Verification of Probabilistic Forecasts

Olivier Talagrand

Summer School
Advanced Mathematical Methods
to Study Atmospheric Dynamical Processes and Predictability

Banff International Research Station
for Mathematical Innovation and Discovery

Banff, Canada
14 July 2011

With acknowledgments to F. Atger, G. Candille and L. Descamps

With acknowledgments too to participants in Interest Group 5 of THORPEX Working Group on *Predictability and Dynamical Processes*.

As far as we can can tell, there will always be significant uncertainty on the future state of the atmosphere. How does that uncertainty evolve in time ? For instance, how fast does it increase ?

Deterministic dynamical system. State vector $\boldsymbol{x} = (x_1, x_2, \ldots, x_N)^{\mathrm{T}}$. Evolves in time according to equation

$$\frac{dx}{dt} = F(x,t)$$

or, componentwise

$$\frac{dx_i}{dt} = F_i(x,t) \qquad\qquad i=1,2,...,N$$

Probability Density Function (*PDF*) $p(\boldsymbol{x},t)$ for state vector. Evolves in time according to equation

$$\frac{\partial p}{\partial t} + div(pF) = 0$$

or

$$\frac{\partial p}{\partial t} + \sum_i \frac{\partial (pF_i)}{\partial x_i} = 0$$

which expresses conservation of probability in the flow F. It is fundamentally the same equation as the 'continuity' equation, which expresses conservation of mass in physical motion. It is called in the present context the *Liouville equation*.

If evolution is discretized in time, *viz.*,

$$\boldsymbol{x}^{k+1} = G_k(\boldsymbol{x}^k)$$

where k is a time index, then PDF evolves according to

$$p(\boldsymbol{x}^{k+1}, k+1) = det(DG_k(\boldsymbol{x}^k))\ p(\boldsymbol{x}^k, k)$$

where DG_k is the Jacobian of the mapping G_k.

If basic evolution equation contains a stochastic term, *viz*.,

$$\frac{dx}{dt} = F(x,t) + \eta(x,t)$$

where the noise $\eta(\boldsymbol{x},t)$ is random, unbiased, white in time, with covariance matrix

$$Q_{ij}(\boldsymbol{x},t) \equiv E\,[\eta_i(\boldsymbol{x},t)\ \eta_j(\boldsymbol{x},t)]$$

the PDF evolves according to the so-called *Fokker-Planck equation*,

$$\frac{\partial p}{\partial t} + \sum_i \frac{\partial(pF_i)}{\partial x_i} = \frac{1}{2}\sum_{i,j}\frac{\partial^2}{\partial x_i \partial x_j}[pQ_{ij}]$$

(aka the *second Kolmogorov equation* in the Russian literature)

In case the noise is not white in time, or is not additive, no simple equation describes the temporal evolution of the PDF, but that evolution is unambiguously defined.

Ensemble Prediction, in which one (or several) numerical models for the evolution of the flow are integrated for different initial or lateral boundary conditions and noise realizations, provides an affordable approximation for approximately solving the Liouville or Fokker-Planck equation for the PDF of the state of the atmospheric flow. Temporal dependence in the noise can easily be introduced in EPSs, provided it is explicitly quantified.

Operational Ensemble Meteorological Prediction was initiated in 1992 by the *National Centers for Environmental Prediction* (NCEP, USA) and the *European Centre for Medium-range Weather Forecasts* (ECMWF).

As of now, 10 meteorological centres are producing daily global operational ensemble predictions. Most of these predictions (about 200) are stored on the *THORPEX Interactive Grand Global Ensemble* (TIGGE) database, accessible at the address http://tigge.ecmwf.int/. In addition, a number of other centres are running regional Ensemble Prediction Systems (EPSs).

Three systems (those of ECMWF, of the Meteorological Service of Canada and of the UK Meteorological Office) include perturbations intended at representing the effects of model errors. The other systems evolve all ensemble elements with the same deterministic model (this may not be true any more as of July 2011).

Simple example (D. Richardson)

If temperature goes below freezing, road traffic is disrupted. Cost L.

Preventive action (gritting the road) is possible at cost C ($< L$).

Forecast is uncertain. Is is appropriate to take preventive action, or not ?

Two options

- Take no action. If freezing occurs with frequency p, long term expected loss pL

- Take action. Cost C

Conclusion. Take action iff anticipated probability of occurrence of freezing

$$p > C/L$$

That however requires that

- the anticipated probability of occurrence is *reliable*, in the precise sense that freezing occurs with frequency p in the circumstances when it is anticipated to occur with probability p.

- the conditions are such that the time required for achieving the expected minimization of loss is in a sense short enough (shorter for instance than the time over which the prediction system will significantly evolve).

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).

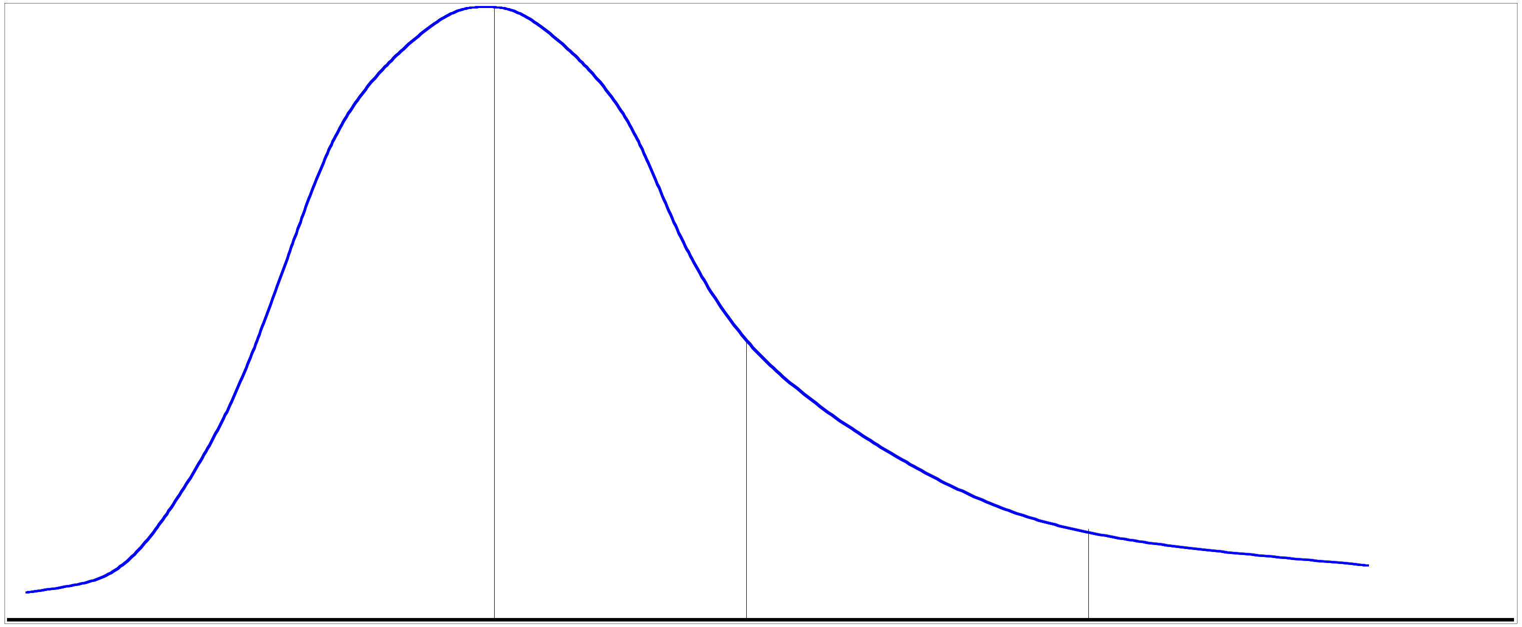
Point of view taken here

Purpose of probabilistic prediction is to describe our uncertainty on the future state of the atmosphere

Question

How is it possible to objectively (and, if possible, quantitatively) evaluate whether that purpose has been achieved ? In particular, how is it possible to objectively compare the performance of two different probabilistic prediction methods ?

In the following, we discuss evaluation of ensemble prediction systems. But most of the methods that will be presented could be used to evaluate any system for probabilistic prediction. And all can also be used for evaluation of ensemble assimilation systems, such as EnKF.



Difficulties

o The predicted object (a probability distribution) and the observed object (a point observation) are not of the same nature. How can they be compared ?

o The predicted object is not better known afterwards than it was beforehand.

o The predicted object, which is meant to describe our uncertainty on the future state of the atmosphere, has actually no objective existence.

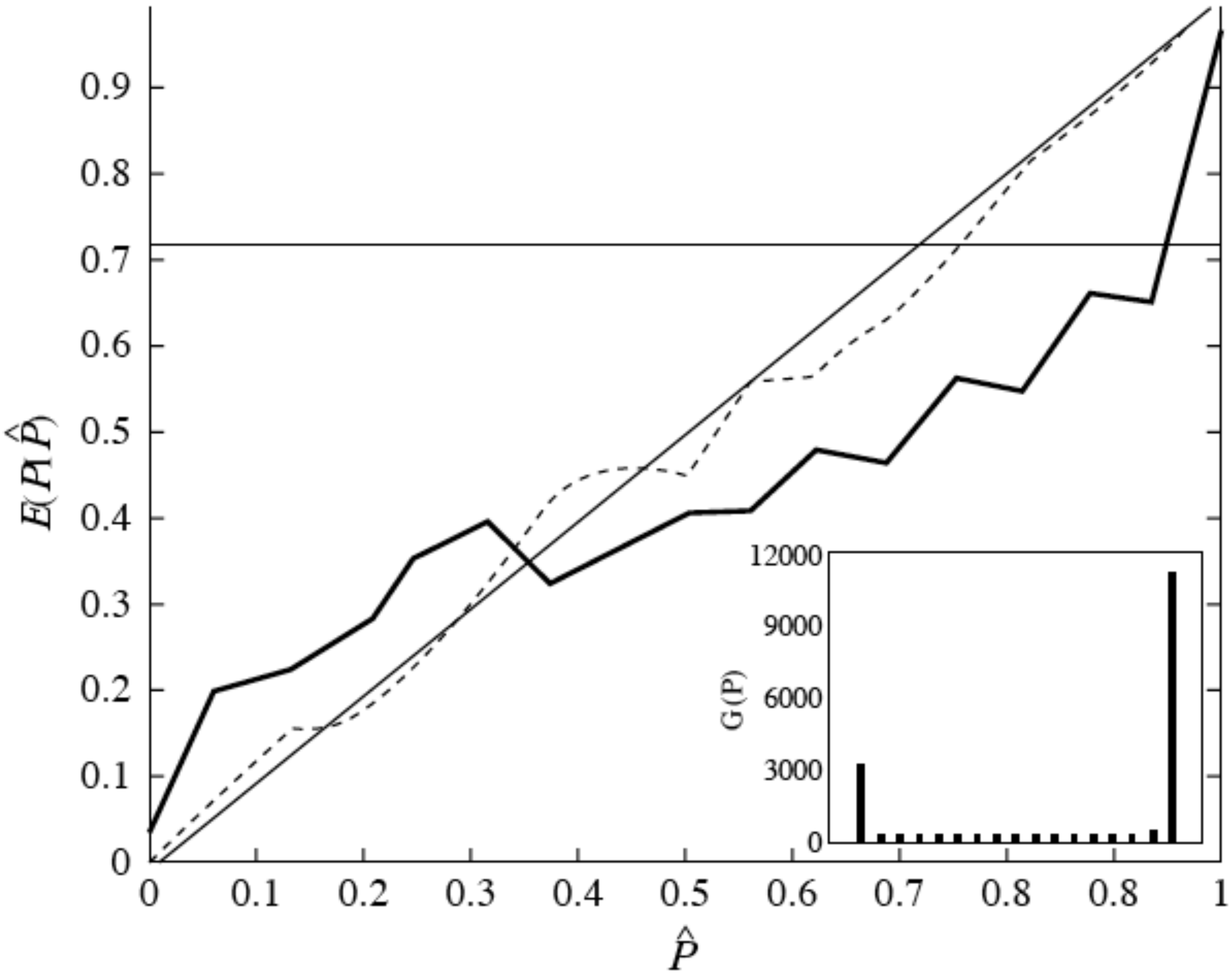
As a consequence, validation of ensemble prediction can only be statistical.

