On Some Aspects of Validation of Probabilistic Prediction

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

[1.] *Laboratoire de Météorologie Dynamique, École Normale Supérieure, Paris, France*
[2.] *Service Météorologique du Canada, Dorval, Canada*
[3.] *Current address : CNRM/GMAP, Météo-France, Toulouse, France*

Abstract

A number of questions relative to objective validation of ensemble prediction are discussed. It is argued that the quality of ensemble prediction (and more generally, of probabilistic prediction) lies in the conjunction of two properties. These are reliability on the one hand (i.e., consistency between predicted probabilities and observed frequencies of occurrence), and resolution on the other, (i.e., the property that reliably predicted probability elements are distinctly different from climatology). The impact of the size of the predicted ensembles, as well as of the size of the verifying sample, is discussed. Concerning the former, evidence is presented that no practical gain can be achieved from ensemble size beyond a few tens of units. As for the latter, it imposes strong limitations to objective validation of probabilistic prediction of (even moderately) rare events, as well as to validation of probabilistic prediction in multi-dimensional state space. Other aspects, such as the impact of errors in the verifying observations, and the definition of initial ensembles, are briefly discussed.

1. Introduction

In spite of the importance it has acquired over the last 15 years or so, ensemble prediction is still a relatively new technique, and a number of questions remain as to how it can best be implemented and evaluated. These notes present a short review of the notions that lie at the basis of evaluation of ensemble prediction, and of the methods that are used for that evaluation. Particular aspects, such as the consequences of the finiteness of the predicted ensembles and of the validating sample, are discussed at some length.

These notes do not aim at presenting an exhaustive discussion of the vast problem of evaluation of ensemble prediction, and they are certainly biased to some extent towards the authors' personal interests. While some material is new, a significant part of what is presented below has already been published elsewhere, possibly in a slightly different form. We refer in particular to Murphy (1973), Wilks (1995), Anderson (1996, 1997), Talagrand *et al*. (1999), Toth *et al*. (2003), Candille *et al*. (2007).

2. Reliability and resolution

It will be convenient for a start to discuss two questions that seem to be rather fundamental. The first question is that of the precise goal, or goals, that can can be assigned to ensemble prediction, and of what can be expected from ensemble prediction. The second question is that of the means that can be used for objectively (and, as far as possible, quantitatively) evaluating the quality of an Ensemble Prediction System, *i*. *e*., assessing the degree to which it achieves the intended purposes(s).

Which goal(s) can be assigned to ensemble prediction? Various goals have been considered in the literature for ensemble prediction. One is to reduce the forecast error, for instance by taking the mean of the predicted ensemble, thereby producing an individual forecast that can be hoped to be more accurate (at least statistically) than other individual forecasts that can be obtained. Other goals are simply to produce bounds on the meteorological quantities of interest, or else 'scenarii' as to what the future evolution of the meteorological flow will be. Still another goal is to produce an *a priori* estimate of the uncertainty on the state of the flow ('*forecast the forecast error*'). A similar, but in a sense more precise goal is to predict

quantitative probabilities, for instance for the occurrence of a particular event. Quantitative probabilities, if they are reliable, can for instance help to minimize long-term loss in situations that involve a financial risk in relation with weather.

All these goals are distinct, but can all be considered as particular aspects of a more general goal, which is simply to predict probabilities or probability distributions. If a probability distribution is available for a particular variable, the least-variance estimator for that variable is the expectation of the distribution. This defines a simple way for obtaining an accurate (or at least as accurate as possible given the uncertainty defined by the probability distribution) estimate of the quantity under consideration. Similarly, the knowledge of a probability distribution for the state of the flow defines bounds on variables (if not strict bounds, at least to within a degree of confidence), and bi- or multimodal distributions define 'scenarii'. And of course a probability distribution for the future state of the flow defines an uncertainty for that future state, as well as quantitative probabilities for the occurrence of events.

