EVALUATION OF PROBABILISTIC PREDICTION SYSTEMS

O Talagrand and R Vautard,
Laboratoire de Météorologie Dynamique du CNRS
Paris, France

and
B Strauss
Météo-France
Illkirch, France

Summary. The performance of Ensemble Prediction Systems and, more generally, of *Probabilistic Prediction Systems* (PPSs), *i. e.* systems that predict a probability distribution function (pdf) for the state of the physical system under consideration, is evaluated in terms of two specific qualities. The first quality is *statistical consistency* between the predicted pdfs and the *a posteriori* verifying observations, and the second quality is *variability* in the predicted pdfs. Concerning probabilistic prediction of binary events, the corresponding Brier score is decomposed, following *Murphy* (1973), into two terms that precisely measure the two above qualities. Concerning prediction of numerical variables, several integrated measures of statistical consistency are considered.
The various diagnostics thus defined are applied to the Ensemble Prediction System of ECMWF. The latter is also compared to an economical 'poor man's' PPS, built on statistical properties of past deterministic forecasts. The poor man's system has much better statistical consistency than the ECMWF EPS, and globally performs better up to a range of about three days.

1. INTRODUCTION

The quality of Numerical Weather Forecasts is highly variable from one meteorological situation to another, and the need for *a priori* accurate assessment of the quality of individual forecasts has been felt for a long time. This has led several major meteorological centres to develop in the last few years *Ensemble Prediction Systems* (EPSs), in which a number of numerical forecasts are performed from initial conditions whose dispersion is meant to represent the uncertainty on the initial state of the flow. The resulting dispersion of the forecasts can then be taken as an indication on the corresponding uncertainty on the future state of the flow. The European Centre for Medium-range Weather Forecasts (ECMWF) and the National Centers for Environmental Predictions (NCEP, Washington, DC, USA), among others, are both running operational Ensemble Prediction Systems. Descriptions of these two systems, as well as a number of diagnostics on their performance, can be found in *Molteni et al.* (1996) and in *Toth and Kalnay* (1997) respectively.

A large experience has now been accumulated on Ensemble Prediction. A significant correlation has been observed between the spread of the ensemble forecasts and the *a posteriori* observed forecast error. And the results of EPSs have on repeated occasions increased the confidence of forecasters in the expected occurrence of an unusual meteorological development. A number of questions nevertheless remain, as concerns in particular the objective assessment of the quality of an EPS. These questions become critical when a change is envisaged on an EPS, and it is necessary to decide whether a gain can be reasonably expected from the change.

The present paper is devoted to the general question of the objective assessment of the quality of Ensemble Prediction Systems. Rather than limiting ourselves to Ensemble Prediction Systems as they are operationally implemented at present, we will more generally consider what we will call *Probabilistic Prediction Systems* (PPSs), *i. e.* systems that do not predict the state of the physical object under consideration, but rather predict a *probability distribution function* (pdf) for that state. In all practical situations, the product of a PPS will consist of a finite number of numerical values. These values may parametrically define the predicted pdf, as do for instance the expectation and covariance matrix of a gaussian pdf. Alternatively, as in an EPS, they may define a finite number of states of the physical system under consideration, meant for instance to be independent realizations of the 'underlying' predicted pdf.

As already mentioned by several authors (see, *e. g.*, *Wilson et al.*, 1996), one basic difficulty in assessing the quality of a PPS is that the predicted object (a pdf over the space of possible states of the physical system under consideration) and the verifying object (an observed state of the system) are not of the same nature. It is therefore not possible, contrary to what happens in deterministic prediction, to assess the value of a prediction from a measured 'distance' between the predicted and verifying objects. Indeed, our opinion is that it is not possible to assess in any way the quality of an individual realization of a PPS.

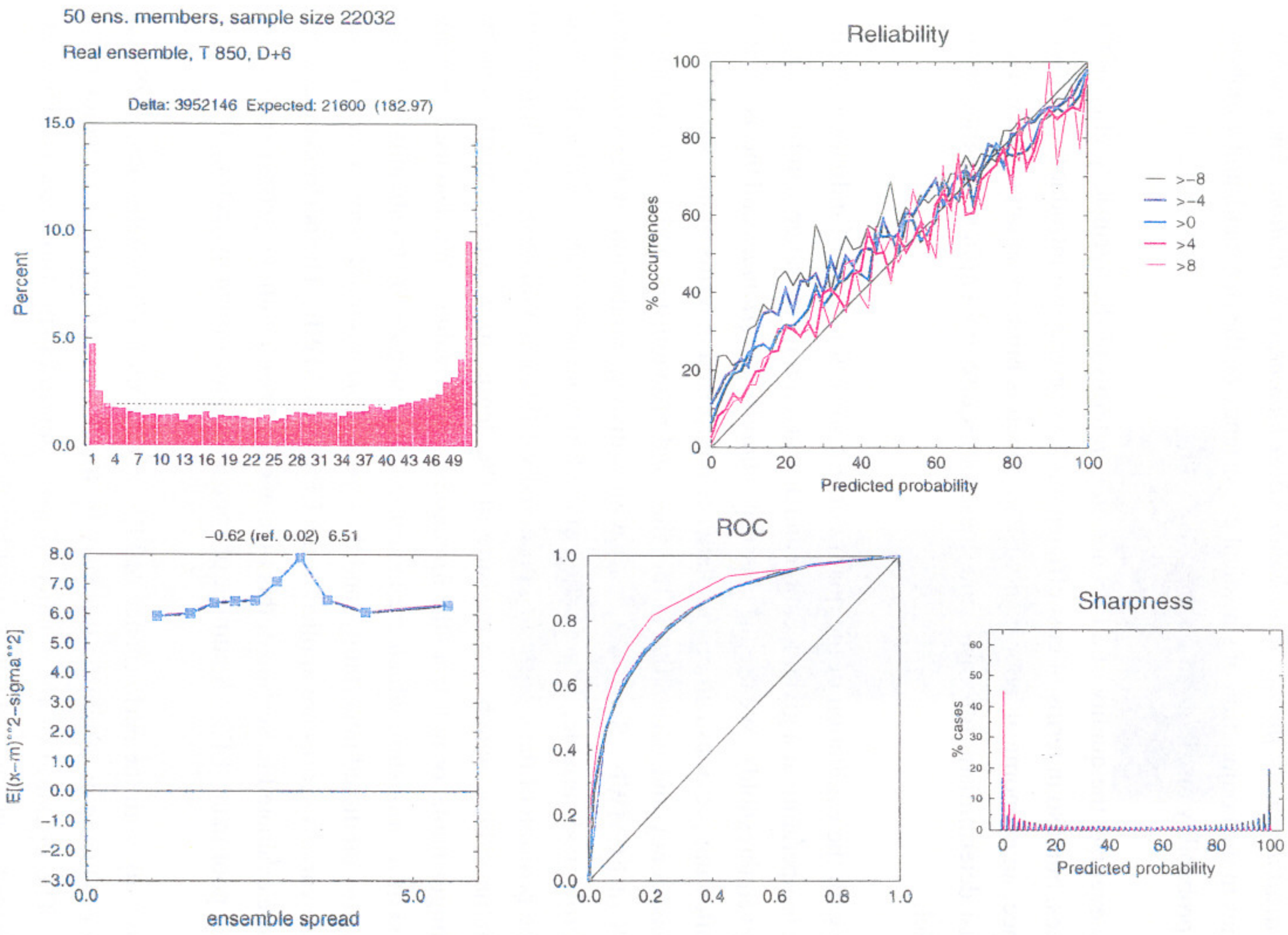
A number of various scores are used in this paper for evaluating the quality of a PPS. These scores are interpreted in terms of the degree to which they measure two mutually independent qualities, already discussed by other authors (see, *e. g.*, *Hsu and Murphy*, 1986). The first quality is that the predicted probabilities are in agreement with the verifying observations. A prediction like '*the probability of rain is 40%*' can be considered as exact only if rain is observed to occur with a frequency of 40% in those circumstances when it is predicted to occur with probability 40%. Agreement of this kind is absolutely necessary, for instance for users who must make a decision on the basis of an objective quantitative risk assessment. The first quality which a PPS must possess is therefore what we will call *statistical consistency* or *reliability*, *i. e.*, agreement between the *a priori* predicted probability distributions and the *a posteriori* observations. That agreement is most generally defined by the following condition: "for each possible probability distribution *f*, the *a posteriori* verifying observations are distributed according to *f* in those circumstances when the system predicts the distribution *f*". A PPS which does not possess that quality is obviously flawed in some sense. On the other hand, statistical consistency, as just defined, is clearly not sufficient for ensuring that a PPS is practically useful. A PPS which would always predict the climatological distribution of the state of the atmosphere would be statistically consistent, but would nevertheless be devoid of any practical utility. The effective usefulness of a statistically consistent PPS therefore depends on the *variability in the predicted probability distributions,* or *resolution.* The practical value of a PPS lies in the conjunction of statistical consistency on the one hand, and variability in the predicted probability distributions on the other.