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

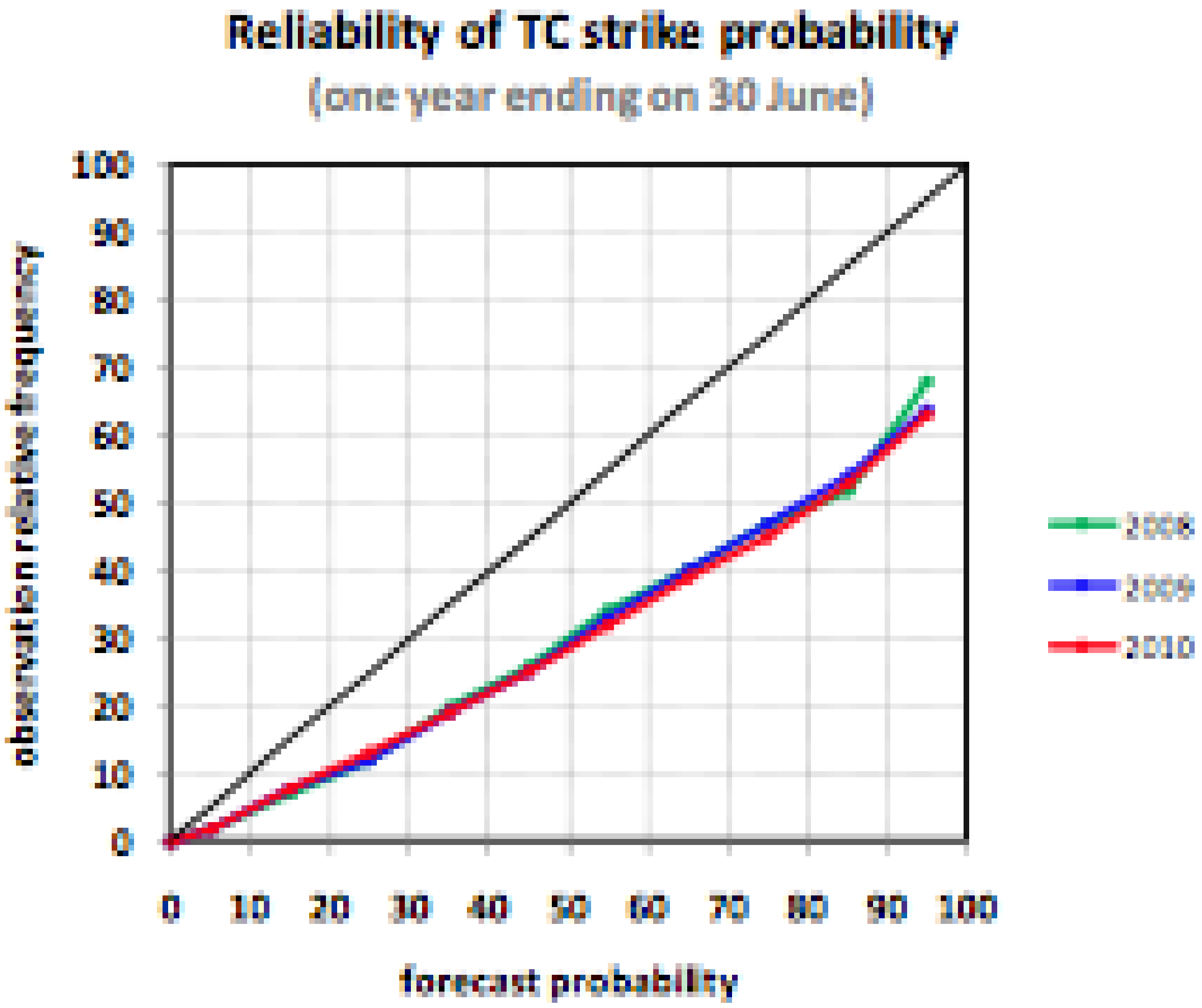
- ***Reliability***

  (*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



Reliability of TC strike probability

(one year ending on 30 June)

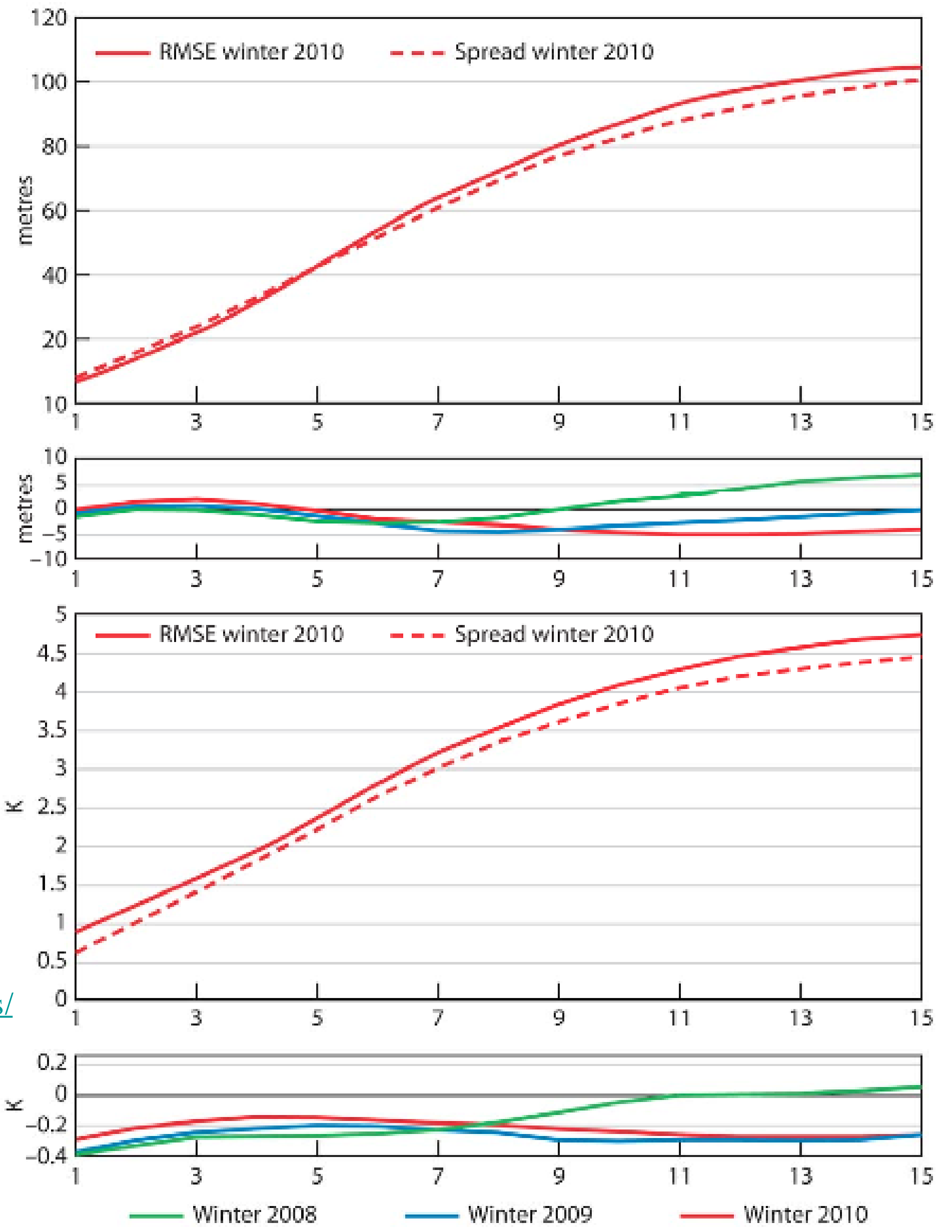
http://www.ecmwf.int/publications/library/ecpublications/_pdf/tm/601-700/tm635.pdf

More generally

- Consider a probability law 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 Random Variable, …



http://www.ecmwf.int/publications/
library/ecpublications/_pdf/tm/
601-700/tm635.pdf

Figure 8: Ensemble spread (standard deviation, dashed lines) and root mean square error of ensemble-mean (solid lines) for winter 2009-2010(upper figure in each panel), complemented with differences of ensemble spread and root mean square error of ensemble-mean for last 3 winter seasons (lower figure in each panel, negative values indicate spread is too small); plots are for 500 hPa geopotential (top) and 850 hPa temperature (bottom) over the extra-tropical northern hemisphere for forecast days 1 to 15.

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

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

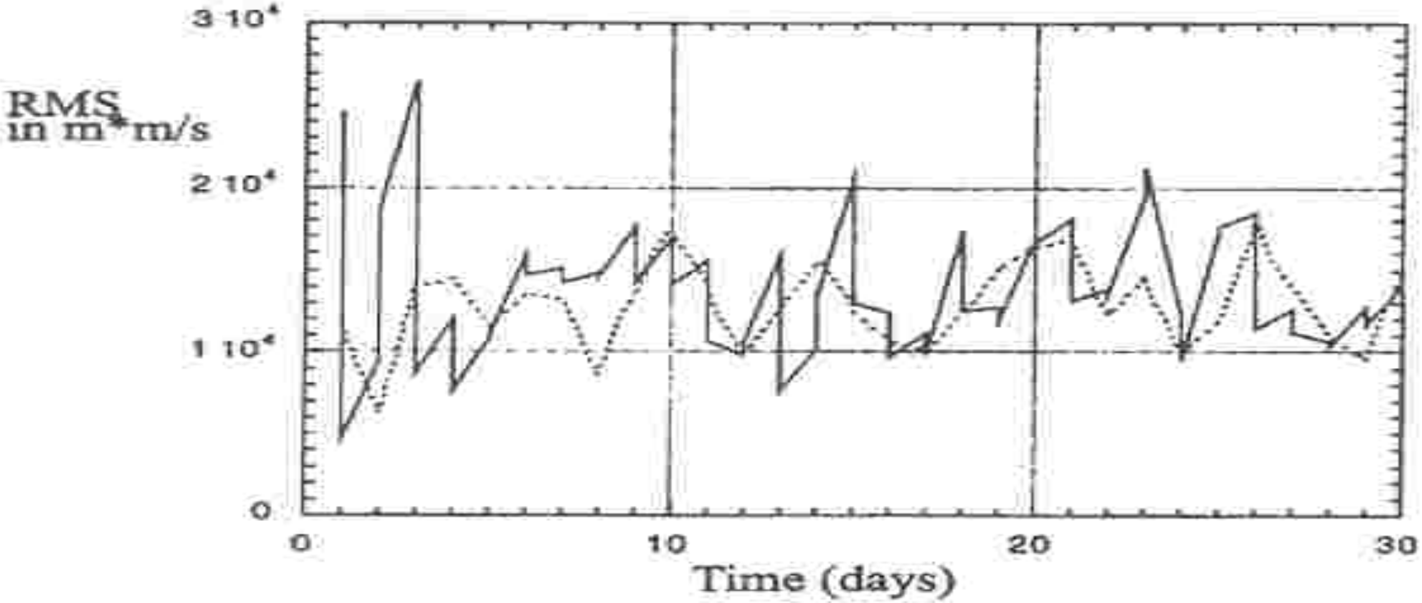


FIG. 12. Comparison of rms error ($m^2 s^{-1}$) between ensemble mean and independent observations (dotted line) and the std dev in the ensemble (solid line). The excellent agreement shows that the SIRF is working correctly.