The position taken in these notes is therefore to consider ensemble prediction as producing probability distributions for the future state of the flow. A deterministic forecast is normally produced from an estimate of the initial state of the flow (the so-called *analysis*), and from the physical laws governing the evolution of the flow. These physical laws are available in practice in the form of a discretized numerical model. If a probability distribution is available for the uncertainty affecting both the analysis and the numerical model, the conditional probability distribution for the future state of the flow is unambiguously defined as a mathematical object. The complete explicit determination of that conditional probability distribution seems however to be totally beyond present technical means, at least in spaces with dimensions as large as those of present Numerical Weather Prediction models. Ensembles produced by an Ensemble Prediction System (EPS) are considered here as substitutes for the underlying inaccessible conditional probability distribution. Ensemble Prediction Systems thus appear as (degraded) forms of *probabilistic prediction*, the purpose of which is to predict probability distributions for numerical variables (or numerical probabilities for events). In particular, the validation tests considered below are tests of the hypothesis that the N elements of a predicted ensemble are independent realizations of a same probability distribution, of which the real state of the flow at verifying time is an (N+1)st independent realization. It is to be noted that this hypothesis may not be fully realistic. For instance, the initial perturbations of the ECMWF EPS consist of pairs of opposite perturbations, and are not mutually independent. For the sake of simplicity, we will however ignore possible mutual dependency between ensemble elements.

Also for the sake of simplicity, and unless explicitly specified otherwise, the expression 'probability distribution' will be used hereafter in a broad sense, to denote not only probability distributions for mono- or multidimensional variables defined on a continuum, but also numerical probabilities for two- or multi-outcome events.

How can one objectively (and quantitatively) evaluate ensemble predictions? The object predicted in probabilistic prediction is a probability distribution for, say, a particular scalar variable, while the verifying object is an observation of that variable. These two objects are of a different nature, and (contrary to what is the case in deterministic prediction, where the predicted and observed objects are of the same nature), cannot be mutually compared. The difference is actually much deeper than that. The conditional probability that is meant to be predicted is not better known afterwards than it was beforehand. Worse than that, it is meant only to describe our uncertainty on the future state of the flow, and has as such no objective existence. Probabilistic prediction is thus of a totally different essence than deterministic prediction, in the sense that it bears on an object that has no objective existence, and cannot be objectively observed.

A first consequence is that it is not possible, except in rather extreme circumstances, to speak of the quality of an individual probabilistic forecast. If rain is predicted to occur with, say, 40%-probability, neither its occurrence nor its non-occurrence makes the prediction a success or a failure. It is sometimes said that the occurrence of an event after it has been predicted to occur with a high probability (say 80%) is a success. That is said particularly when the event under consideration is uncommon. But it would obviously be erroneous to interpret non-occurrence of the event as a failure of the forecast. If the event has been predicted to occur with probability 80%, there is a 20%-probability that it will not occur. It is its relative frequency of occurrence of the event over a large number of situations, not its occurrence or non-occurrence in a particular situation that matters. The only situation in which it can legitimately be said that a probabilistic forecast has been successful is when the forecast ensemble had a very small spread (significantly smaller than a typical uncertainty for the variable under consideration), and the observed verification fell within that small spread. The probabilistic forecast can then be considered as successful in that it has correctly predicted in advance that the prediction would be unusually accurate. Similarly, a probabilistic forecast can legitimately be considered as a failure if the verifying observation falls well outside the range of the predicted uncertainty. But, except in those rather extreme situations, objective evaluation of probabilistic prediction can at best be statistical.

A first statistical property that is required for the quality of a probabilistic prediction system is consistency between the predictions and the observations. Temperature must go below freezing 30% of the times it is predicted to go below freezing with probability 30%. That form of consistency is obviously necessary, for instance for users who, in anticipation of the possible occurrence of a particular meteorological event, must take a decision that involves a financial risk. It is called reliability. Its most general definition is as follows. For any probability distribution F, let us denote $F'(F)$ the conditional observed frequency distribution, given that F has been predicted. Reliability is the property that $F'(F) = F$ for any F.

Reliability can be objectively evaluated by a number of well-known diagnostics that will not be discussed here. A limited list includes *reliability diagrams*, *rank histograms*, the *Reduced Centred Random Variable*, and the (appropriately named) *reliability component* of the *Brier* and *Brier-like scores*. Those various diagnostics, although they all measure some aspect of reliability, are not exactly equivalent. They are described, and their specific significance and properties are discussed, in a number of publications. We mention Murphy (1973), Wilks (1995), Talagrand *et al*. (1999), Toth *et al*. (2003), Candille and Talagrand *et al*. (2005).

Assume a particular probability distribution F has been predicted often enough so that the corresponding conditional observed frequency distribution $F'(F)$ is known from the verifying observations. The next time the system predicts F, the prediction can be replaced by $F'(F)$. That *a posteriori calibration* makes the system reliable. Lack of reliability, under the hypothesis of stationarity of the performance of the system, can be corrected to the same degree it can be diagnosed. But the diagnosis obviously requires a large enough validation sample.