It is desirable to evaluate the quality of a PPS not only in terms of the intrinsic quality of the results it produces, but also in terms of cost efficiency. To that end, we introduce, as a useful baseline reference, an economical 'poor man's EPS', which is built on an appropriate use of analogues in past deterministic forecasts, and does not require any explicit integration of a forecasting model.

Section 2 deals with the evaluation of the quality of statistical forecasts of individual binary events (*e. g.* the temperature at a given location and at a given forecast range will be larger or smaller than a given threshold). We use the classical *reliability diagrammes* and the associated *Brier score.* Following previous authors, the latter is interpreted in terms of the two qualities of statistical consistency and variability in the predicted probabilities, and is applied to the operational EPS of ECMWF. Section 3 deals with statistical prediction of the value of a particular meteorological variable, or 'prediction of the forecast uncertainty'. We use there the histograms of the position of the *a posteriori* observed verification with respect to the *a priori* predicted ensemble values, which is a measure of the degree to which the verification is statistically distinguishable, or not, from the forecast ensemble values. We also use estimates of the forecast-skill relationship, which are measures of the degree of statistical consistency between the *a priori* predicted uncertainty, and the *a posteriori* observed error in the forecast. Here again, these various scores are applied to the EPS of ECMWF. The performance of the latter is once more evaluated in Section 4, this time in comparison with the performance of the above mentioned poor man's EPS. A number of conclusions are drawn in Section 5.

This paper deals only with the performance of PPSs as predictors of probabilities. PPSs can also be used for producing deterministic forecasts, by for instance taking the expectation of the predicted pdf. The question of the quality of such deterministic forecasts is extremely interesting in its own right, but will not be considered here.

We also mention a work recently done by *Atger* (1998), which, although it uses different diagnostics, leads to conclusions which are very similar to ours concerning the performance of present EPSs. Analogous conclusions have also been obtained recently by *Ziehmann* (1998).



BS & BSSS	0.072	0.148	0.164	0.100	0.043	0.252	0.300	0.328	0.368	0.342
calibrated	0.070	0.142	0.158	0.097	0.042	0.279	0.329	0.354	0.382	0.350

Fig. 1 Various scores for the ECMWF EPS prediction of the 850 hPa-level temperature over the midlatitude Northern Hemisphere for the period 1 January - 10 April 1997, at lead time 6 days. Detailed explanations are given in the text. The four rows of numbers at the bottom right of the figure show various Brier scores. For each row, the five values refer to the five events E_τ described in the text, with the thresholds τ taking the values t = -8, -4, 0, 4, 8 K in that order from left to right. Top left row: raw Brier score *B* (eq. 2. 1). Top right row: Brier Skill Score BSS (eq. 2.3a). Bottom left row: variability component of the raw Brier score B_V (eq. 2.2a). Bottom right row: variability component of the Brier Skill Score (1 - BSS_V, where BSS_V is defined by eq. 2.3c).

2. PROBABILISTIC FORECAST OF INDIVIDUAL EVENTS

We recall that the reliability curve relative to the statistical prediction of a given event E is obtained by plotting, as a function of the predicted probability p, the actually observed frequency of occurrence $p'(p)$ of E in the circumstances when p has been predicted. Statistical consistency is expressed by the equality $p'(p) = p$. The right top panel of Figure 1 shows five reliability curves relative to the prediction at the six-day range, by the ECMWF EPS, of the 850-hPa temperature deviation from its climatological average. The five curves are relative to events E_τ of the form 'the temperature deviation is larger than τ', where the threshold τ takes the values $\tau = -8, -4, 0, 4, 8$ K. Statistics have been accumulated over 34 realizations of the EPS in the period from 1 January to 10 April 1997, and over 648 points located on a 5 x 5 degree grid between 30N and 70N around the globe, so that the effective sample size is $M = 22032$. A number of already well identified features of the ECMWF EPS are clearly visible. The reliability curves are close to the diagonal $p' = p$, suggesting a high degree of statistical consistency. The slopes of all five curves are slightly less than one, *i. e.* large values of probability of occurrence are overpredicted, and small values are underpredicted. It is also seen that reliability curves are noisy, suggesting that, although the size of the sample over which statistics have been accumulated is rather large, each probability p has not been predicted often enough for the value of the corresponding frequency of occurrence $p'(p)$ to be statistically stabilized. The predicted probabilities are of the form $p = n/N$ ($n = 0, ..., N$), where $N = 50$ is the size of the forecast ensembles. The lower right panel of Figure 1 (labelled 'Sharpness') shows the distribution of the predicted probabilities for the three thresholds $\tau = -4$, 0 and 4 K. It is seen that even for the middle value $\tau = 0$ (shown in green), extreme probabilities tend to be predicted more often. This shows a rather large variability in the predicted probabilities.

One classical measure of the quality of statistical prediction of a given event E is given by the Brier score (*Brier*, 1950)

$$B = \frac{1}{M} \sum_{i=1}^{M} (p_i - o_i)^2 \qquad (2.1)$$

where M is the number of realizations of the PPS over which statistics have been accumulated. For each realization i, p_i is the predicted probability, while o_i assumes the value 1 or 0 depending on whether the event E has been observed to occur, or not to occur. The Brier score is equal to 0 for a perfect deterministic system, which always correctly predicts the occurrence, or non-occurrence, of E.

It is convenient to use here a continuous, rather than discrete, representation of the probabilities, and we will denote by $g(p)$ the frequency with which the probability p has been predicted. For $p_i = p$, the quantity $(p_i - o_i)^2$ takes the value $(p - 1)^2$ with frequency $p'(p)$, and the value p^2 with frequency $1 - p'(p)$. The contribution of the probability p to the Brier score is therefore equal to $(1- p')p^2 + p'\,(1- p)^2$, and

$$B = \int_0^1 [(1- p')p^2 + p'\,(1- p)^2] g(p)\,\mathrm{d}p$$

This is in turn easily transformed into (see *Murphy*, 1973)

$$B = B_c + B_v \tag{2.2a}$$

where

$$B_c = \int_0^1 (p' - p)^2 g(p)\,\mathrm{d}p \tag{2.2b}$$

and

$$B_v = \int_0^1 p'(1 - p') g(p)\,\mathrm{d}p = -\int_0^1 (p' - p_c)^2 g(p)\,\mathrm{d}p + p_c(1 - p_c) \tag{2.2c}$$

In the latter expression

$$p_c = \int_0^1 p'(p)\, g(p)\,\mathrm{d}p$$

is the climatological frequency of occurrence of E (different PPSs would produce different distributions $g(p)$ and $p'(p)$, but p_c would remain the same). B_c, which is 0 for $p'(p) = p$, is clearly a measure of the statistical consistency of the system. As for B_v, it is a measure of the dispersion, about p_c, of the actually *a posteriori* observed frequencies of occurrence p' of the event *E*. B_v decreases from $p_c(1-p_c)$ for 'climatological' systems (*i. e.* systems that are not able to predict the occurrence of *E* more accurately than with its climatological frequency of occurrence) to 0 for a system in which p' takes only the values 0 and 1 (such a system, if it is statistically consistent, is then a perfect deterministic system). If accumulation of statistics has shown that the event *E* occurs with frequency $p'(p)$ when it is predicted by the PPS to occur with probability p, the right thing to do, the next time the system 'predicts' p, is obviously to actually predict $p'(p)$ instead (see, *e. g.*, *Zhu et al.*, 1996). This *a posteriori* correction renders statistically consistent an originally inconsistent system. It amounts to shifting horizontally to the diagonal all points of the reliability diagramme. As for the Brier score, it is reduced in that transformation to B_v. The latter quantity therefore measures the variability of the *a posteriori* predictable probabilities of occurrence p'.

The Brier score *B* thus decomposes into two terms which independently measure the two qualities whose conjunction has been considered above as making the effective value of a Probabilistic Prediction System, *viz.* statistical consistency between the predicted probabilities and the *a posteriori* observations on the one hand, and variability in the predicted probabilities on the other.

In order to compare the accuracies of the statistical predictions of different events, it is convenient to normalize the Brier score with the value it takes for a climatological PPS. We will in the following use the so-called Brier Skill score

$$BSS = 1 - B/[p_c(1-p_c)] \qquad (2.3a)$$

The value of the Brier Skill score *increases* from 0, for a climatological PPS, to 1 for a perfect deterministic system. We will also use other scores, and in particular

$$BSS_c = B_c\, [p_c(1-p_c)] \qquad (2.3b)$$

which is a normalized measure of the statistical consistency of the system under consideration, and

$$BSS_v = B_v\, [p_c(1-p_c)] \qquad (2.3c)$$

which is a normalized measure of the variability in the probabilities it predicts. The values of both scores BSS_c and BSS_v decrease with increasing quality of the system.