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

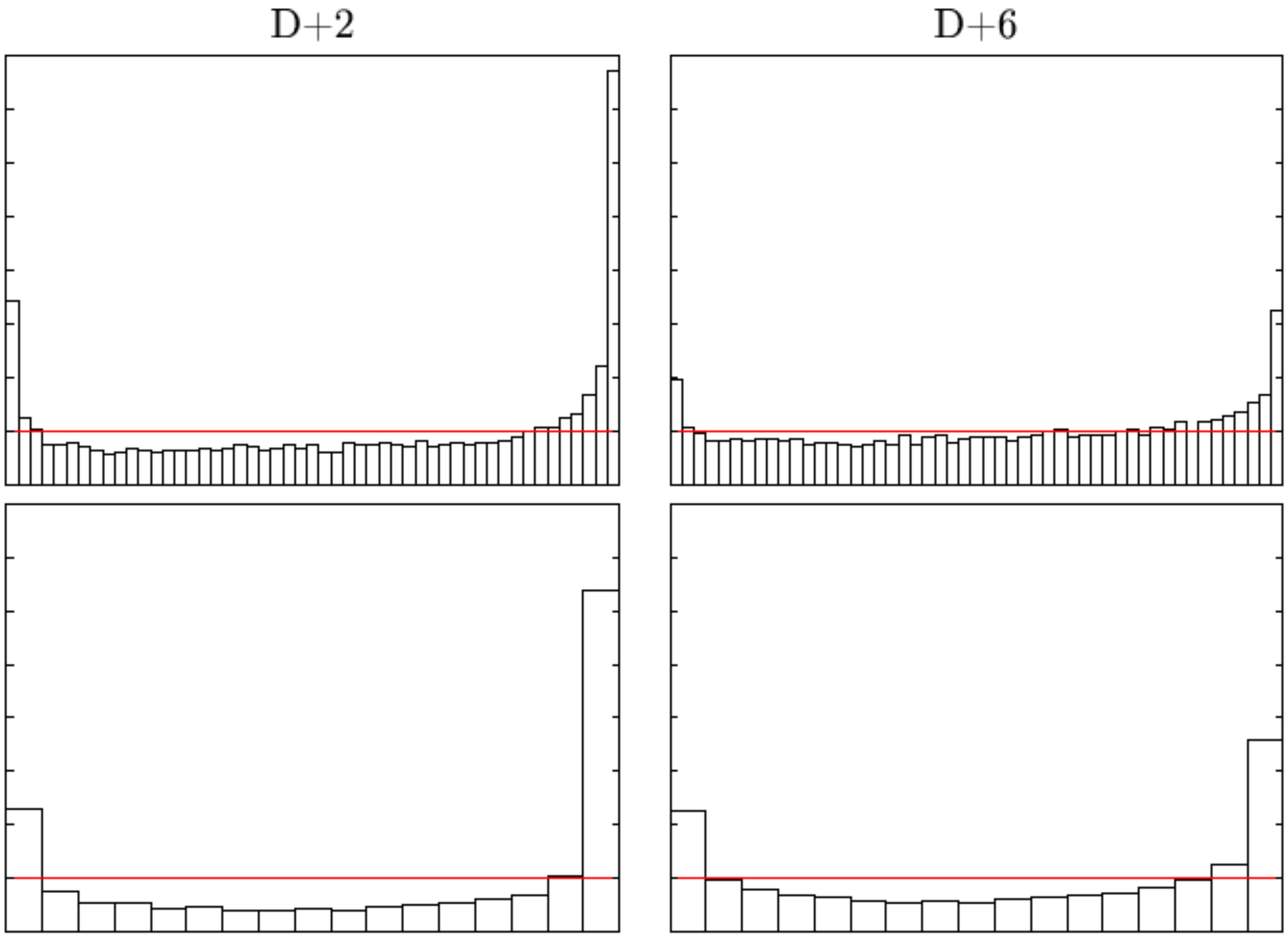
Rank Histograms

For some scalar variable x, N ensemble values, assumed to be N independent realizations of the same probability distribution, ranked in increasing order

$$x_1 < x_2 < \ldots < x_N$$

Define $N+1$ intervals.

If verifying observation ξ is an $N+1$st independent realization of the same probability distribution, it must be statistically undistinguishable from the x_i‘s. In particular, must be uniformly distributed among the $N+1$ intervals defined by the x_i‘s.



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

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

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. Resolution measures real intrinsic value of prediction system, *i. e.*, what remains when system has been made reliable by *a posteriori* calibration.

  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 validation sample is available, each of these qualities can be objectively and quantitatively measured by a number of different, not exactly equivalent, scores.

Brier Score (Brier, 1950), relative to binary event $\mathcal{E}$

$$\mathcal{B} \equiv E[(p - p_o)^2]$$

where p is predicted probability of occurrence, $p_o = 1$ or 0 depending on whether $\mathcal{E}$ has been observed to occur or not, and E denotes average over all realizations of the prediction system.

Decomposes into

$$\mathcal{B} = E[(p-p')^2] \ - \ E[(p'-p_c)^2] \ + \ p_c(1-p_c)$$

where $p_c \equiv E(p_o) = E(p')$ is observed frequency of occurrence of $\mathcal{E}$.

First term $E[(p-p')^2]$ measures reliability.

Second term $E[(p'-p_c)^2]$ measures dispersion of *a posteriori* calibrated probabilities p'. The larger that dispersion, the more discriminating, or resolving, and the more useful, the prediction system. That term measures the *resolution* of the system.

Third term, called *uncertainty* term, depends only on event $\mathcal{E}$, not on performance of prediction system.

Remark. All the above remains valid if p_o takes values different from 0 or 1.

Brier Skill Score

A system which always predicts climatological frequency of occurrence p_c (fully reliable, but no resolution) has Brier score $p_c(1-p_c)$

$$\mathcal{B}_{SS} \equiv 1 - \mathcal{B}/p_c(1-p_c)$$

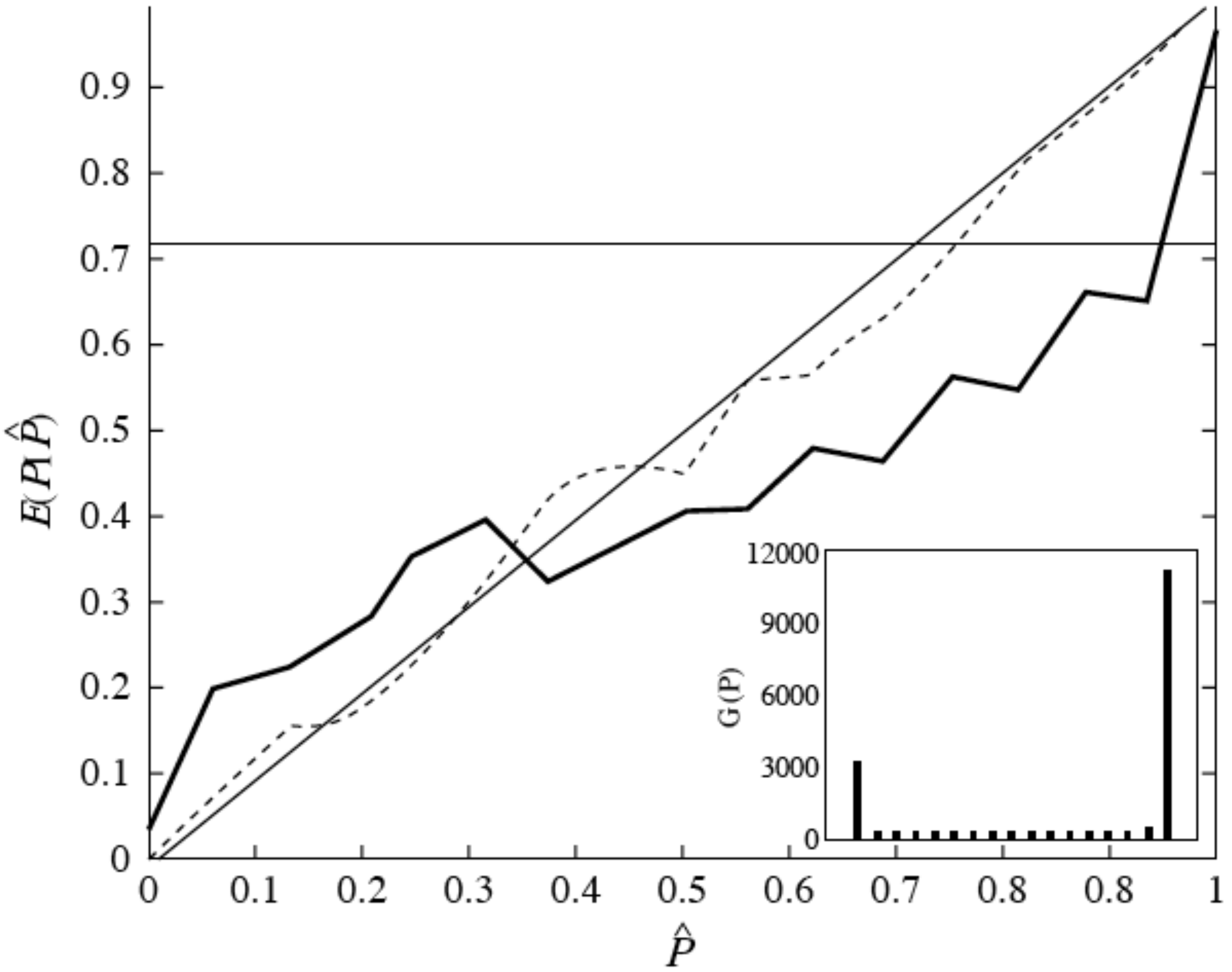
(positively oriented)

