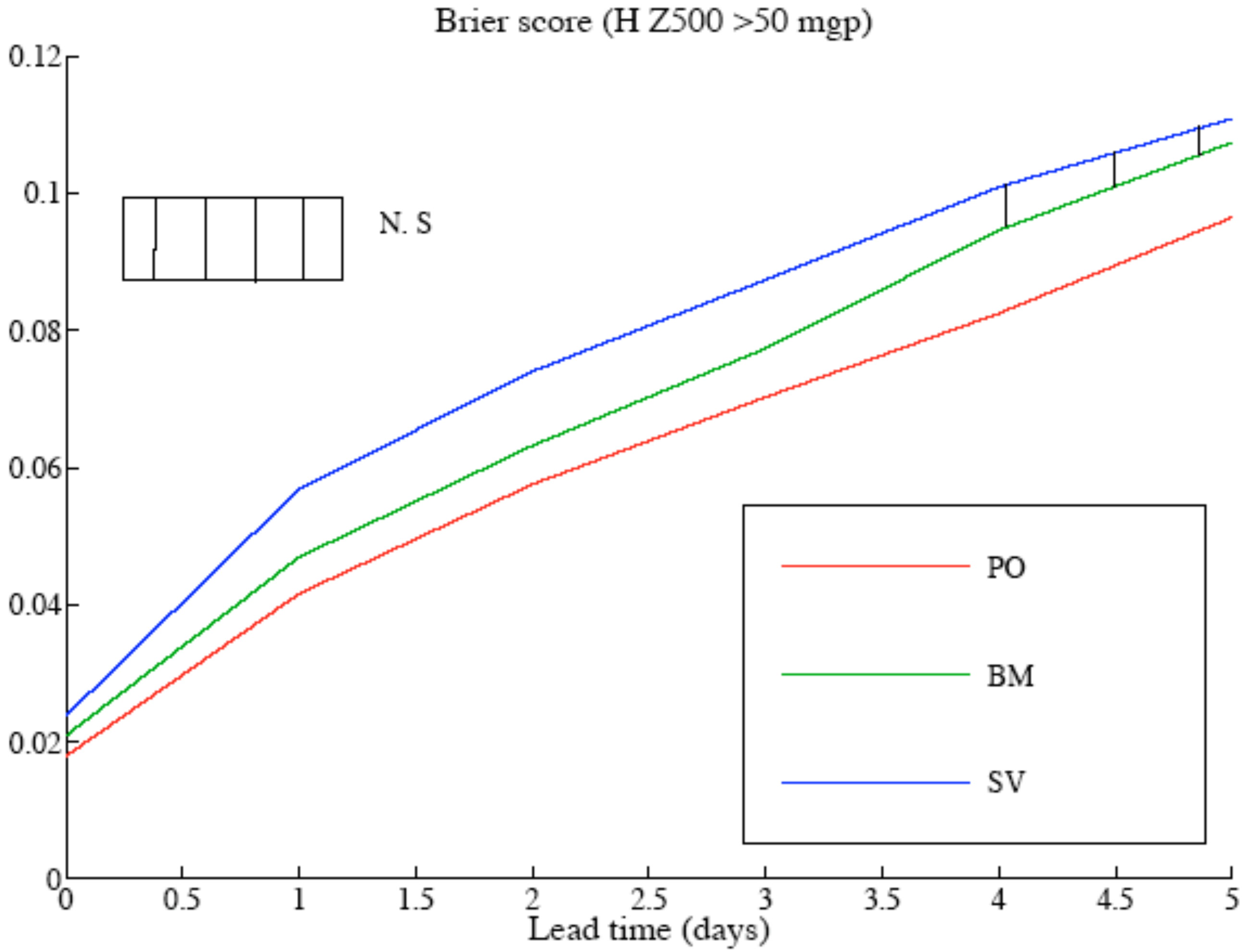


Figure 6: Evolution of the Brier score, as a function of lead time, for the four different methods: EnKF (solid line), ETKF (dotted line), BM (dashed line), SV (dash-dotted line) ; panel a: QG model; panel b: Lorenz model.

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



Arpège model (Météo-France)

Table 1: AREA UNDER THE ROC CURVE AT VARIOUS LEAD TIMES FOR EVENT EV1 AND THE THREE METHODS (500-hPa geopotential).

Lead time (days)	PO	BM	SV
0	0 .98	0 .95	0 .93
1	0 .94	0 .90	0 .88
2	0 .91	0 .87	0 .85
3	0 .87	0 .83	0 .82
4	0 .82	0 .77	0 .76
5	0. 76	0. 71	0. 71

Table 2: SAME AS Table 1, BUT FOR EVENT EV2.