The values of the Brier Skill score *BSS* corresponding to the five reliability curves of Figure 1 are given at the bottom of the Figure (upper right sequence of values). The best performance is for the threshold $\tau = 4$K. The line below gives the value of the Brier Skill score for the *a posteriori* corrected probabilities (*i. e.* the quantity $1 - BSS_v$, with the above notations). It is seen that the improvement from *BSS* is numerically small. However, owing to the lack of reference and to the fact that statistical consistency and variability of predicted probabilities are intrinsically different properties, it is difficult to judge whether there is any real significance in the fact that the change from *BSS* to $1 - BSS_v$ is numerically small.

The accuracy with which the quality of a PPS can be evaluated is unavoidably limited in practice by various sources of uncertainty. Two such sources are the error which is necessarily present in the verifying observations, and the finiteness of the number of realizations over which the performance of the system is evaluated. In the case of an Ensemble Prediction System, an additional source of uncertainty is the finiteness of the ensembles from which predicted probabilities are estimated. Assessing the effect of these various sources of uncertainty is essential for giving significance limits to the various scores one can compute for evaluating the quality of a PPS. We will here describe the effect, on the Brier score, of the finiteness of the forecast ensembles of an EPS. In the limit of an infinite number of realizations of the system, and of perfect observations, the Brier score B_N of an EPS based on forecast ensembles of size N is equal to

$$B_N = B + \frac{1}{N}\int_0^1 p(1-p)g(p)\,\mathrm{d}p \qquad (2.4)$$

where B is the 'exact' score, which would be obtained in the limit of infinite N. It is seen that increasing N, with ensemble elements being drawn from the same statistical population, will always result in a numerical decrease of the Brier score, *i. e.* in an *increase* of the quality of the system. Although we think this increase is real, it only results from the fact that increasing N has the effect of smoothing noise due to the finiteness of the ensembles. It does not correspond to any improvement in the intrinsic quality of the system, measured by the limit value B. The second term on the right-hand side of eq. (2.4) is a measure of the dispersion of the *a priori* predicted probabilities p (see eq. 2.2c for comparison). It results that the numerical impact of increasing N will be larger when the predicted probabilities have small dispersion (small sharpness) than when they have large dispersion.

The panel denoted *ROC* in Figure 1 shows the so-called *Relative Operating Characteristics* curves associated with the three events E_τ, with τ = -4, 0, 4 K (same colour code as for the reliability diagrammes). For each event, a curve is built through a process, described in detail in *Stanski et al.* (1989), based on a stratification by observations of the results of the PPS. As a consequence, the ROC curve depends only on the dispersion of the effectively predictable probabilities $p'(p)$, and not on the statistical consistency of the system. Each ROC curve joins the lower left corner and the upper right corner of the square in the figure. The extreme cases of a climatological forecast and of a perfect deterministic forecast respectively correspond to the diagonal of the square and to the curve consisting of the left and upper sides of the square. The resolution of a probabilistic prediction system for the occurrence of an event E can therefore be measured by the area below the corresponding ROC curve. That area is a different measure than the quantity BSS_v introduced above, in the sense that there is not a numerical one-to-one relationship between BSS_v and the ROC area. But it is seen from the ROC panel that the two measures lead to the same qualitative conclusions, namely that the performance of the ECMWF EPS is similar for the three events under consideration, with a slight but distinct advantage for the case τ = 4K (red curve).

3. PROBABILISTIC FORECAST OF INDIVIDUAL VARIABLES

We now consider a scalar variable $x(t)$, such as temperature or geopotential at a given point, to be probabilistically forecast at time t. The product of the forecast is then a one-dimensional pdf. If a deterministic forecast is also available (such a probabilistic forecast could for instance be the expectation of the predicted pdf), one can say that what the probabilistic forecast produces is an *a priori* estimation of the forecast error, or more precisely an *a priori* estimation of the forecast uncertainty.

In the case of an EPS, the predicted pdf for the variable x will be defined by the N values x_i (i = 1, ..., N) produced by the ensemble forecasts. Ranking these values in increasing order defines N+1 intervals. If the verifying observation x_v is an additional independent realization of the same pdf which has produced the x_i 's, x_v will be statistically undistinguishable from the x_i 's, and will therefore fall with equal frequency 1/(N+1) in each of the intervals defined by the x_i 's. The histogram of the position of x_v with respect to the x_i 's therefore defines a measure of the statistical consistency of the EPS. A perfectly consistent system will produce a flat histogram. The left top panel of Figure 1 shows the histogram for the data sample already considered in the previous section. It is seen that the verifying observation falls in each of the extreme intervals much more frequently than in the middle intervals, which is just an indication of the often observed fact that the spread of the ECMWF forecast ensembles is too small. A more accurate quantitative diagnotic is given by the sum of the squared differences, over all

N+1 intervals, between the population s_j of each interval and its expected value $M/(N+1)$, *viz.*

$$\Delta = \sum_{j=1}^{N+1} (s_j - \frac{M}{N+1})^2$$

It is easily seen that, if x_v is distributed uniformly over the N+1 intervals, the expected value of D is $MN/(N+1)$, *i. e.* 21600 in the present case. The actual value is 3952146, almost 200 times as large.

Another measure of the statistical consistency of a PPS can be obtained from the quantity

$$D = ENSK - ENSP \qquad (3.1)$$

In this expression, *ENSK* is the 'ensemble skill', *i. e.* the squared difference $(x_v - m)^2$, where m is the expectation of the predicted pdf, while *ENSP* is the 'ensemble spread', *i. e.* the variance of the predicted pdf. The expectation of D, for a given pdf, is 0. Denoting by <> averages taken over a large number of realizations of the system, which sample the various predicted pdfs, the quantity

$$<D> = <ENSK> - <ENSP> \qquad (3.2)$$

must therefore be 0. A similar argument shows that the quantity

$$y = \frac{x_v - m}{ENSP^{\frac{1}{2}}} \qquad (3.3)$$

averaged over a large number of realizations of the system, must have mean 0 and variance 1.

A basic product that most users will expect from a probabilistic prediction system, before quantified probabilities, is an estimate, even if only qualitative, of the confidence to be given to the forecast. Basically, one will want to be sure that, if the spread of a predicted pdf is small, then the corresponding uncertainty on the forecast is small, while the uncertainty is large if the spread is large. This 'spread-skill relationship' is of course only one aspect of statistical consistency, which can be globally measured by the statistical properties of the quantities D and y above. Globally integrated quantities may not however be a very good measure of spread-skill relationship, since it is conceivable that inconsistencies, such as small predicted spread in case of large uncertainty and large predicted spread in case of small uncertainty, will compensate in global statistics. A more accurate diagnostic of spread-skill relationship is obtained by computing the statistics of D or y independently for limited subsets of the value of the predicted spread. The left bottom panel of Figure 1 shows the average of D (in unit K^2), computed for different values of the spread of the forecast ensemble (the spread being defined here as the difference between the largest and the smallest values in the ensemble). The significantly positive value of D shows again that the spread of the forecast ensembles is too small in comparison with the deviation of the verifying observation from the ensemble mean. But an interesting fact is that the average of D is essentially independent of the spread itself, which means that the amount by which the actual spread is underestimated is statistically independent of the value of the spread. A consequence is that there is a positive correlation between predicted spread and forecast skill.

4. COMPARISON WITH A 'POOR MAN'S PPS'

4.1 Definition of a 'poor-man's' PPS

The objective of this Section is to propose a simple and costless PPS that can serve as a reference for evaluation of EPSs. This 'poor-man's' PPS is also an EPS, and assumes the knowledge of only past control forecasts and their associated verifications (if possible observations, if not analyses).

We consider as above a variable $x(t)$ to be probabilistically forecast at time t. We assume that we have a set of K previous control forecasts $x_c(t_k)$ from the same prediction model, and the corresponding verification values $x_v(t_k)$, with k = 1, ... , K. At time t, we are only in possession of the present control forecast $x_c(t)$. One simple way of constructing an ensemble prediction with N members is to extract from the past record the N nearest-neighbour values of $x_c(t)$ (N must be much smaller than K), and take as predictions the corresponding verification values. We will denote by I the subset of these N indices between 1 and K. The ensemble prediction of the value $x(t)$ is therefore $\{x_v(t_k), k \in I\}$. This set of predictions can be manipulated just as any prediction ensemble. Note that by construction, this poor-man's EPS is close to being statistically consistent, unless the sample of past values is not long enough.

In the following, we will compare the ECMWF EPS with the poor-man's EPS using the measures previously defined in this paper. That is, we will compare reliability diagrammes, Brier scores and spread-skill relationships. The comparison is done on the T850 field over the extratropical Northern hemisphere as before. In the poor-man's EPS application, one has to use a cross-validation scheme in order to increase the size of the data which the subset I is to be extracted from. For a given day, the "training" sample from which I is taken simply consists of all gridpoints on all other days. The number of nearest neighbours is taken exactly as the ECMWF ensemble size, *i. e.* N=50.