and components

$$\mathcal{B}_{rel} \equiv E[(p-p')^2]/p_c(1-p_c)$$

$$\mathcal{B}_{res} \equiv 1 - E[(p'-p_c)^2]\,/p_c(1-p_c)$$

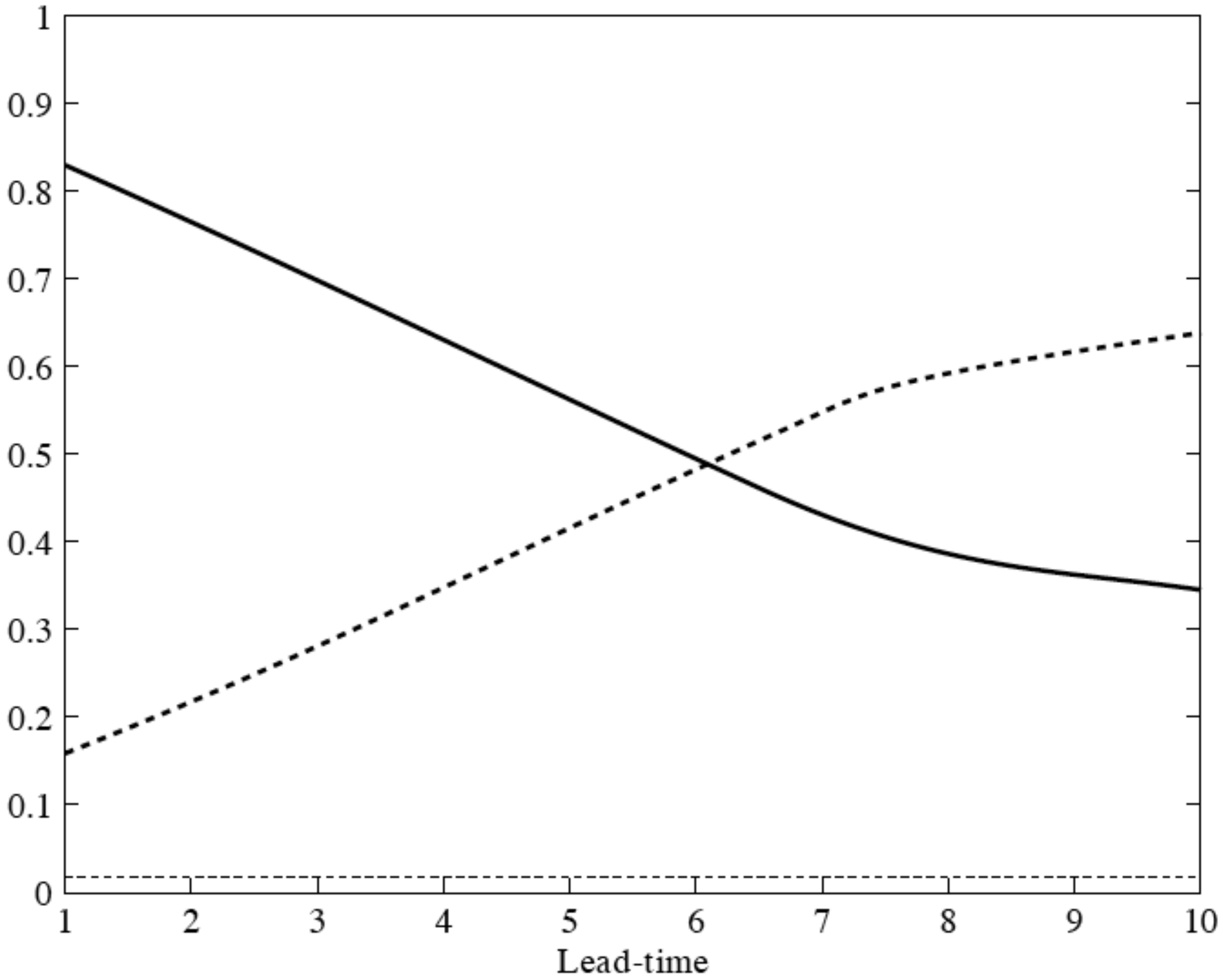
(negatively oriented)



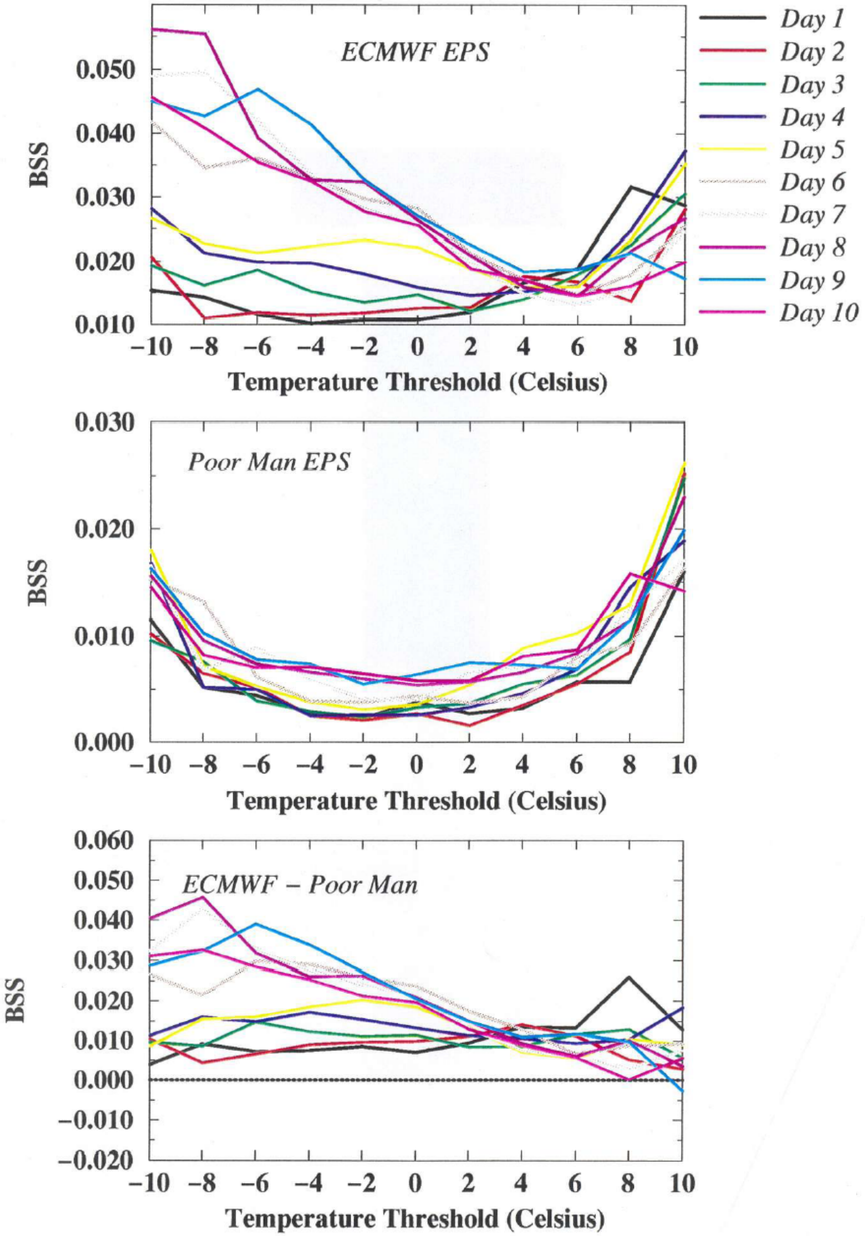
Brier Skill Score (positively oriented) = 0.677

Reliability (negatively oriented) = 0.027

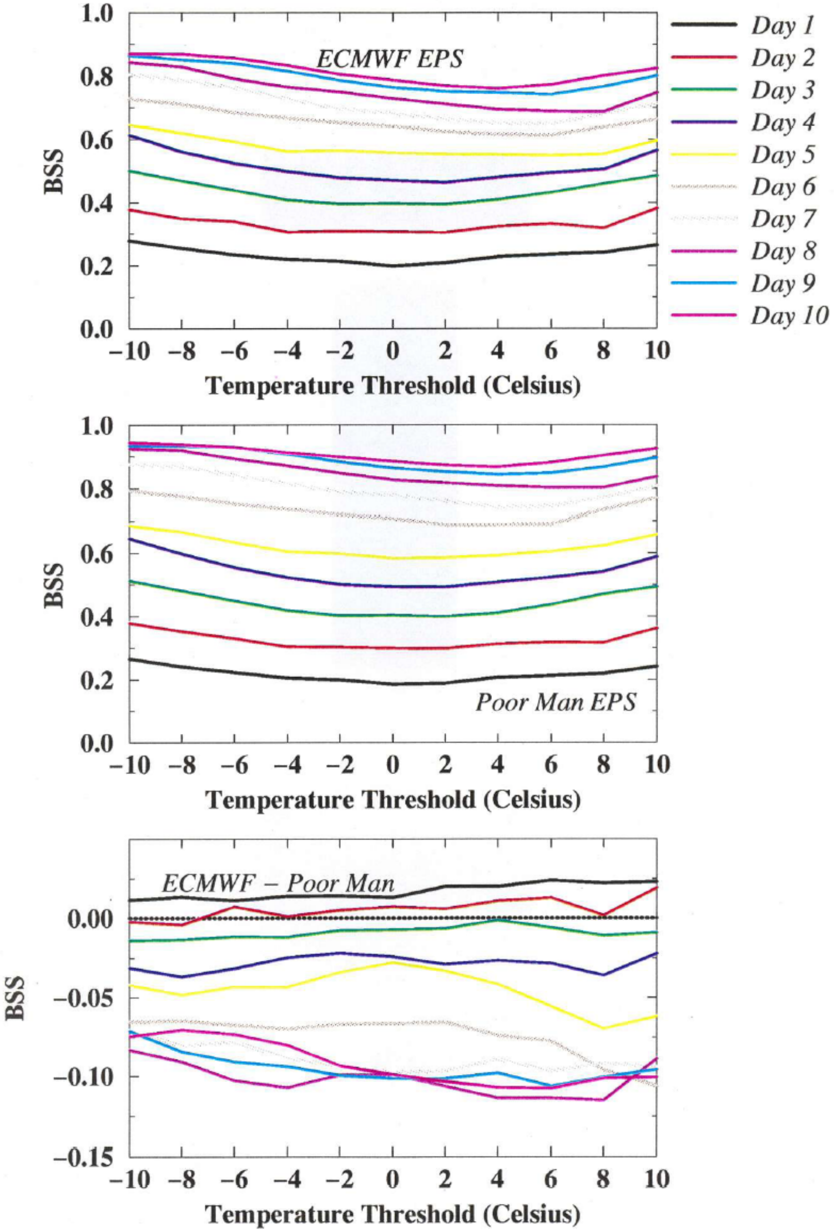
Resolution (negatively oriented) = 0.296



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



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



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

Properness of Score

Forecaster whose performance is evaluated by Brier score. Class of situations $\mathcal{C}$ in which, to the best of forecaster's knowledge, event $\mathcal{E}$ is going to occur with frequency p‘ ($E_{\mathcal{C}}(o) = p$‘). What must forecaster predict in those situations ?

Assume forecaster predicts probability p

$$(p - o)^2 = (p\text{-}p')^2 + 2\,(p\text{-}p')(p'\text{-}o) + (p'\text{-}o)^2$$

If forecaster is correct in his belief, middle term on right hand side will cancel on taking conditional expectation $E_{\mathcal{C}}$. There will remain