Lead time (days)	PO	BM	SV
0	0 .97	0 .94	0 .91
1	0 .92	0 .88	0 .86
2	0 .89	0 .84	0 .82
3	0 .84	0 .78	0 .77
4	0 .78	0 .72	0 .72
5	0 .71	0 .67	0 .66

Arpège model, Météo-France

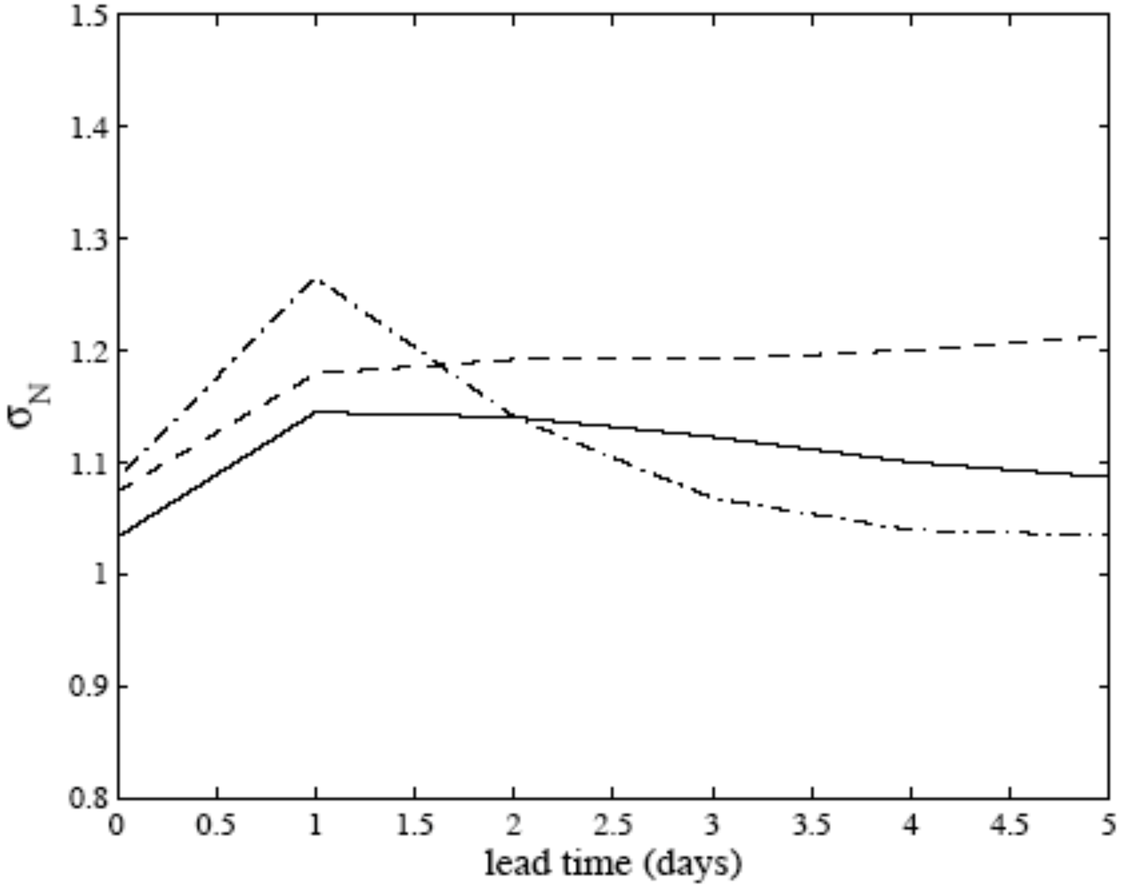


Fig. 5. Evolution of the Standard deviation of the Reduced Centered Random Variable (850-hPa temperature), as a function of lead time, for the three different methods: PO (solid line), BM (dashed line), SV (dash-dotted line)

Descamps and Talagrand, 2008

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. And diagnostics intended at identifying past unstable modes of the flow as such are not as efficient as ensemble assimilation.