4.2 Prediction of individual events

4.2.1 *Brier Scores*

As before, we consider the event that the temperature anomaly is larger than a given threshold τ. We first compare the Brier skill scores (*BSS*), for lead times ranging from 1 to 10 days, and for values of the threshold τ ranging from -10K to 10K. Figure 4.1 shows the results. The top panel shows the *BSS* of the ECMWF EPS. The *BSS* decreases with increasing lead time and is fairly insensitive to the temperature threshold, at least at short lead times. Note however that probabilistic prediction of extreme temperature values has slightly lower skill. In the case of the operational EPS, the too small spread must of course degrade the forecast for large values of the threshold τ. More generally, we can expect prediction of extreme values to be more sensitive to model errors as well as to sampling errors. Note also a dissymmetry between positive and negative temperature anomalies, especially for long lead times, where the best score is obtained for the threshold τ = 4K. This fact, already noticed on Figure 1, probably results from model systematic errors.

The middle panel shows the *BSS* of the poor-man's scheme. Quite surprisingly, the same qualitative values and behaviour are found as for the ECMWF *BSS*, including the better performance, at large lead times, for the threshold τ = 4K. In order to compare the two EPSs, we show, in the bottom panel, the difference *BSS*(ECMWF) – *BSS*(Poor Man). At short lead times, *i. e.* for days 1, 2 and 3, the poor man's EPS beats the ECMWF EPS. By contrast, the poor man's EPS is clearly beaten in the medium range (days 4 to 10).

We now decompose the *BSS* into the two components (statistical consistency BSS_C, and variability BSS_V) presented above. Figure 4.2 shows, in the same format as Figure 4.1, the values of BSS_C for the two EPSs, and Figure 4.3 the values of BSS_V. First of all note that for both systems the values of BSS_C are one order of magnitude smaller than the values of BSS_V, meaning that the EPS skill is not greatly affected by inconsistency (a similar remark has already been made above concerning the numerical values shown in Figure 1). The ECMWF values are generally increasing with lead time, especially for negative anomaly thresholds, again reflecting model systematic errors. Interestingly, all ECMWF consistency curves seem to intersect near the +5K threshold, where best consistency for long lead times is achieved. This fact must be related to the best overall performance already noticed for the +4K threshold, but we are not able to give for it a fully satisfactory interpretation.

The poor man's consistency curves all display a convex shape, with higher inconsistency at extreme thresholds. This can be interpreted more easily by sampling effects, and in particular by the relative difficulty of having good extreme statistics from a finite sample of 50 members. Unlike for the global BSS, the poor man's scheme beats the ECMWF scheme at all lead times

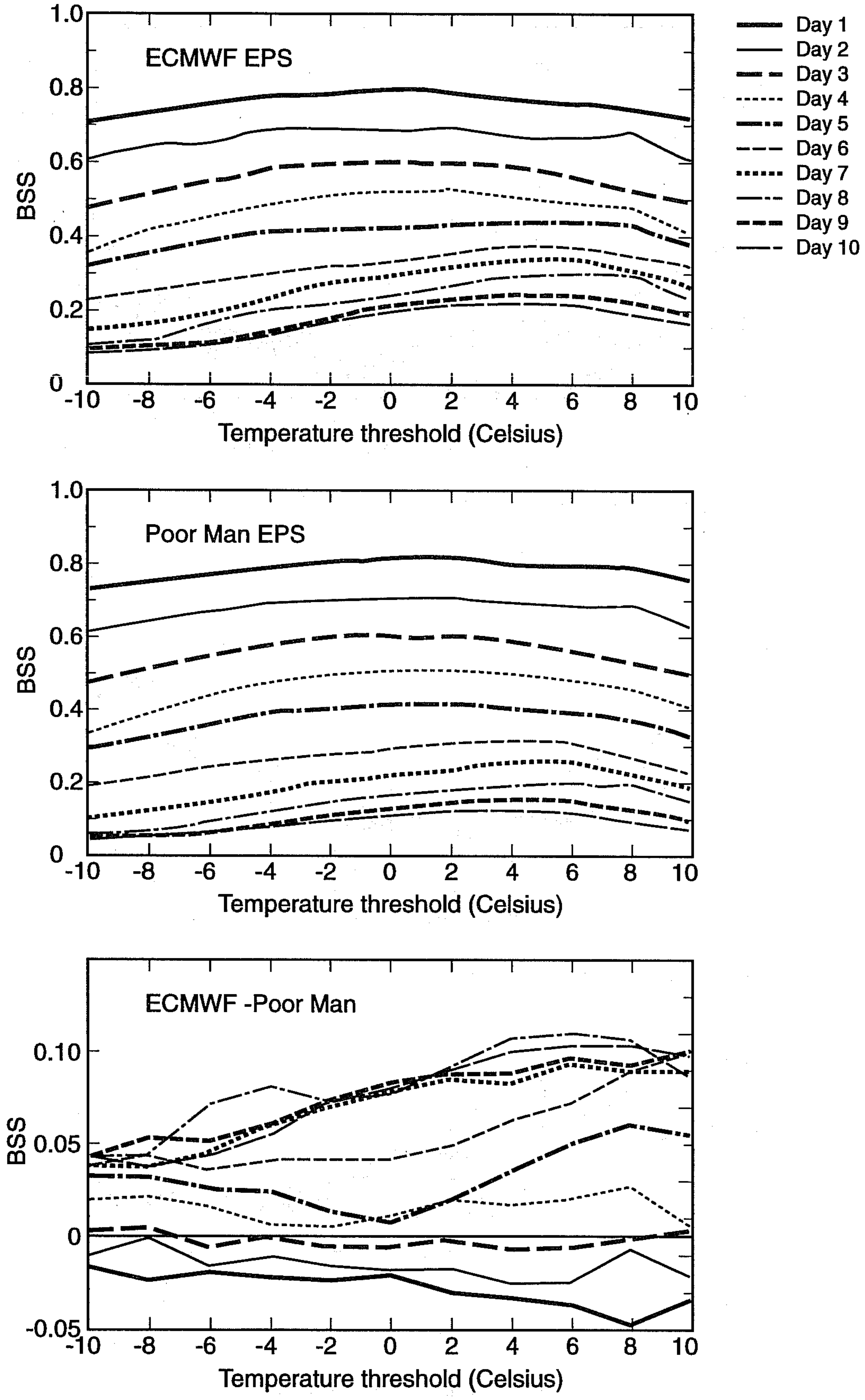


Fig 4.1 Brier Skill Score BSS (positively oriented) as a function of lead time and threshold. Top panel: operational EPS. Middle panel: poor man's EPS. Bottom panel: difference between top and middle panels.

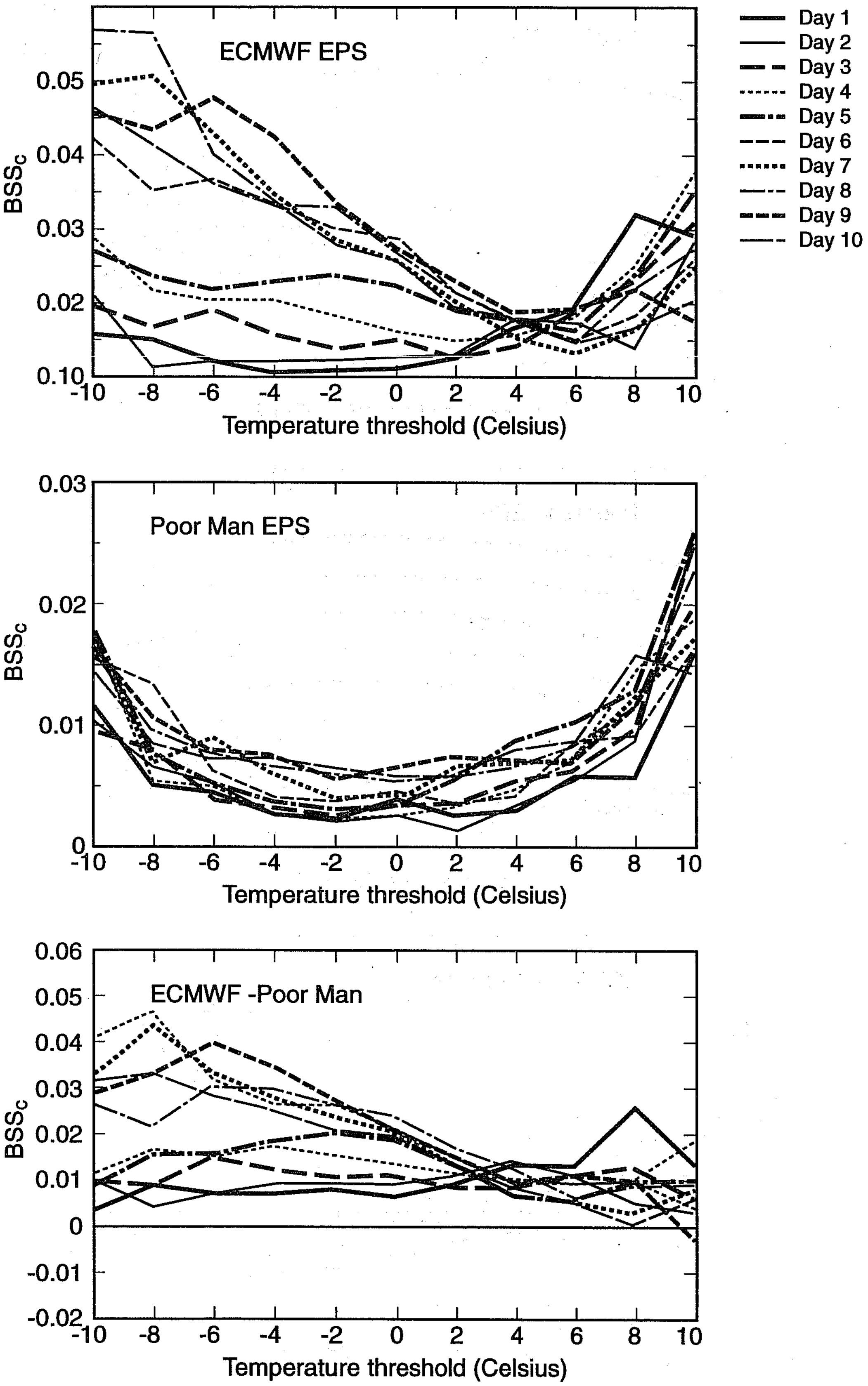


Fig. 4.2 Same as Figure 2.1, but for the consistency component BSS_V (negatively oriented) of the Brier Skill Score.