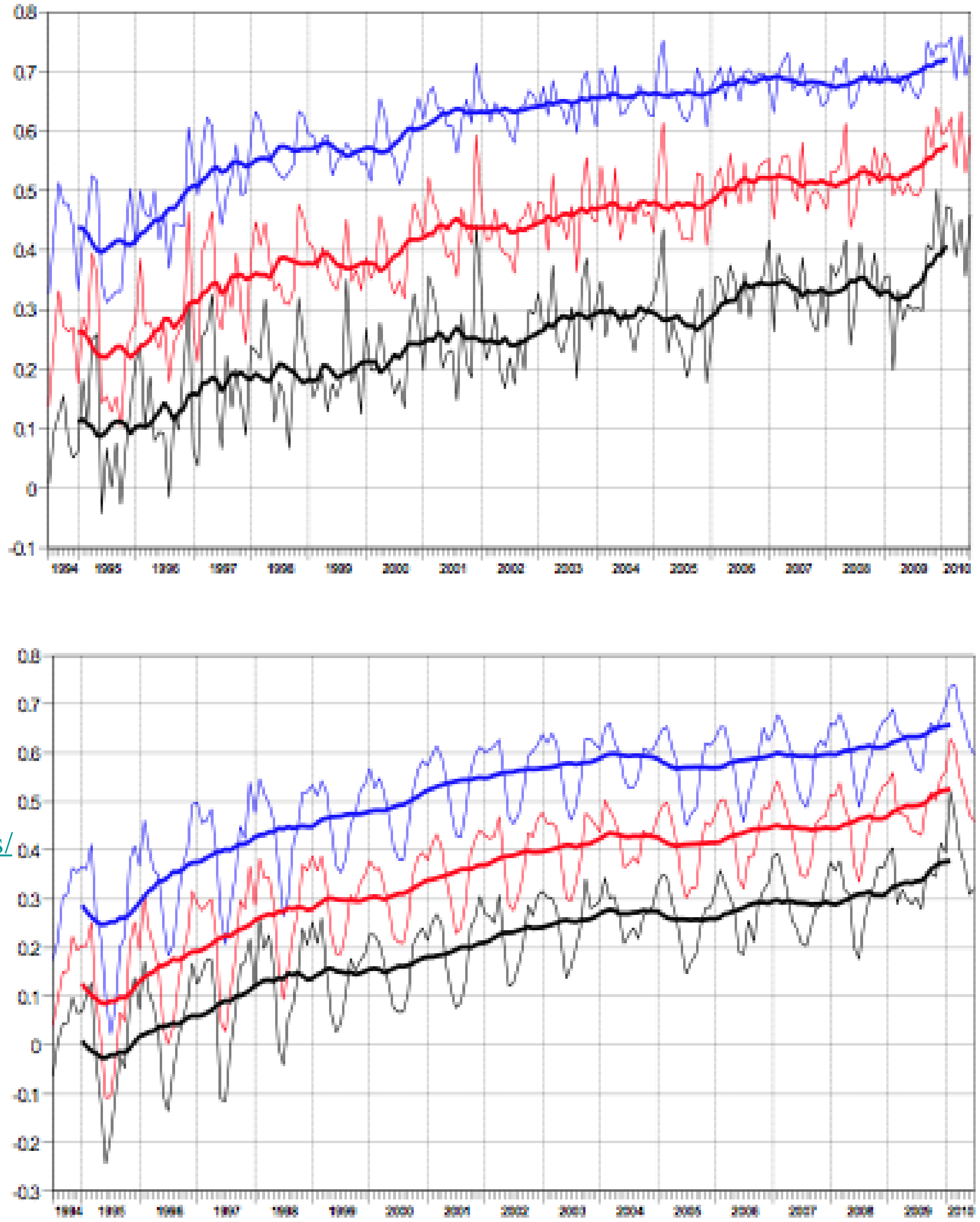
$$E_{\mathcal{C}}[(p - o)^2] = (p\text{-}p')^2 + E_{\mathcal{C}}[(p'\text{-}o)^2]$$

Second term on right hand side is independent of p, while first one vanishes for $p=p'$. The objective interest of forecaster is to be honest, and to predict what is to his best knowledge the probability of occurrence of $\mathcal{E}$. Brier score is *proper* (see also papers by J. Bröcker).

Ranked Probability Score (RPS)

Sum of Brier scores for events $x > \xi_k$ for a number of prespecified thresholds ξ_k. Introduces proximity to the thresholds in the score.

Continuous Ranked Probability Score (CRPS). The same, replacing sum over a finite number of thresholds by an integral over all doamin of variation of variable under consideration (measure wrt which integral is taken matters).



http://www.ecmwf.int/publications/library/ecpublications/_pdf/tm/601-700/tm635.pdf

Figure 7: Monthly score and 12-month running mean (bold) of Ranked Probability Skill Score for EPS forecasts of 850 hPa temperature at day 3 (blue), 5 (red) and 7 (black) for Europe (top) and the northern hemisphere extratropics (bottom).

http://www.ecmwf.int/publications/
library/ecpublications/_pdf/tm/
601-700/tm635.pdf

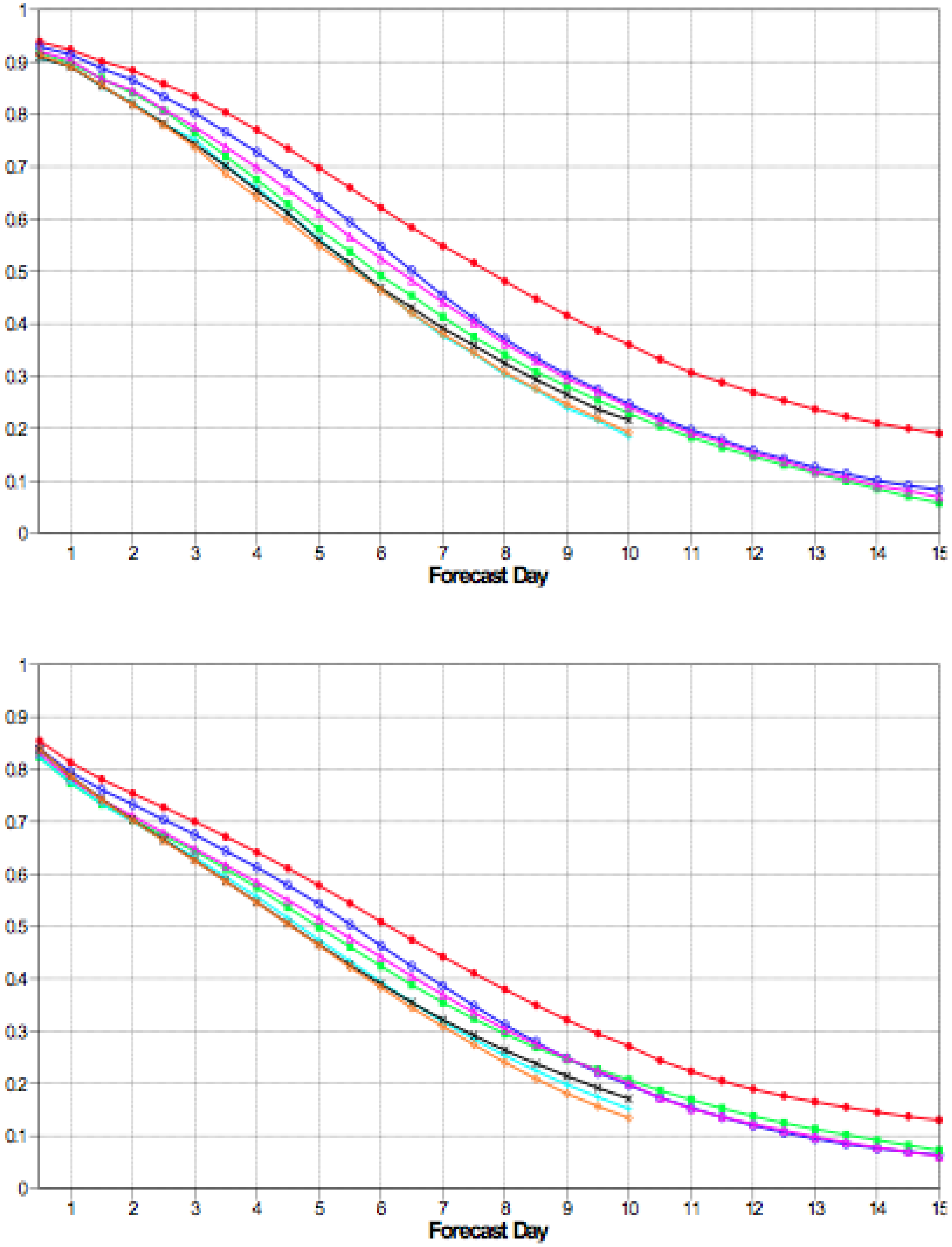


Figure 9: Ranked probability skill score for 500 hPa height (top) and 850 hPa temperature (bottom) EPS forecasts for winter (December-February) over the extra-tropical northern hemisphere. Skill from the EPS day 1-15 forecasts is shown for winter 2009-10 (red), 2008-09 (blue), 2007-08 (green) and 2006-07 (magenta). The EPS only ran to 10 days in previous years: 2005-06 (cyan), 2004-05 (black), 2003-04 (orange).

Generalized Brier Scores

Set of binary events $\mathcal{E}_i$ (may overlap or not, may be exhaustive or not). Any score of the form

$$\mathcal{B} \equiv \Sigma_i \ \alpha_i E[(p_i\text{-}o_i)^2] = \Sigma_i \ \alpha_i \ \mathcal{B}_i$$

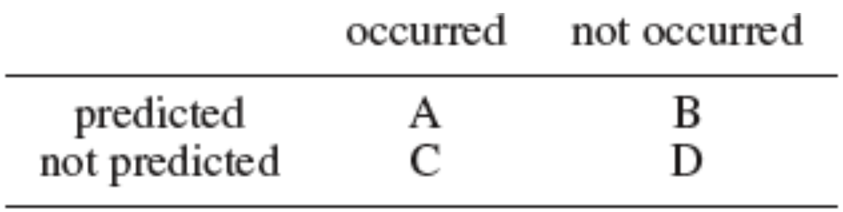
where the α_i's are positive, and $\mathcal{B}_i$ is the Brier score for event $\mathcal{E}_i$. Examples : *Ranked Probability Score* (*RPS*), *Continuous Ranked Probability Score* (*CRPS*). These are respectively a finite sum and an infinite integral, over thresholds ξ, of Brier scores for the events of the form $x > \xi$ where x is scalar variable.

Have a reliability-resolution decomposition (actually several; see Hersbach, 2000, *Wea. Forecasting*, and Candille and Talagrand, 2005, *QJRMS*), and are proper.

The only scores that are defined as mean of an observation-minus-prediction difference, and are proper ?

Relative Operating Characteristics

Binary event $\mathcal{E}$. Contingency table



	occurred	not occurred
predicted	A	B
not predicted	C	D

‘*Hits*’

$$H \equiv A/(A + C)$$

a posteriori conditional probability that $\mathcal{E}$ had been predicted to occur, given it has occurred

‘*False alarms*’

$$F \equiv B/(B + D)$$

a posteriori conditional probability that $\mathcal{E}$ had been predicted to occur, given it has not occurred.

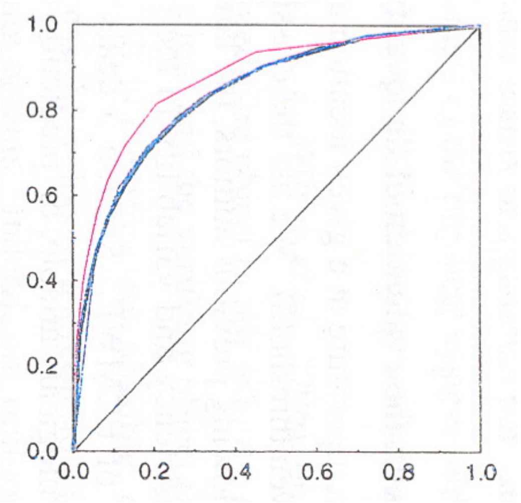
For an accurate system, H must be close to 1, and F to 0 .