Now, reliability, if it is obviously necessary for the usefulness and the value of a probabilistic prediction system, is clearly not sufficient. A system that would always predict the climatological probability distribution would be perfectly reliable in the sense that has just been defined, but would nevertheless be devoid of any usefulness. A second property is therefore that the probability distributions F‘(F) be non trivial. That property can be expressed in several similar (but not necessarily exactly equivalent) forms. It can be expressed by the condition that the distributions $F'(F)$ are different from the climatological distribution, or that they have small individual spread, or large mutual spread. That property will be called here *resolution* (although some authors disagree with that word). Like reliability, it can be quantitatively

measured by a number of (non exactly equivalent) scores such as (what everybody agrees to call) the *resolution component* of the *Brier* and *Brier-like scores*, the *Relative Operative Characteristic* (*ROC*) curve area, or the *mean entropy* (*information content*) of the distributions $F'(F)$. These scores are described and discussed in the same references mentioned above concerning scores for reliability.

3. The size of the predicted ensembles and of the verification sample

Three causes at least can introduce uncertainty on the diagnostics. These are the finiteness of the ensembles, the finiteness of the verifying sample and the noise in the verifying observations. All three causes have been studied, among other places, in Candille (2003). More recently, the question of the finiteness of the verifying ensemble has been discussed in Candille and Talagrand (2005). The question of the observational noise has been discussed in, *e.g*., Saetra *et al*. (2004), Bowler (2006a) and Candille and Talagrand (2008), who have proposed various ways for taking the noise into account.

Concerning the size of the ensembles, an obvious question is *Given the choice, is it better to improve the quality of the forecast model, or to increase the size of the predicted ensembles?* Actually, if the size of the ensembles is the relevant quantity as concerns cost-efficiency, it is the numerical accuracy with which probabilities are predicted that is relevant from the purely scientific point of view. It is possible for instance, using ensembles with dimension 100, to lump elements together so as to predict probabilities with accuracy of, say, 1/20. What we will consider here is the accuracy with which probabilities are predicted. Figure 1, extracted from Talagrand *et al*. (1999), relative to the ECMWF Ensemble Prediction System, shows variations of the Brier Skill Score (a variant of the Brier Score) as a function of the number of elements in the ensemble. The event under consideration is that the 850-hPa temperature is higher than its climatological average. The verification area is the 30N-70N latitudinal belt, and the verification period is January-April 1997. The size *N* of the ECMWF EPS ensembles was 50 at that time (as it still is). For each value *n* on the horizontal axis, the Brier Skill Score shown in the figure was obtained from *n* randomly drawn elements from the 50-ensemble. The left panel of the figure is relative to the operational EPS of ECMWF, the right panel to a 'poor man's EPS' based on a search for analogues in the archives of the high resolution operational deterministic forecast (see Talagrand *et al*., 1999, for more details). For each panel, the 10 curves correspond to the 10 forecast ranges 1, 2, …, 10 days. The Brier Skill Score is positively oriented, and it is seen that the quality of the forecast decreases with increasing lead time. But what we want to particularly stress here is that the quality of the forecast increases with ensemble size, but also saturates very rapidly with that size. The score saturates for size smaller than 10 for 1-day forecasts, and for size smaller than 30 for 10-day forecasts. Talagrand *et al*. (1999) mention that, assuming an infinitely large verification sample, the Brier Skill Score BSS_N varies, as a function of the size *N* of the ensembles, as

$$BSS_N = BSS_\infty - (1/N)A \tag{1}$$

where BSS_∞ is the score that would obtain with infinitely large ensembles, and the positive constant *A* decreases with increasing spread of the predicted probabilities (an exact proof of eq.1 is given in Richardson, 2001). This is in agreement, at least qualitatively, with Fig. 1, where the score saturates more rapidly for short forecast ranges, for which predicted probabilities assume mostly the values 0 and 1, than for longer ranges, for which predicted probabilities assume intermediate values more often.

The rapid saturation of scores with ensemble size in the range of a few tens seems to be very general. Candille (2003) has made systematic diagnostics, with real and simulated data, which all lead to the same conclusion, and the authors do not know of any evidence to the contrary, in particular in the more recent literature. This raises two new questions

a. Is it useful to use ensembles with dimension larger than a few tens of units, at which objective scores are observed to saturate?

b. What is the reason for such a rapid saturation?