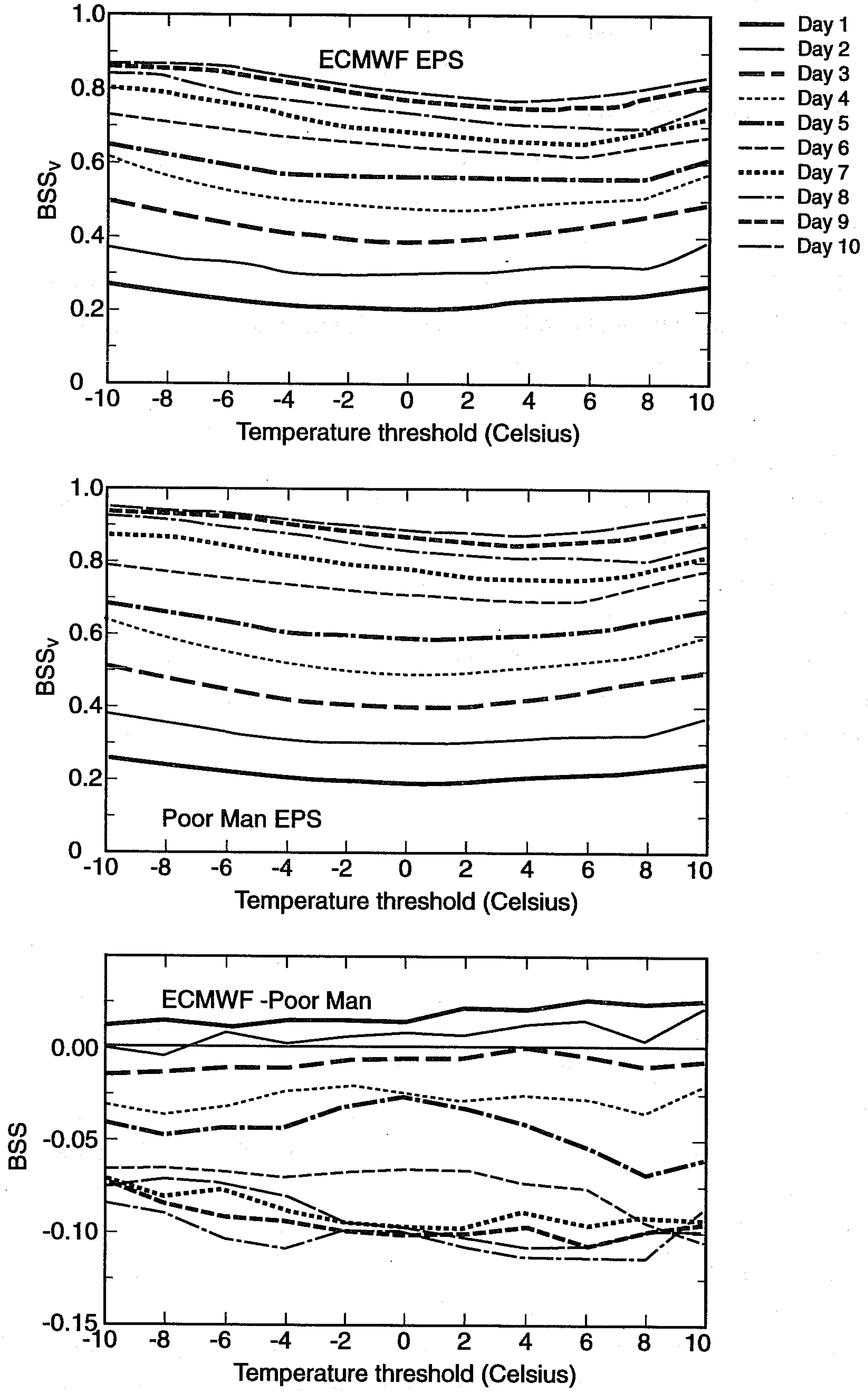


Fig. 4.3 Same as Figure 2.1, but for the variability component BSS_V (negatively oriented) of the Brier Skill Score.

in terms of consistency. This is not really a surprise since the poor man's EPS is specifically designed to be consistent, but it is still interesting to notice that such a good consistency is obtained from only a few "training" days. The use of a larger training set of control forecasts will certainly still increase consistency.

The values of BSS_V (Figure 4.3), which measure the variability of the PPSs, reveal a behaviour very similar to that of BSS (compare with Figure 4.1, keeping in mind that performance increases for increasing BSS, and for decreasing BSS_V). An interesting feature is that the ECMWF scheme now beats the poor man's scheme for lead times larger than 2 days instead of 3 days for the global BSS. We conclude that at day 3, it is only effect of higher consistency which makes the poor man's EPS superior to the ECMWF EPS.

4.2.2 *Reliability diagrammes*

Figure 4.4 shows the reliability diagrammes for both the ECMWF and the poor man's EPSs, for two threshold values, τ = 0K and τ = 8K and three lead times 1, 6 and 10 days. For τ = 0K, both the ECMWF and Poor Man EPSs are fairly consistent. Notice however that ECMWF reliability diagrammes have a substantial bias at low probabilities. This bias is amplified for extreme temperature forecasts and for short lead times, while for longer lead times the curve remains close to the diagonal (except for the noise). The poor-man's EPS having less variability for τ = 8K, it has no events of large probabilities especially at long lead times, hence some curves of the bottom right panel of Figure 4.4 "stop". An interesting fact comes from the poorer quality of the ECMWF EPS at short lead times, linked to its poorer reliability.

4.2.3 *Dependence on the size of forecast ensembles*

One particularly interesting question is whether one should continue increasing the size of the ensembles or rather concentrate efforts on other points. Figure 4.5 attempts to address this issue. We display the global BSS values as a function of the number of members N, for the median threshold (τ = 0K) and the extreme threshold (τ = 8K). One argument for the extension of the ensemble size is the better estimation of probabilities of extreme events. We should therefore see in Figure 4.5 a larger sensitivity to N for the threshold τ = 8K than for the threshold τ = 0K. Such is not the case. It is to be noticed that convergence is actually reached quite quickly at all lead times, for, say, N = 20-30. Notice that the sensitivity to N depends much more on lead time than on threshold. It is much larger for long than for short lead times. This is in agreement with one conclusion drawn above from eq. (2.4), according to which the sensitivity to N depends on the dispersion of the *a priori* predicted probabilities. The dispersion is smaller for long lead times, which corresponds, as seen in Figure 4.5, to larger