Relative Operating Characteristics curve

Shows variations of $H(s)$ as a function of $F(s)$, where, for each threshold $s, 0 < s < 1$

- $H(s)$ is *a posteriori* conditional probability that predicted probability p was larger than s, given that $\mathcal{E}$ has occurred

- $F(s)$ is *a posteriori* conditional probability that predicted probability p was larger than s, given that $\mathcal{E}$ has not occurred



ROC curve is globally invariant in *a posteriori* calibration $p \rightarrow p'$. Area below curve is a measure of rersolution.

http://www.ecmwf.int/publications/
library/ecpublications/_pdf/tm/
601-700/tm635.pdf

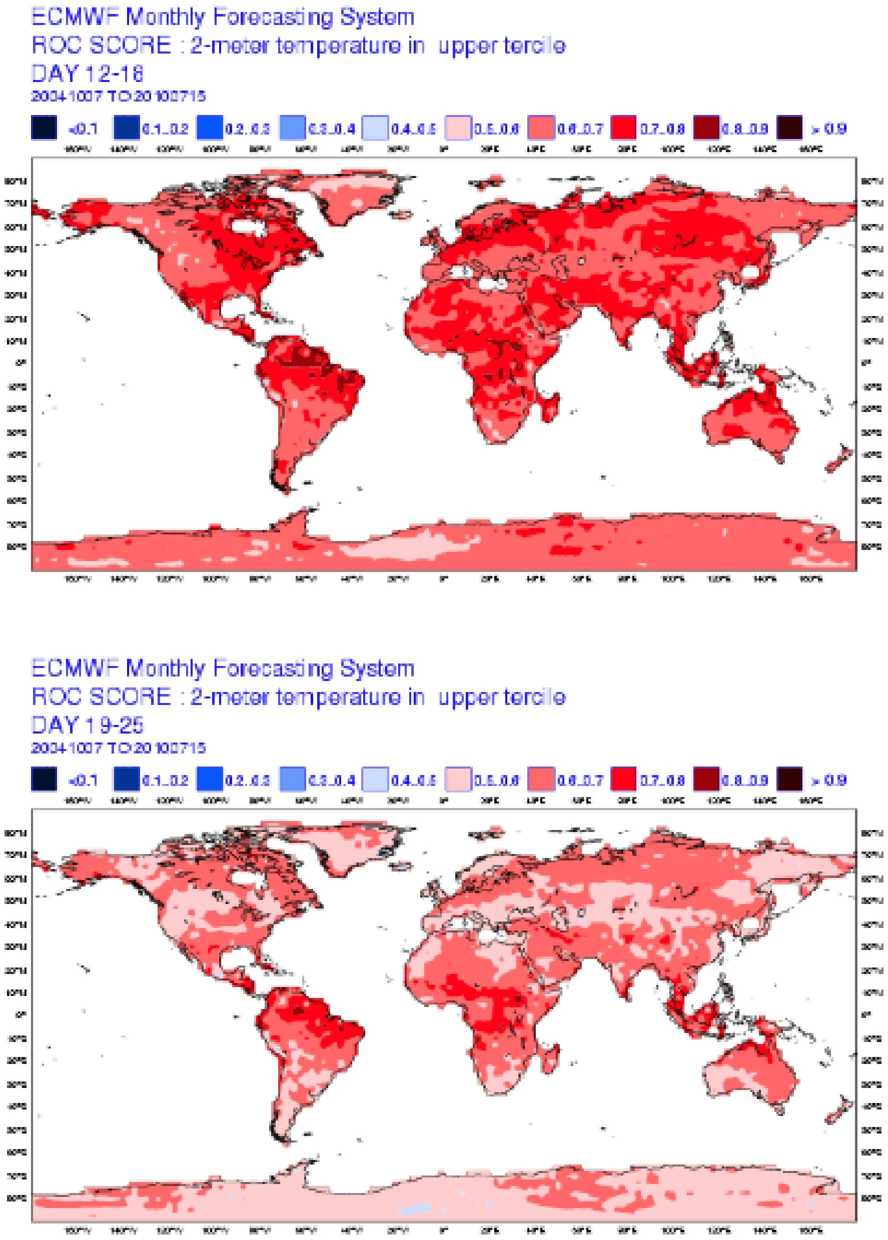


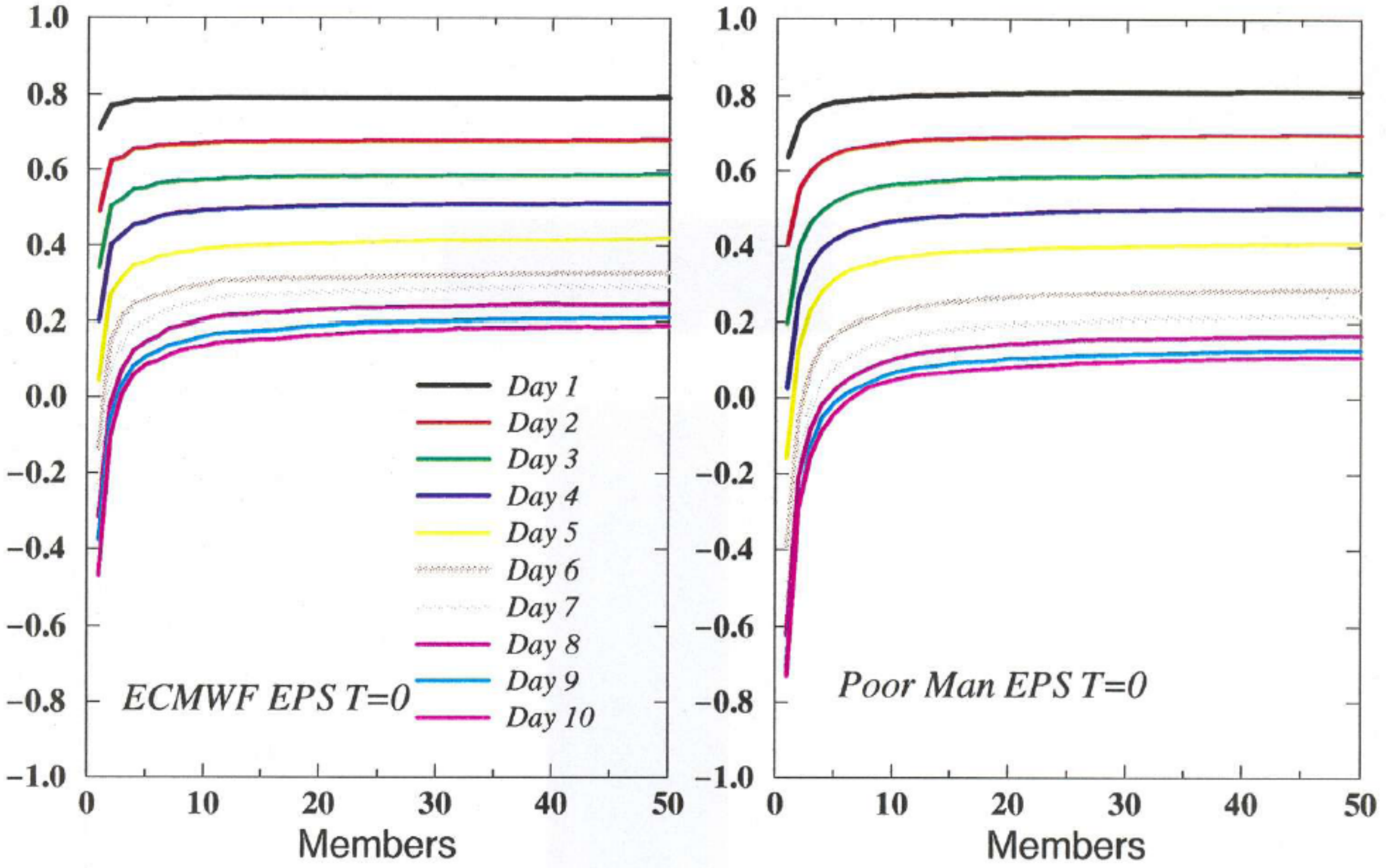
Figure 29: Monthly forecast verification. Spatial distribution of ROC area scores for the probability of 2 m temperature anomalies being in the upper third of the climatological distribution. The sample comprises all forecasts issued between 7 October 2004 and 15 July 2010 for two 7-day forecast ranges: days 12-18 (top) and days 19-25 (bottom). Red shading indicates positive skill compared to climate.

Statistical performance of a probabilistic prediction system is entirely determined by

- Relative frequency $g(F)$ with which each probability law F is predicted.

- Mapping $F \rightarrow F'(F)$

If system is reliable, then $F'(F) = F$ for any F, and statistical performance of system is entirely determined by relative frequency $g(F)$.



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

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

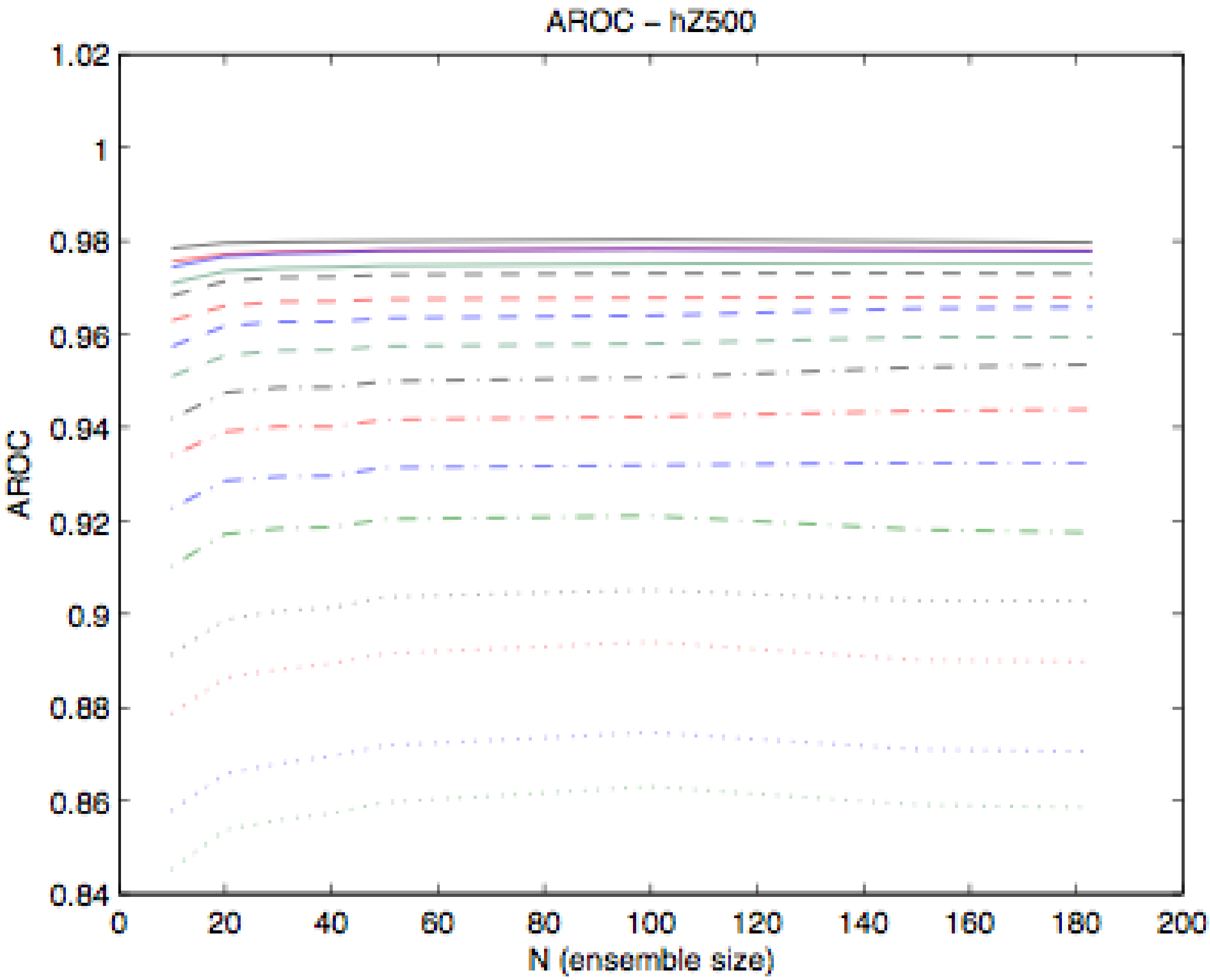
Theoretical estimate (raw Brier score)