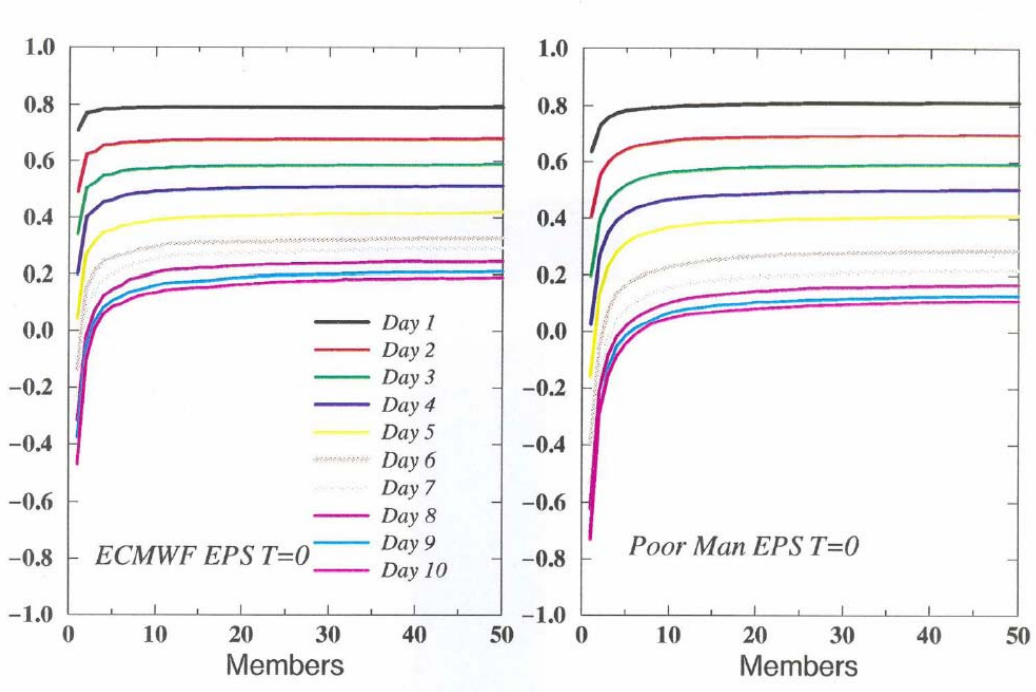

Figure 1: Brier Skill Score for ensemble prediction of the 850-hPa temperature as a function of the ensemble size. Left panel : ECMWF operational EPS. Right panel : 'poor man's EPS'. See text for details (extracted from Talagrand et al., 1999).

Concerning the first question, let us suppose an EPS produces ensembles with dimension N=200, which allows in principle to predict probabilities with accuracy 1/200. It is hard to imagine which user will ever be interested in knowing *a priori* probabilities with such an accuracy, and in knowing in advance whether the predicted probability for, say, rain is 132/200 or 133/200. In that respect, an accuracy somewhere in between 1/50 and 1/10 seems to be largely sufficient for all practical purposes.

And even if an accuracy of 1/200 is considered to be useful, then comes the question of the size of the verification sample that would be necessary to check that the difference between forecasts 132/200 and 133/200 is significant. That is a difficult question, which does not seem to have been fully answered yet. Consider the simpler question of objectively assessing the reliability of a probability forecast of $1/N$ for an event $\mathcal{E}$. In order to check whether that forecast is statistically correct, it must have been made αN times, where α is at the very least of the order a few units, and the event $\mathcal{E}$ must have occurred about α times out of those αN times. Let us assume one 10-day ensemble prediction is made every day, so that 10 successive forecasts are produced for a given date. Let us assume in addition that, among the 10 forecasts that precede an occurrence of $\mathcal{E}$, the N+1 probabilities i/N (i = 1,…, N) are statistically predicted with the same frequency $1/(N+1)$. The probability that the particular probability $1/N$ has been predicted over the 10 days preceding a given occurrence of $\mathcal{E}$ is then equal to about $10/N$. Let in addition T be the mean time between two successive occurrences of $\mathcal{E}$. The minimum waiting time for the event to have occurred α times, while having also been predicted to occur with probability $1/N$, is then

$$\alpha TN / 10 \qquad (2)$$

For an event that occurs 4 times a year (hardly a rare event) and for $\alpha = 4$ (not a very stringent requirement), the waiting time is 5 years for N = 50, and 20 years for N = 200. The prediction system will have significantly evolved before such waiting times have elapsed.