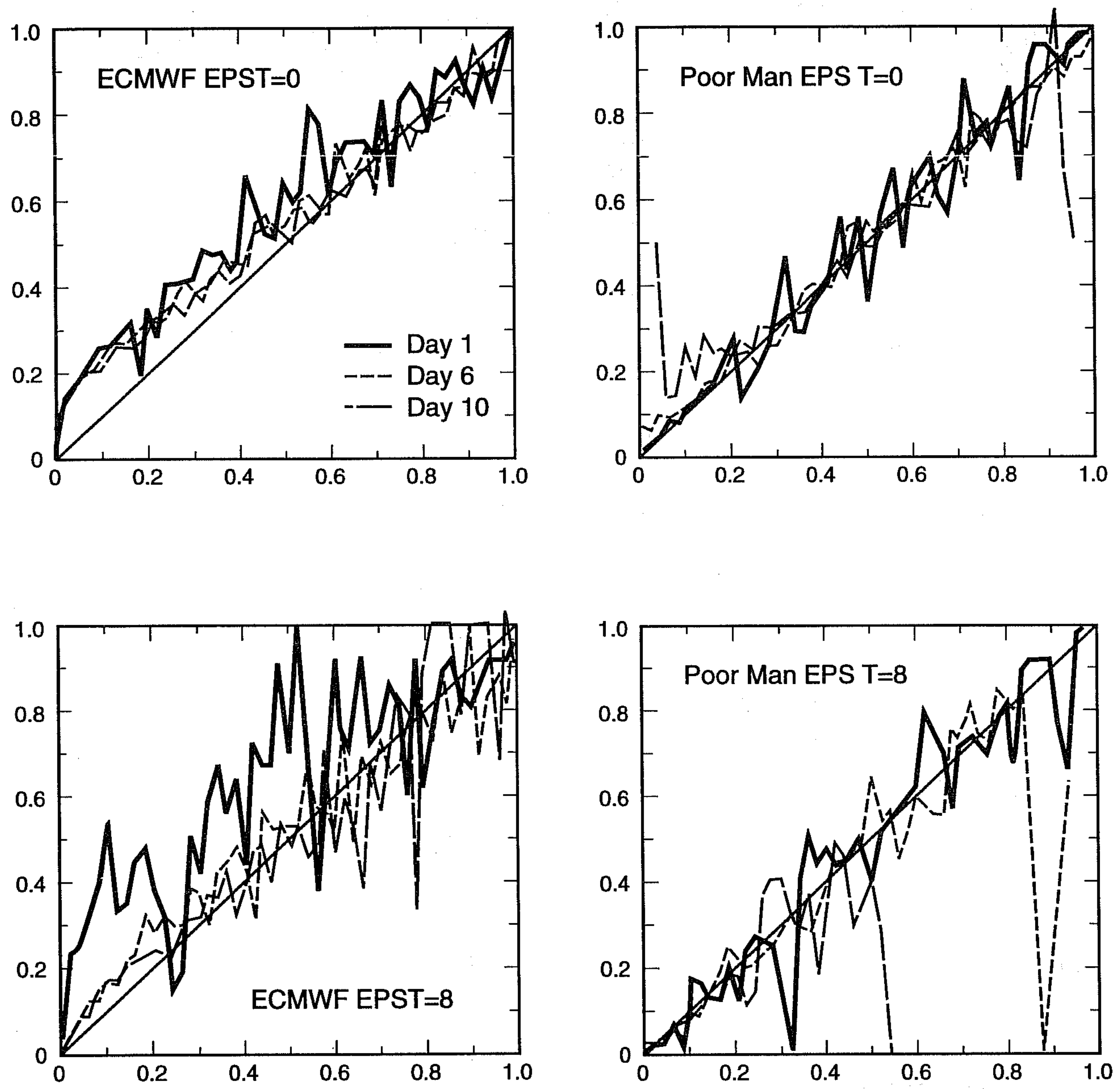


Fig 4.4 Reliability diagrams for lead times 1, 6 and 10 days, for thresholds 0 and 8 K (top and bottom panels respectively) and for the operational and poor man's EPSs (left and right panels respectively).

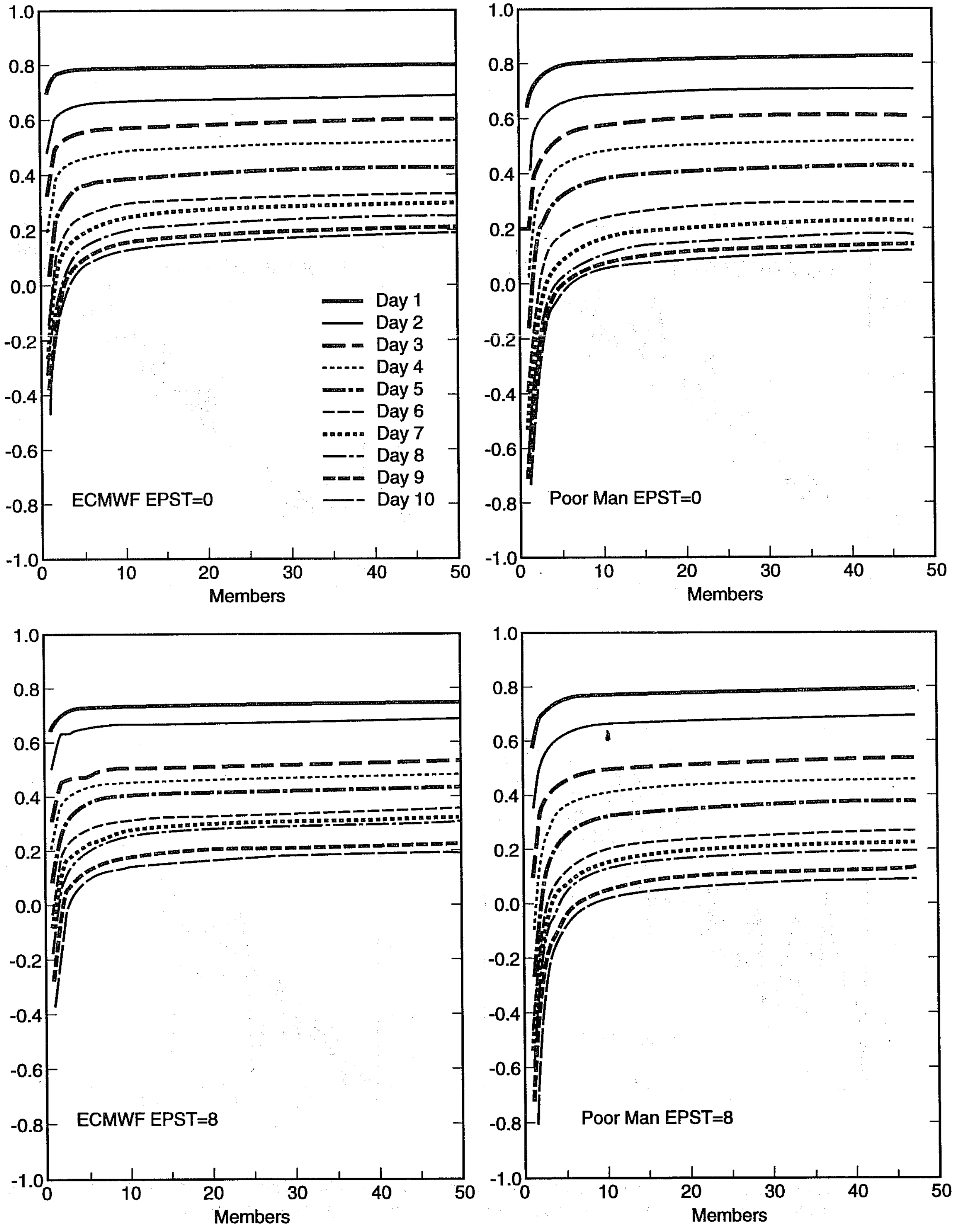


Fig. 4.5 Variations of the Brier Skill Score BSS as a function of the number of ensemble elements, for thresholds 0 and 8 K (top and bottom panels respectively) and for the operational and poor man's EPSs (left and right panels respectively).

sensitivity to N. This is true of both the operational and the poor man's EPSs.

4.3 Prediction of a continuous variable and forecast skill

Figure 4.6 displays various integrated measures of the spread-skill relationship (which, as already said, is one aspect of statistical consistency) for both the ECMWF and the poor man's EPSs. The quantity

$$ENC = \frac{<ENSK>}{<ENSP>} - 1$$

is the quantity $<D>$ above (eq. 3.2), divided by $<ENSP>$. Statistical consistency requires ENC to be equal to 0, with a positive value indicating too small a spread of the forecast ensembles. The quantity

$$COC = \frac{<COSK>}{<COSP>} - 1 \ ,$$

which is specific to PPSs based on a prior 'control' forecast, is defined as ENC, the control forecast being used instead of the forecast ensemble mean for computing the skill of the forecast and the dispersion of the ensemble. In the case of the ECMWF EPS, the control forecast is a better forecast on average than a randomly chosen element of the forecast ensemble (and not as good a forecast as the mean of the ensemble, with which it coincides at the start of the forecast period), but it is difficult to quantitatively assess how it could be used for defining the predicted pdf. However, statistical consistency implies that two quantities of the form $<COSK>$ and $<COSP>$ must be equal, independently of what was used as a reference to estimate the 'skill' of the forecast and the 'spread' of the ensemble.