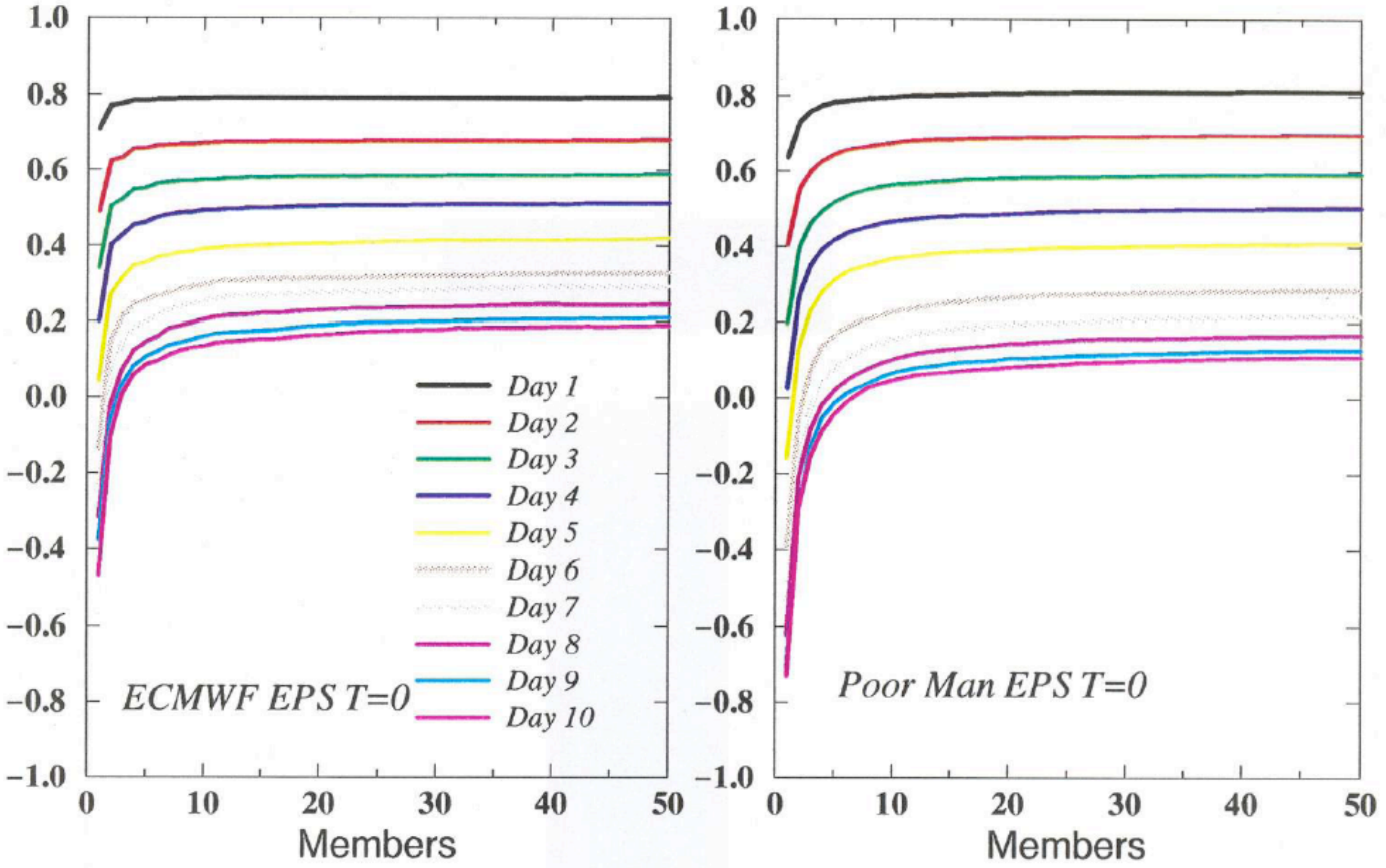
Conclusion in agreement with other studies (Anderson, *MWR*, 1997, Hamill *et al.*, *MWR*, 2000, Wang and Bishop, *JAS*, 2003, Bowler, *Tellus*, 2006).

On the other hand, Buizza (IUGG, Perugia, 2007) has presented results of comparisons made at ECMWF, in which the best results are obtained with SVs.

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 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)

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

Questions

- 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 ?

- 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 α $N/10$ 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 (α = 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 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

$$\begin{pmatrix} N+L \\ L \end{pmatrix}$$

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) ?