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

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

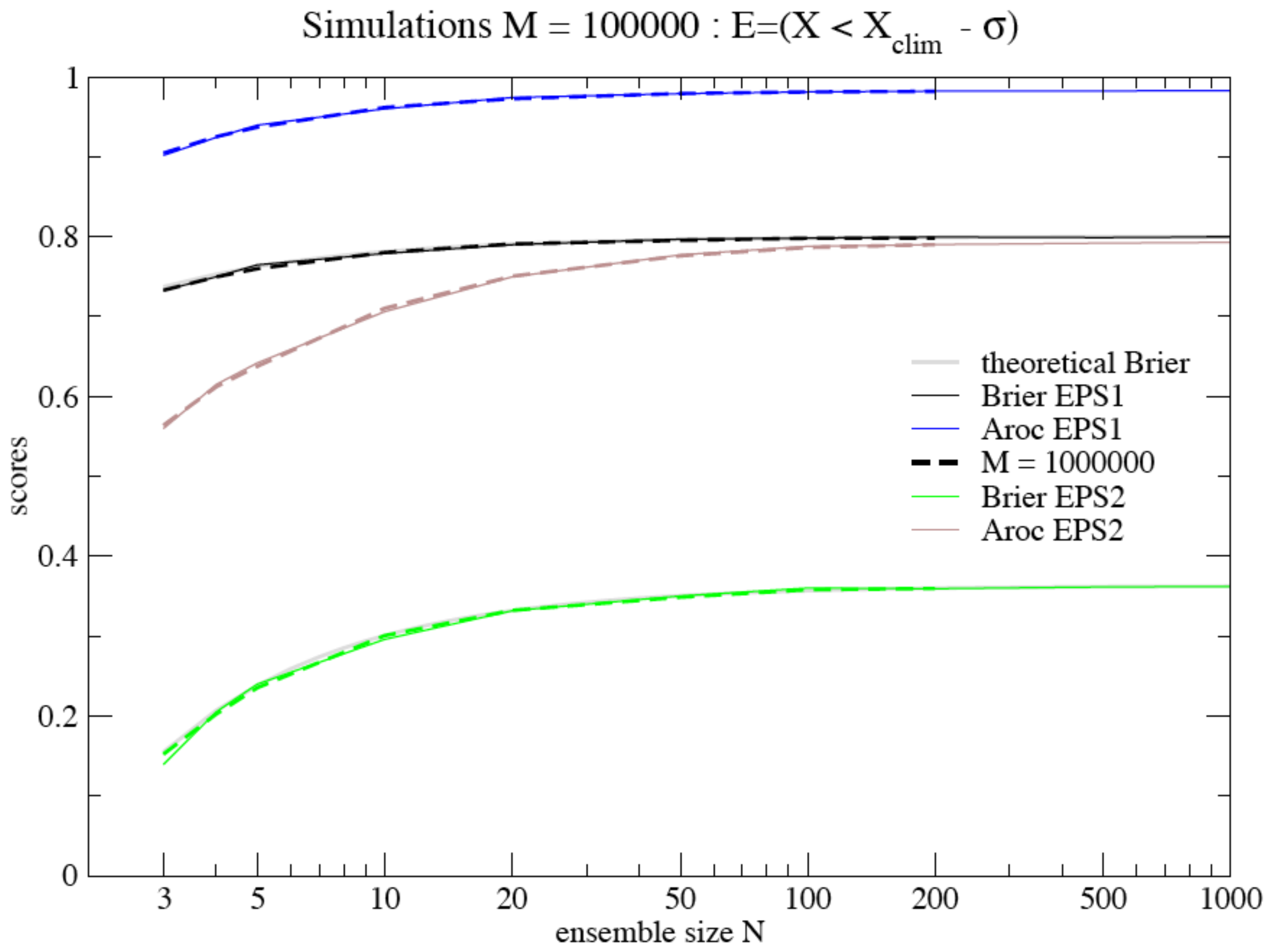


Figure 1: Impact of N on Brier Skill Score and ROC area

G. Candille, 2009

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. Known-to-be-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

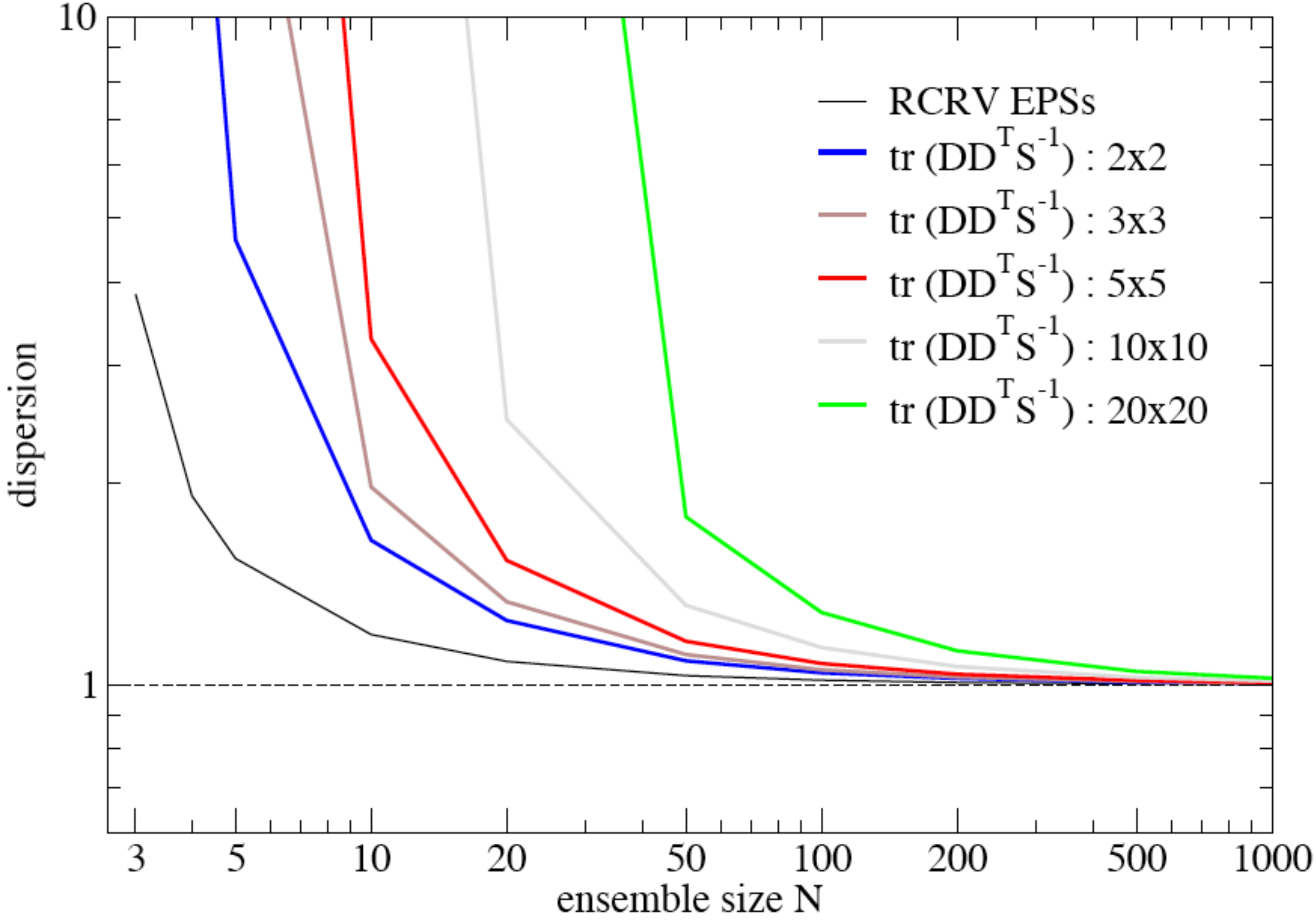
(i) Saturation is determined by the number of unstable modes in the system. Situation might be different with mesoscale ensemble prediction.

(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.



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.

Conclusion on ensemble size

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

Best Linear Unbiased Estimate

State vector x, belonging to *state space* $\mathcal{S}$ ($\dim\mathcal{S} = n$), to be estimated.
Available data in the form of

- A '*background*' estimate (*e. g.* forecast from the past), belonging to *state space*, with dimension n

  $$x^b = x + \zeta^b$$

- An additional set of data (*e. g.* observations), belonging to *observation space*, with dimension p

  $$y = Hx + \varepsilon$$

Assume $E(\zeta^b) = 0$, $E(\varepsilon) = 0$, $E(\zeta^b \varepsilon^{\mathrm{T}}) = 0$ (not mathematically restrictive)
and set $E(\zeta^b \zeta^{b\mathrm{T}}) \equiv P^b$ (also often denoted B), $E(\varepsilon\varepsilon^{\mathrm{T}}) \equiv R$

Best Linear Unbiased Estimate (continuation 1)

If error $(\zeta^{bT}, \varepsilon^{T})^{T}$ gaussian, then $P(\boldsymbol{x} \mid \boldsymbol{x}^b, \boldsymbol{y}) = \mathcal{N}[\boldsymbol{x}^a, \boldsymbol{P}^a]$. with

$$\boldsymbol{x}^a = \boldsymbol{x}^b + P^b H^{\mathrm{T}} [HP^bH^{\mathrm{T}} + R]^{-1} (\boldsymbol{y} - H\boldsymbol{x}^b)$$
$$P^a = P^b - P^b H^{\mathrm{T}} [HP^bH^{\mathrm{T}} + R]^{-1} HP^b$$

or equivalently

$$\boldsymbol{x}^a = \boldsymbol{x}^b + P^a H^{\mathrm{T}} R^{-1} (\boldsymbol{y} - H\boldsymbol{x}^b)$$
$$[P^a]^{-1} = [P^b]^{-1} + H^{\mathrm{T}} R^{-1}H$$

Best Linear Unbiased Estimate (continuation 2)

Lump $\boldsymbol{x}^b$ and $\boldsymbol{y}$ into

$$z = \Gamma x + \zeta$$

with $$z = \begin{pmatrix} x^b = x + \zeta^b \\ y = Hx + \varepsilon \end{pmatrix}$$

and $$\Gamma = \begin{pmatrix} I_n \\ H \end{pmatrix} \qquad \zeta = \begin{pmatrix} \zeta^b \\ \varepsilon \end{pmatrix}$$

Set $$S \equiv E(\zeta\zeta^{\mathrm{T}}) = \begin{pmatrix} P^b & 0 \\ 0 & R \end{pmatrix}$$

Then

$$\boldsymbol{x}^a = (\Gamma^\mathrm{T} S^{-1} \Gamma)^{-1} \Gamma^\mathrm{T} S^{-1} \boldsymbol{z}$$

$$P^a = (\Gamma^\mathrm{T} S^{-1} \Gamma)^{-1}$$

Conversely, if data of the form

$$z = \Gamma x + \xi, \qquad \xi \sim \mathcal{N}[0, S],$$

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

provided $rank \Gamma = n$ (the data vector z contains information on every component of the state vector $\boldsymbol{x}$)

This provides a ready recipe for obtaining a sample of conditional probability distribution $P(\boldsymbol{x} \mid \boldsymbol{z})$

- Perturb data vector according to the error probability distribution

$$\boldsymbol{z} \rightarrow \boldsymbol{z}' = \boldsymbol{z} + \zeta', \quad \zeta' \sim \mathcal{N}[0, S]$$

- Do analysis

$$x'^{a} = (\Gamma^{\mathrm{T}} S^{-1} \Gamma)^{-1} \Gamma^{\mathrm{T}} S^{-1} \boldsymbol{z}'$$

x'^{a} is distributed according to $P(\boldsymbol{x} \mid \boldsymbol{z}) = \mathcal{N}[\boldsymbol{x}^{a}, P^{a}]$

Question. To which extent does this result hold true in case of non-gaussianity and /or non-linearity ?