An additional measure can be calculated for the poor man's EPS. Indeed, control skill can be estimated directly from the nearest neighbours by calculating the average square error of the nearest neighbour forecasts themselves, *i. e.* by defining

$$FCSK = \frac{1}{N} \sum_{k \in I} (x_c(t_k) - x_v(t_k))^2$$

while the “control spread” would be given, for the poor man's scheme, by

$$COSP = \frac{1}{N} \sum_{k \in I} (x_c(t) - x_v(t_k))^2$$

Then we can define again a global measure

$$FCC = \frac{<COSK>}{<FCSK>} - 1$$

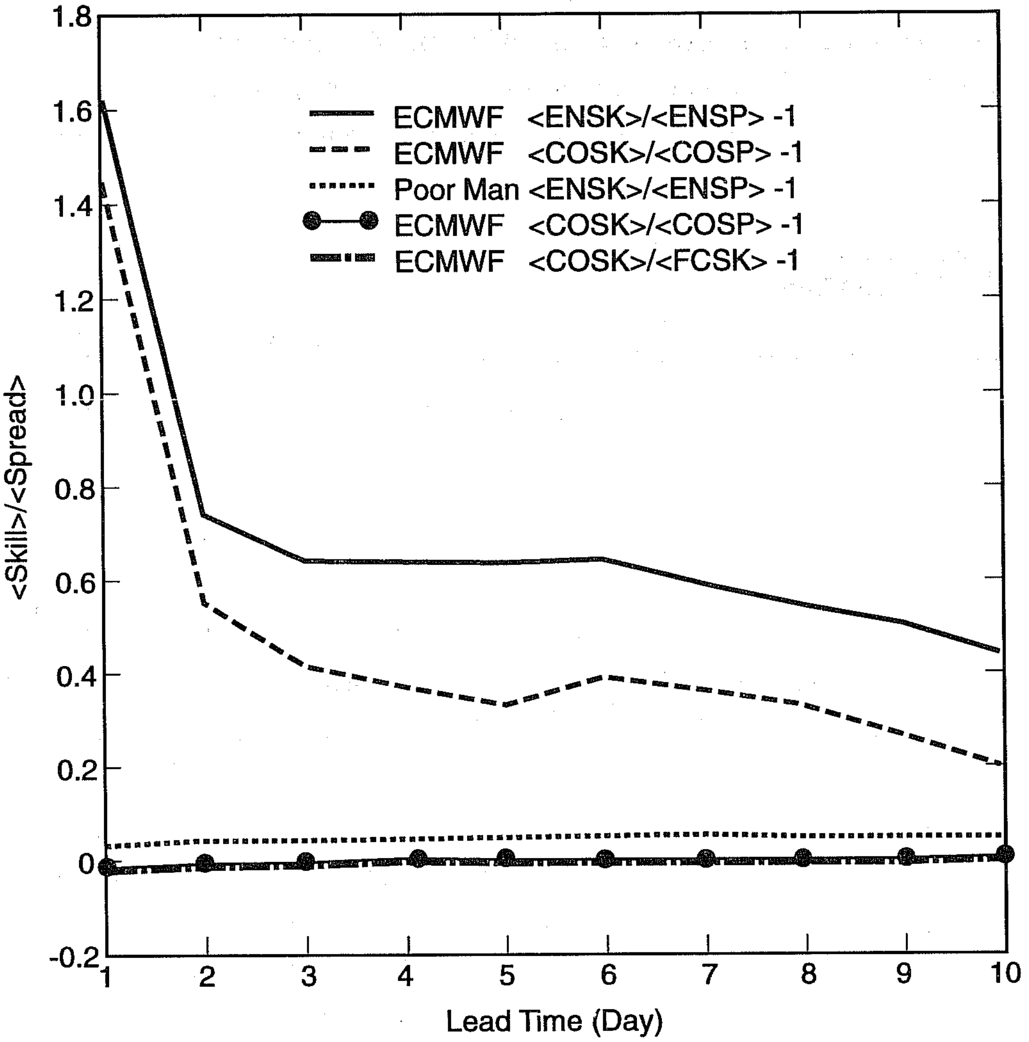


Fig. 4.6. Variations with lead time of various measures of the spread-skill relationship, for both the operational and the poor man's EPSs. Detailed explanations in the text.

Figure 4.6 shows the values of *ENC*, *COC* and *FCC* as functions of lead time for the two EPSs. The poor man's EPS is absolutely consistent in the sense of the different measures, while the ECMWF EPS displays a quite strong inconsistency, especially at short lead times. The spread at day 1 is about 2.5 times too small while it is about 1.5 times too small at later lead times. Notice also the large difference between day 1 and day 2. One can reasonably suspect the singular vector formulation to be responsible for this. Indeed, since perturbations are projected onto the fastest growing directions, ECMWF has to keep very small amplitudes for initial perturbations in order to obtain a reasonable spread at optimization time (2 days). Hence the spread/skill ratio is way too small before day 2. The ECMWF EPS was of course not designed for day 1-day 2 forecasts, but one can however question the underlying methodology, since in principle a perfect PPS should have a perfect spread-skill relationship at all lead times.

In Figure 4.7, we examine the detailed spread-skill relationships, for all lead times. Individual values of *ENSP*, *COSP* and *FCSK* are classified into 16 increasing-value equally-populated categories, and for each category the average of *ENSK* is plotted against the average of *ENSP* (resp. *COSK* against *COSP*, and *COSK* against *FCSK* for the poor man's EPS). Both EPSs bear more or less the same deficiencies (and qualities). The alignment along the diagonal is generally satisfactory except for small lead times (say below day 4). However the too small mean spread is reflected here by most of the curve points lying above the diagonal for the ECMWF EPS. For the latter, large-spread forecasts are generally more skilful, in terms of skill prediction, than small-spread forecasts. *One should therefore rely more on ensembles with higher spread than on ensembles with small spread.* Another general feature is the relative flatness of the spread-skill curve, especially at small lead times, revealing a lack of spread-skill relationship. Surprisingly, the poor man's scheme also suffers from this deficiency, which is not visible from figure 4.6. More experiments would be required, using simpler models, in order to fully understand this behaviour.

It is interesting that, judging from Figure 4.7, both EPSs behave in rather similar ways. This does not mean that individual forecast pdfs are similar. In order to illustrate this point, we address the question of whether the spread of the ensembles has any dependency on the temperature anomaly itself. Figure 4.8 shows *COSP* as a function of the control forecast of the anomaly temperature for both EPSs and for day 6. The poor man's spread clearly has higher values at extreme temperatures, while this is much less obvious for the ECMWF spread. Indeed, the former depends only on the temperature forecast itself (and therefore has a link to the true temperature), while the latter depends on all flow variables. Roughly speaking, the poor man's EPS only tells that when a control forecast gives an extreme value, the

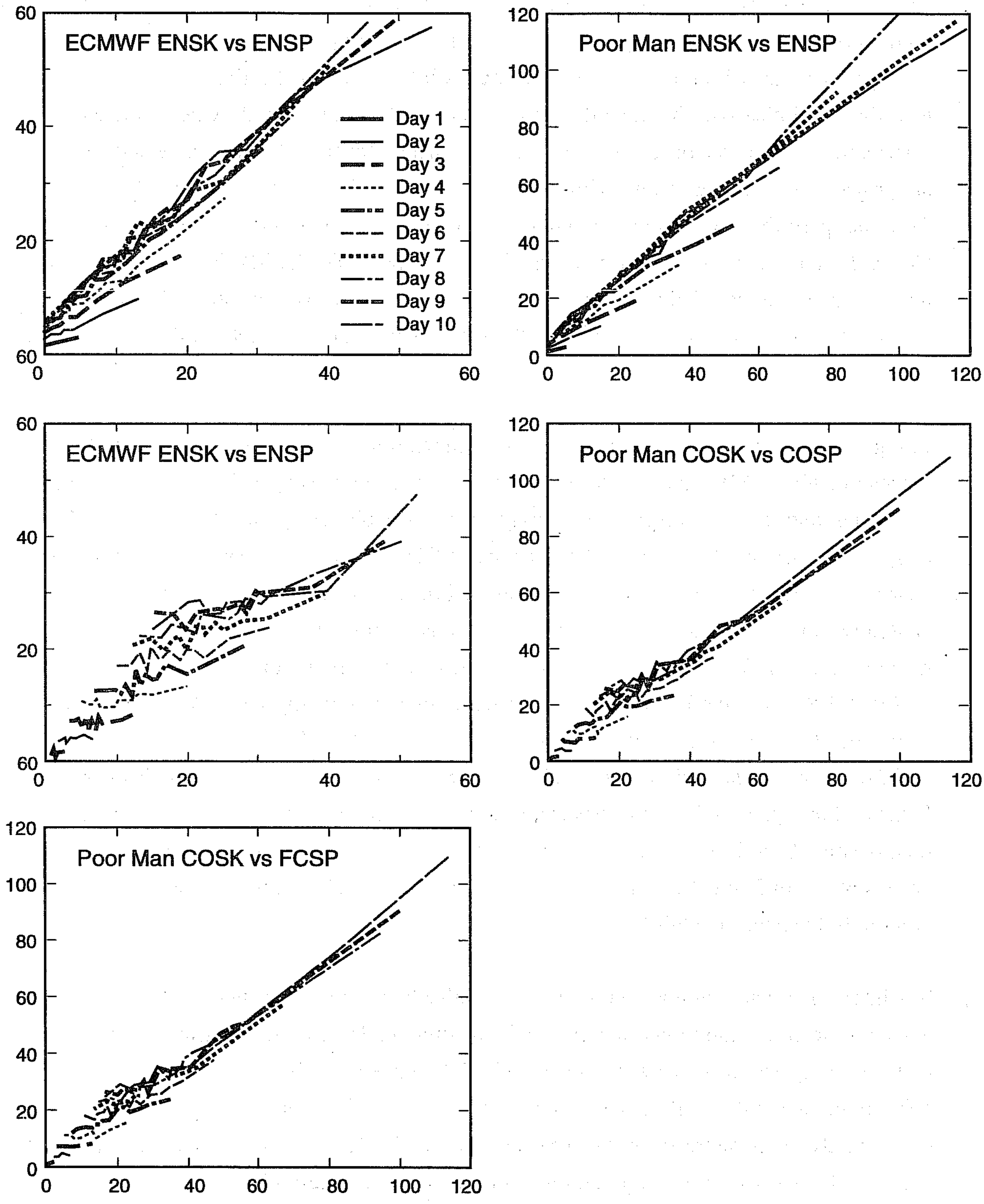


Fig. 4.7. Visualization of various measures of the spread-skill relationship, for both the operational and the poor man's EPSs. Detailed explanations in the text.