The estimate (2) is very crude, and probably very conservative in that it only defines a minimum waiting time. Candille (2003), using a different but more refined argument, evaluates the size of the verification sample that is necessary for checking significance of binary event probabilistic prediction with accuracy 1/N (*e. g.*, checking that the difference between predictions 132/200 and 133/200 is significant). He finds that the size of the verification sample varies like $BN\ln N$, where the constant B is of the order $O(100)$. This grows more rapidly than (2). Without going into a more detailed (and more difficult) analysis, there are obviously very strong limitations to objective validation of probabilistic prediction of even moderately rare events. That conclusion must be tempered in several respects. Gross lack of reliability can be detected from a small verifying sample for intermediate probabilities (if an event is predicted every day to occur with probability ½, and does not occur over one month, then the prediction is obviously unreliable). And an event that is relatively rare at a given location can be frequent over a whole hemisphere or the entire globe. If geographical location does not matter, the time T in eq.(2) can be short, even for an event that is rare in a given place. Validation is also possible, along what was done by Hamill and Whitaker (2007), through reforecasts over past situations. However, in view of the fact that the observing system evolves very rapidly, what can be obtained from reforecasts may not be representative of the performance of EPSs in the future. In addition, climate is expected to change significantly in the coming decades. This will further degrade the possible utility of long period past validation samples.

Although the present picture is not entirely clear, there is substantial evidence, coming from two totally different sources (the observed performance of existing EPSs on the one hand, and considerations on the size of the verification sample that would be necessary to objectively evaluate large dimension ensemble predictions) that very little gain, if any, is to be achieved by using ensembles with dimension larger than a few tens of units. Is that conclusion is confirmed, it provides a clear answer to the question raised above as to whether it is better to increase the ensemble size and to improve the quality of the prediction model. Beyond an ensemble size of a few tens, it is preferable to improve the quality of the model.

The second question raised above is why scores for evaluation of EPSs saturate in the range 30-50. The same phenomenon has been observed by Candille (2003) with low-dimensional systems that bear no relationship with atmospheric dynamics, so that it seems the reason is not to be sought in specific features of the atmosphere (such as for instance the number of unstable modes at a given time in the flow), but rather in some general properties of probability theory. A possible explanation is that probability distributions predicted by EPSs are generally of a simple, unimodal form. In addition, evaluation is most often performed on binary events (or events with a small number of possible outcomes), or on one-dimensional scalar variables. If an ensemble size of, say, 50, may seem ridiculously small for sampling a probability distribution in a space with dimensions 10^6 or more, it is certainly not for sampling a one-dimensional probability distribution, and the gain obtained by increasing the ensemble size must rapidly saturate in a small dimension space.

4. Is validation possible in large state dimensions?

The last remark raises in turn a new question. Is it possible at all to evaluate prediction of large dimension probability distributions? According to the definitions that were given above for reliability and resolution, full assessment of a probabilistic prediction system first requires, for each predicted probability distribution F, to wait until it has been predicted a large enough number of times so that the corresponding conditional frequency distribution $F'(F)$ can be reliably assessed. Consider the case of probabilistic prediction of the

500-hPa geopotential field over the midlatitude winter Northern Atlantic. To fix ideas, assume the prediction is made over a 5x5-degree^2 grid covering the area (10-80W, 20-70N), which corresponds to 165 gridpoints. Assume the ensemble size is N=5, and 'probabilities' are defined at each gridpoint by the positions of the N predicted values for that gridpoint with respect to L=2 predefined thresholds. That is a very crude way of defining a one-dimensional probability distribution, and the whole thing is perfectly feasible with present computing power. It is actually very economical in comparison to the computing requirements of the present EPSs of major meteorological centres. The number of ways of positioning N numerical values with respect to L thresholds is equal to the binomial coefficient

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

i. e., 21 for N=5 and L=2. This leads to the absurd value $21^{165} \approx 10^{218}$ for the number of probability distributions that the system can in principle predict over the 165-point grid. Most of those probability distributions will of course never be predicted, but one can seriously doubt whether any of them will ever be predicted twice. This clearly shows that objective validation of probability distributions in large dimensional spaces is simply impossible. Drastic reduction of the effective dimension of the predictions is necessary for validation. In the present case, that would mean lumping gridpoints together over at least large fractions of the geographical domain of the prediction. That is what is done, out of necessity, is validation of existing EPSs, but any possibility of validating probabilistic prediction of spatial patterns in the flow is lost.

Consider the much more modest goal of 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. That would be hardly sufficient for accurate validation, especially if one takes into account the fact that the prediction system will probably have evolved to some extent over a 3-year period.

There are obviously very strong limitations, even for small ensemble sizes, to the validation of multi-dimensional probabilistic forecasts. These limitations result from the ineluctably small dimension of the validation sample (at least in comparison with the dimension that would be required for accurate validation), and will not be relaxed by technical progress.