M. Jardak (2011)

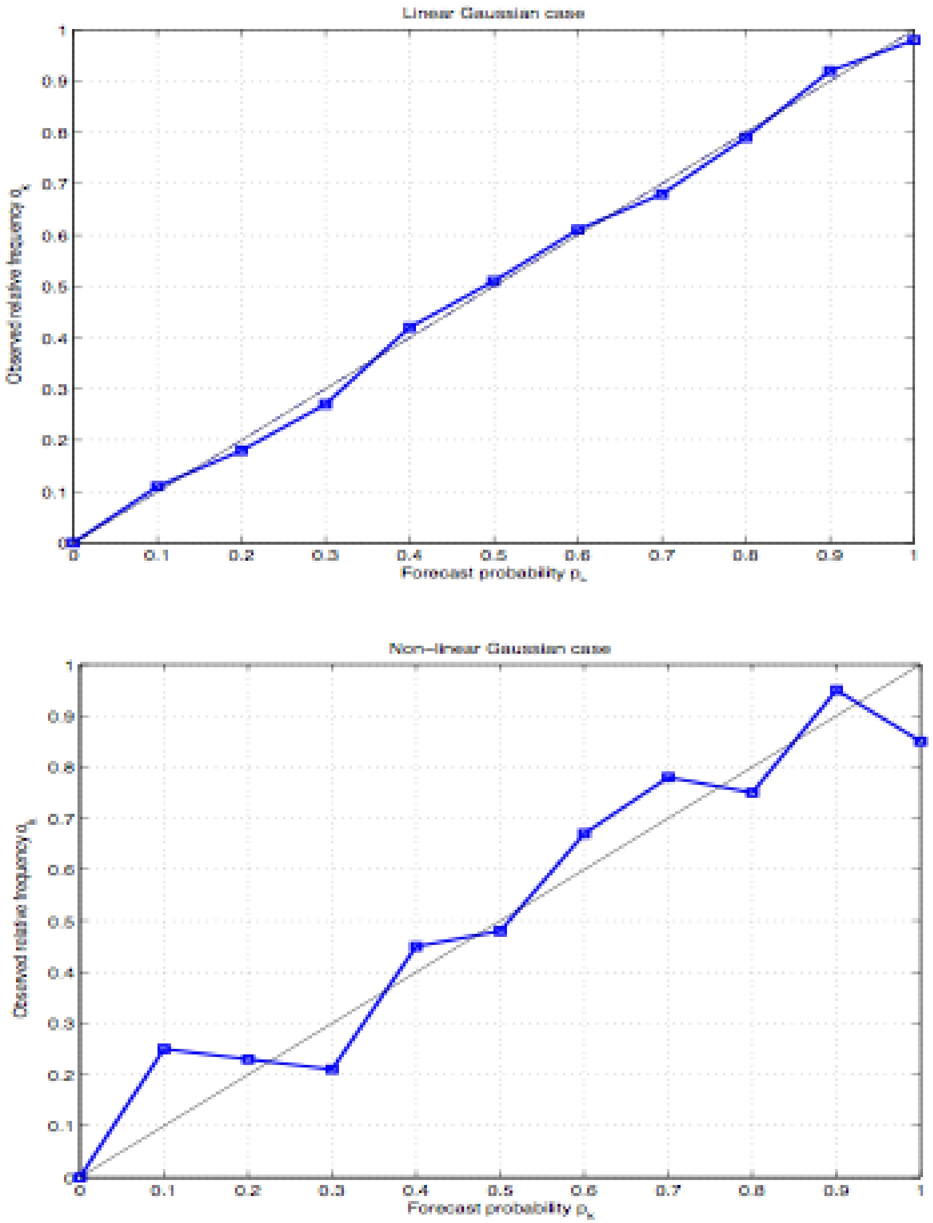
Kuramoto-Sivashinsky equation

$$u_t + uu_x + u_{xx} + u_{xxxx} = 0$$

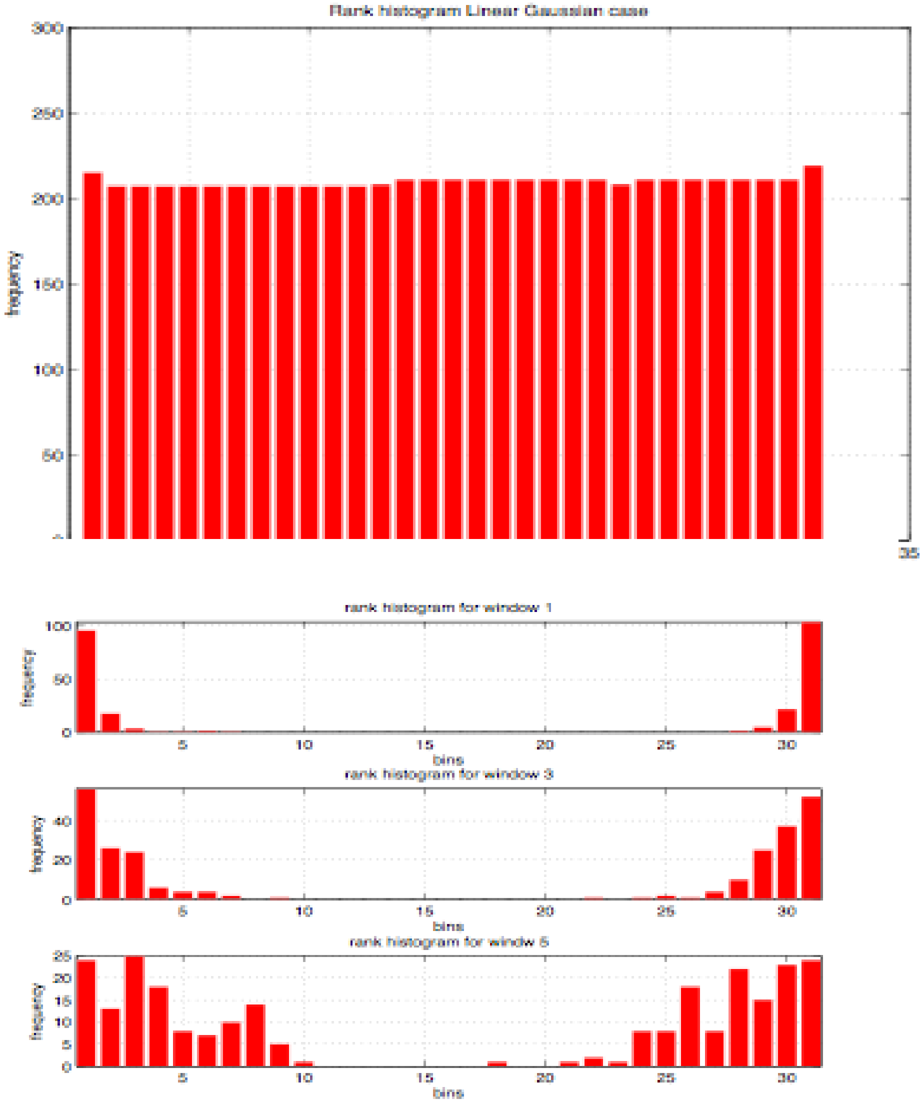
with periodic conditions in x.

Ensemble variational assimilation has been implemented in linearized $i.\ e.$, equation linearized in the vicinity of one nonlinear solution) and nonlinear cases.

There is no test of bayesianity (and there cannot be). But bayesianity implies reliability, as defined above, and non-reliability therefore implies non-bayesianity.

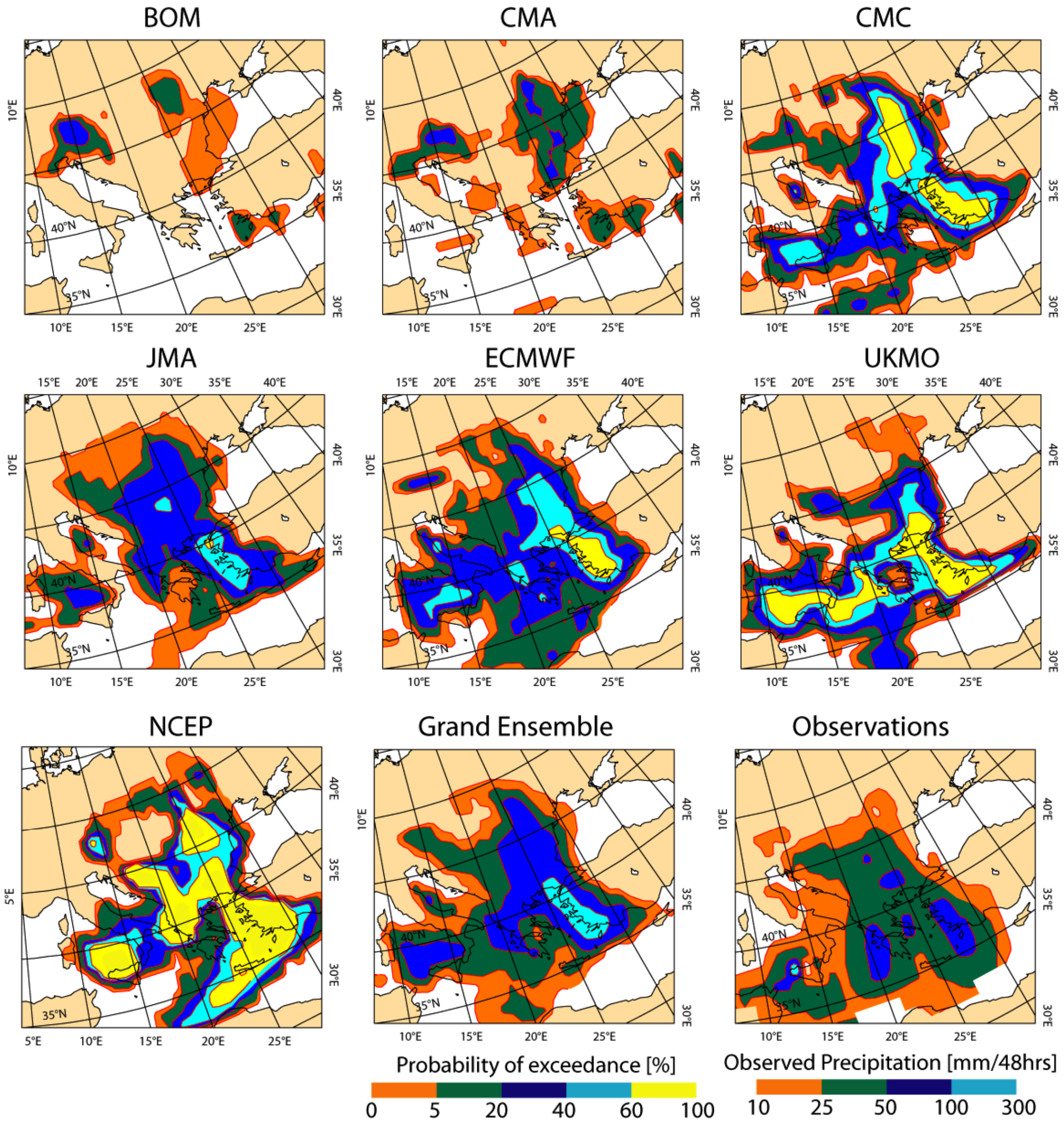


Top : linearized case. Bottom : nonlinear case



Top : linearized case. Bottom : nonlinear case

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



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

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 ?

Conclusions

Reliability and *resolution (sharpness)* are the attributes that make the quality of a probabilistic prediction system (my opinion at least …) . 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.

Must we tend to a situation where the output of prediction (and assimilation too) will be a probability distribution ?