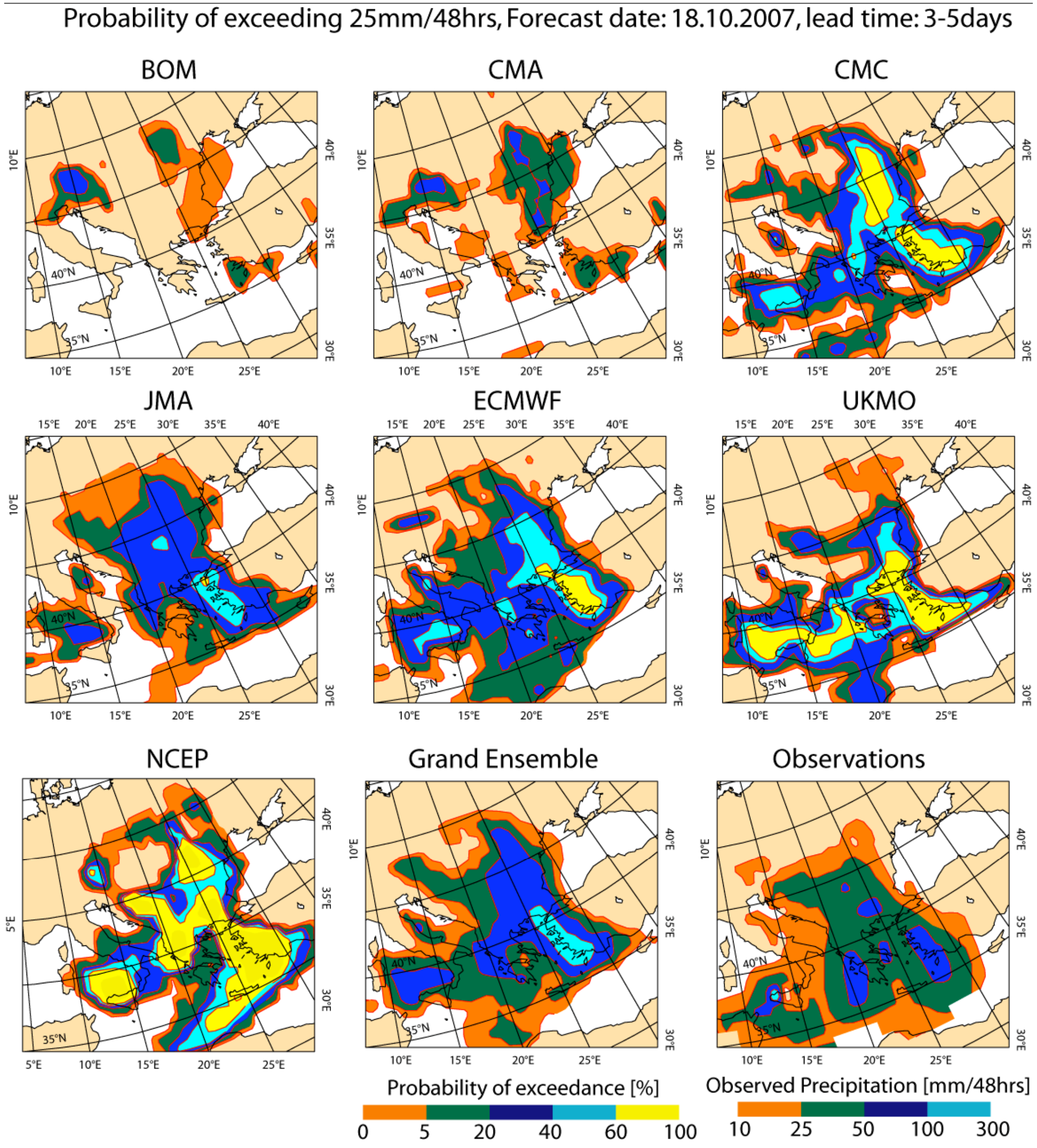
$21^{165} \approx 10^{218}$ possible probability distributions.

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.


Probability of exceeding 25mm/48hrs, Forecast date: 18.10.2007, lead time: 3-5days
BOM
CMA
CMC
JMA
ECMWF
UKMO
NCEP
Grand Ensemble
Observations
10°E
15°E
20°E
25°E
30°E
35°E
40°E
35°N
40°N
Probability of exceedance [%]
0
5
20
40
60
100
Observed Precipitation [mm/48hrs]
10
25
50
100
300

Pappenberger *et al*., 2008, *Geophys. Res. Lett.*

Conclusions

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.

Conclusions (2)

Present situation is somewhat hybrid, the predicted ensemble being a kind of auxiliary to a statistically more accurate higher resolution forecast. This is actually cause of confusion, when the high resolution forecast disagrees from a large subset of the ensemble.

How to include the high resolution forecast into the definition of the predicted probability distribution ?

Must we aim at a situation where the predicted object will be a probability distribution ?