Further work is certainly necessary to assess more precisely what the limitations to probabilistic prediction are, but the discussion of this section suggests that, in addition to the previous conclusion that ensemble sizes beyond a few tens of units are useless, objective validation of large dimensional probabilistic prediction is simply impossible.

5. The impact of observational errors

The verifying observations will most often, if not always, be affected by errors, and it is important to determine the impact of those errors on the validation. One possibility is to perturb the ensemble elements with random noise that is meant to simulate the uncertainty in the observations. This approach has been studied by, *e.g.*, Saetra *et al*. (2004). It does not however evaluate how accurately the prediction predicts the reality, but how it predicts the observations. Candille and Talagrand (2008) have observed that most of validation scores for ensemble prediction do not require the validating observation to be a uniquely determined numerical value, but can very well deal with observations that are considered as numerical probabilities (for events) or a probability distributions (for numerical variables). The Brier and the Brier-like

scores can thus be extended to the case of noisy observations, as can the ROC area score. The former keep their standard reliability-resolution decomposition in that new setting. The Reduced Centred Random Variable can also be extended to the case of noisy observations. Numerical simulations show that, generally speaking (but not systematically), the scores are degraded in comparison to what they would be in the absence of observational noise. However, that approach may not be fully satisfactory in that it is also an assessment of the degree to which the prediction fits the observation, not the reality. In the case of the Brier score, the estimated performance can actually decrease if the forecast error becomes smaller than the observational error (which is perfectly conceivable, especially for short range forecasts).

Still another possibility is suggested by the work of Bowler (2006a). Let $F'(F)$ be the conditional probability density function for reality, given that the probability density function *F* has been predicted (we are here more precise than we were above, in that we specifically consider *density* functions). If the observational error is statistically independent of reality, the observed conditional density function will be the convolution product

$$F'' = F' * G$$

where *G* is the density function of the observational error. The Fourier transform $\mathcal{F}\left[F' * G\right]$ of a convolution product is the numerical product of the Fourier transforms of the factors.

$$\mathcal{F}\left[F' * G\right] = \mathcal{F}\left[F'\right] \times \mathcal{F}\left[G\right]$$

This defines in principle a way for determining F' from F'' and *G*, thereby filtering out the effect of observational errors. It remains however to see if that is practical, and in particular if the size of the validation sample will be large enough to allow reliable estimation of *F*'.

6. The definition of initial conditions

A question that is still debated is the best definition of the initial conditions for ensemble prediction. ECMWF uses a combination of singular modes of the flow (Molteni *et al*., 1996). The singular modes are the modes which are most unstable, in a precise sense, over a given time period. They are independent of any estimate of the initial uncertainty on the state of the flow. The National Centers for Environmental Prediction (USA) use the 'bred modes', which are obtained through a procedure which can be described as a crude approximation of the assimilation process (although observations are not used), and leads to initial perturbations which are concentrated in the directions that have been unstable in the recent past (Toth and Kalnay, 1997). Other methods, like the so-called Perturbed Observations method (Houtekamer *et al*., 1996), the Ensemble Kalman Filter (EnKF, Evensen, 2003) and the Ensemble Transform Kalman Filter (ETKF, Bishop *et al*., 2001), are assimilation methods, and produce an ensemble that is meant to sample the uncertainty on the flow at the initial time of the prediction. The Ensemble Kalman Filter is used at present by the Meteorological Service of Canada for the definition of the initial conditions of its EPS.

A number of systematic comparisons have been performed, on models of varying complexity (and most often with simulated data), in order to assess the relative quality of those different methods. An incomplete list is Houtekamer and Derome (1995), Anderson (1997), Hamill *et al*., (2000), Wang and Bishop (2003), Bowler (2006b) and Descamps and Talagrand (2007). Figure 2, obtained from experiments performed with the NWP primitive equation model ARPÈGE of Météo-France, shows, as a function of lead time, the statistical quality of ensemble predictions for three different initialization methods. More precisely, it shows the (negatively oriented) Brier score for the event that the 700-hPa relative humidity differs from its climatological mean by more than one standard deviation (that event occurs with frequency 0.38). The comparison is made here between the methods of perturbed observations (full curve), of bred vectors (dashed

curve) and of singular vectors (dash-dotted curve). It is shown that the quality of the predictions (and of the initial conditions too) decreases in that same order.

Similar results are obtained with other variables and validation scores, both as concerns reliability and resolution. More generally, the results of figure 2 are typical of those that have been obtained in the various comparison studies mentioned above. Methods that are intended, through a process of ensemble assimilation, at producing a sample of the initial uncertainty on the state of the flow, most often perform better than methods that only intend at identifying the unstable modes of the flow. For instance, Descamps and Talagrand (2007) have performed comparisons between EnKF, ETKF, bred modes and singular vectors. The results vary somewhat with the model that is used, the variable on which the validation is done, and the score used for the validation, but the performance almost systematically ranks as EnKF > ETKF > bred modes > singular vectors. All those results are in full agreement with a notion which seems obvious to the authors, namely, that if a probabilistic prediction system is evaluated by how well it samples the uncertainty on the future state of the flow, then the best initial conditions are those that sample best the initial uncertainty. In other words (and although there is an obvious link between present uncertainty and instabilities that have developed in the recent past), it is better to diagnose directly where the uncertainty lies than to identify the unstable components of the flow.

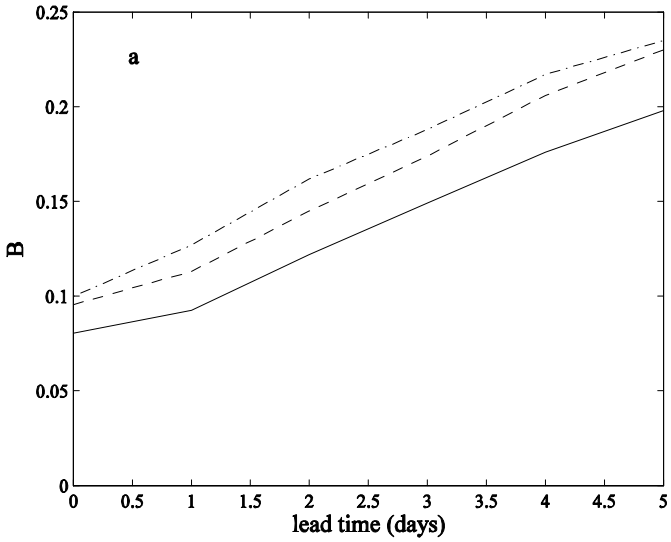

Figure 2: Brier Score for ensemble prediction of the 750-hPa relative humidity with the ARPÈGE model (simulated data ; see the text for details). Intial ensembles are defined by the Perturbed Observations method (full curve), bred modes (dashed curve) and singular vectors (dash-dotted curve).

Now, Buizza (*pers. com.*) has performed comparison experiments with the ECMWF EPS (and in particular with real data) that lead to an opposite conclusion, in that the best performance is obtained with singular vectors. Obviously, more work is still to be done on the best definition of the initial conditions of probabilistic prediction.

7. Conclusions

A number of questions relative to the validation of ensemble prediction have been briefly discussed in these notes. Although further work is clearly necessary on many aspects, several conclusions can be drawn. Two totally different lines of argument (the observed performance of EPSs on the one hand, and the size of the validation sample that would be required on the other) strongly suggest that no gain can be obtained by increasing numerical accuracy of predicted probabilities beyond $1/N$, where N is of the order of a few tens of units. Very strong limitations also exist on objective evaluation of probabilistic prediction of even

moderately rare events, as well as on objective evaluation of probabilistic prediction in multi-dimensional state spaces. Those limitations need to be more precisely assessed. Finally, the question of the best definition of the initial ensembles remains open. The authors' view is that it is preferable to identify where the initial uncertainty lies than to identify the unstable components of the flow *per se*. But additional work is certainly necessary on that aspect too.

8. References

Anderson, J. L., 1996, A Method for Producing and Evaluating Probabilistic Forecasts from Ensemble Model Integrations, *J. Climate*, **9**, 1518-1530.

Anderson, J. L., 1997, The Impact of Dynamical Constraints on the Select□ion□ of Initial Conditions for Ensemble Predictions: Low-Order Perfect Model Results, *Mon. Wea. Rev.*, **125**, 2969-2983.

Bishop, C. H., B. J. Etherton and S. J. Majumdar, 2001, Adaptive Sampling with the Ensemble Transform Kalman Filter. Part I: Theoretical Aspects, *Mon. Wea. Rev.*, **129**, 420–436.

Bowler, N. E., 2006a, Explicitly Accounting for Observation Error in Categorical Verification of Forecasts, *Mon. Wea. Rev.*, **134**, 1600–1606, doi: 10.1175/MWR3138.1.

Bowler, N. E., 2006b, Comparison of error breeding, singular vectors, random perturbations and ensemble Kalman filter perturbation strategies on a simple model, *Tellus*, **58A**, 538-548, doi: 10.1111/j.1600-0870.2006.00197.x

Candille, G., 2003, *Validation des systèmes de prévisions météorologiques probabilistes* (in French), Doctoral Dissertation, Université Pierre-et-Marie-Curie, Paris, France, 143 pp..

Candille, G., C. Côté, P. L. Houtekamer and G. Pellerin, 2007, Verification of an Ensemble Prediction System against Observations, *Mon. Wea. Rev.*, **135**, 2688-2698.

Candille, G., and O. Talagrand, 2005, Evaluation of probabilistic prediction systems for a scalar variable, *Q. J. R. Meteorol. Soc.*, **131**, 2131-2150, doi: 10.1256/qj.04.71.

Candille, G., and O. Talagrand, 2008, Impact of Observational Error on the Validation of Ensemble Prediction Systems, *Q. J. R. Meteorol. Soc.*, in press.

Descamps, L., and O. Talagrand, 2007, On Some Aspects of the Definition of Initial Conditions for Ensemble Prediction, *Mon. Wea. Rev.*, **135**, 3260–3272, DOI: 10.1175/MWR3452.1.

Evensen, G., 2003, The Ensemble Kalman Filter: Theoretical Formulation and Practical Implementation, *Ocean Dynamics*, **53**, 343-367.

Hamill, T. M., and C. Snyder, R. E. Morss, 2000, A Comparison of Probabilistic Forecasts from Bred, Singular-Vector, and Perturbed Observation Ensembles, *Mon. Wea. Rev.*, **128**, 1835-1851.

Hamill, T. M., and J. S. Whitaker, 2007, Ensemble calibration of 500 hPa geopotential height and 850 hPa and 2-meter temperatures using reforecasts. *Mon. Wea. Rev.*, **135**, 3273-3280, DOI: 10.1175/MWR3468.1.

Houtekamer, P. L., and J. Derome, 1995, Methods for Ensemble Prediction, *Mon. Wea. Rev.*, **123**, 2181-2196.

Houtekamer, P. L., L. Lefaivre, J. Derome, H. Ritchie and H. L. Mitchell, 1996, A System Simulation Approach to Ensemble Prediction, *Mon. Wea. Rev.*, **124**, 1225-1242.

Marshall, J., and F. Molteni, 1993, Toward a Dynamical Understanding of Planetary-Scale Flow Regimes, *J. Atmos. Sci.*, **50**, 1792-1818.

Molteni, F., R. Buizza, T. N. Palmer, and T. Petroliagis, 1996: The ECMWF Ensemble Prediction System: methodology and validation. *Quart. J. Roy. Meteor. Soc*., **122**, 73119.

Murphy, A. H., 1973, A new vector partition of the probability score, *J. Appl. Meteor*., **12**, 595-600. Richardson, D. S., 2001, Measures of skill and value of ensemble prediction systems, their interrelationship and the effect of ensemble size, *Q. J. R. Meteorol. Soc*., **127**, 2473-2489

Saetra, Ø., H. Hersbach, J.-R. Bidlot and D. S. Richardson, 2004, Effects of Observation Errors on the Statistics for Ensemble Spread and Reliability, *Mon. Wea. Rev*., **132**, 1487-1501.

Talagrand, O., R. Vautard and B. Strauss, 1999, Evaluation of Probabilistic Prediction Systems, in *Proceedings of Workshop on Predictability* (October 1997), ECMWF, Reading, England, 1-25, available at the address http://www.ecmwf.int/publications/library/do/references/list/16233.

Toth, Z., and E. Kalnay, 1997, Ensemble Forecasting at NCEP and the Breeding Method, *Mon. Wea. Rev*., **125**, 3297-3319.

Toth, Z., O. Talagrand, G. Candille and Y. Zhu, 2003, Probability and Ensemble Forecasts, in *Forecast Verification. A practitioner's Guide in Atmospheric Science*, I. Jolliffe and D. B. Stephenson, editors, John Wiley & Sons, Ltd, Chichester, United Kingdom, ISBN: 0-471-49759-2, 137-163, available at the address http://wwwt.emc.ncep.noaa.gov/gmb/ens/ens_info.html.

Vautard, R., 1990, Multiple Weather Regimes over the North Atlantic: Analysis of Precursors and Successors, *Mon. Wea. Rev*., **118**, 2056-2081.

Wang, X., and C. H. Bishop, 2003, A Comparison of Breeding and Ensemble Transform Kalman Filter Ensemble Forecasts Schemes, *J. Atmos. Sci*., **60**, 1140-1158.

Wilks, D. S., 1995, *Statistical Methods in the Atmospheric Sciences: an Introduction*, International Geophysics Series, Vol. 59, Academic Press, 464 pp..