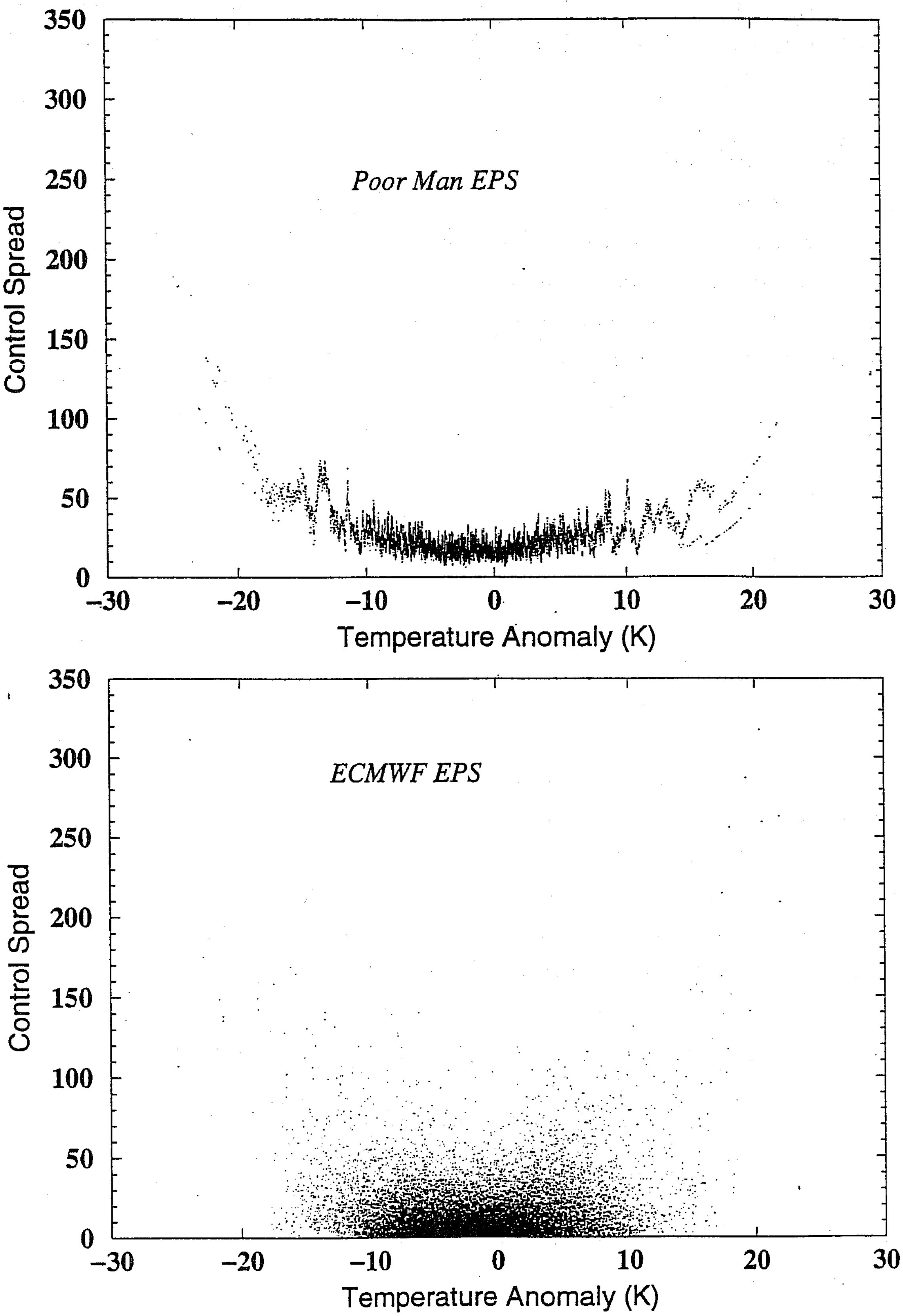


Fig. 4.8. Spread of the forecast ensembles (with respect to the control forecast) as a function of the observed temperature anomaly, for the operational and the poor man's EPSs (bottom and top panels respectively).

uncertainty is large. By contrast, the ECMWF EPS gives a substantial number of large-spread values near the mean and also a number of low-spread values at extreme temperatures.

5. CONCLUSIONS

A number of diagnostics of the performance of PPSs (most of them classical and already widely used) have been considered in this paper. They have been systematically discussed in terms of the degree to which they measure the two qualities that we consider as making the value of a PPS, namely statistical consistency between predicted probabilities and *a posteriori* observations on the one hand, and variability in the predicted probabilities on the other. Concerning probabilistic prediction of occurrence of binary events, the decomposition (2.2a-c) of the Brier score defines a simple and adequate measure of these two qualities. Since statistical consistency can always be achieved by *a posteriori* correction of the predicted probabilities (provided a sample of realizations of the EPS is available that is large enough for ensuring reliable statistics on the performance of the system), the term B_C can be made equal to 0, and the performance of the system is in practice measured by the variability term B_V.

The quality of the prediction of the forecast uncertainty on a given numerical variable or, equivalently, of the prediction of the probability distribution for that variable, has been discussed in both Sections 3 and 4. A number of diagnostics of statistical consistency (histogrammes of the position of the observed value with respect to the predicted ensemble values, as well as several quantitative measures of the agreement between the predicted spread and the observed skill of the forecast) have been considered. One definite advantage of these diagnotics is that, in all cases, what a perfectly statistically consistent PPS would produce is known, so that one knows what to expect from them. As useful as they are, they must however be interpreted with discernment. They are based on integrals computed over all (or at least a large number of) realizations of the EPS, and mutual compensation can occur between individually inconsistent subsets of realizations of the system. In particular, *a posteriori* 'calibration' intended at establishing statistical consistency, as measured by one of these diagnostics, may be illusory. The situation is in this respect totally different for the Brier score, which is the integral of a positive quantity, in which mutual compensation cannot occur. An interesting development will be to extend the Brier score, defined by eq. (2.1) for individual events, to probabilistic prediction of numerical variables. This extension is theoretically possible.

The comparison between the operational and the poor man's EPSs, presented in Section 4, is very instructive. According to the scores used there, the poor man's system performs better than the ECMWF operational system for forecast ranges of up to 2 days. A similar conclusion, based on the comparison of the performance of the ECMWF EPS (and also of the NCEP EPS) with a 'reference pdf' somewhat different from our poor man's scheme, has been reached by Atger (1998). Similar results have also been obtained by Ziehmann (1998). Concerning the ECMWF EPS, a reason for its relatively poorer performance at short range might lie in the fact that the initial perturbations of the ensemble forecasts are chosen along the dominant singular vectors of the flow.

6. REFERENCES

Atger, F, 1998: The skill of Ensemble Prediction Systems, submitted for publication in *Mon. Wea. Rev.*.

Brier, G W, 1950: Verification of forecasts expressed in terms of probability, *Mon. Wea. Rev.*, **78**, 1-3.

Hsu, W, and A H Murphy, 1986: The attributes diagram. A geometrical framework for assessing the quality of probabilistic forecasts, *International Journal of Forecasting*, **2**, 285-293.

Molteni, F, R Buizza, T N Palmer and T Petroliagis, 1996: The ECMWF Ensemble Prediction System: Methodology and validation, *Q. J. R. Meteorol. Soc.*, **122**, 73-119.

Murphy, A H, 1973: A new vector partition of the probability score, *J. Appl. Meteor.*, **12**, 595-600.

Stanski, H R, L J Wilson and W R Burrows, 1989: Survey of common Verification Methods in Meteorology, Research Report 89-5, Service de l'Environnement Atmosphérique, Downsview, Canada, 114 pp..

Toth, Z, and E Kalnay, 1997: Ensemble forecasting at NCEP and the breeding method. *Mon. Wea. Rev*, **125**, 3297-3319.

Wilson, L J, B Strauss and A Lanzinger, 1996: *On the Verification of Weather Element Forecasts from an Ensemble Prediction System*, Preprints, 15th Conference on Weather Analysis and Forecasting, Norfolk, VA, Amer. Meteor. Soc., Boston, MA, pp. J83-J86.

Zhu, Y, G Iyengar, Z Toth, S M Tracton and T Marchok, 1996: *Objective Evaluation of the NCEP Ensemble Forecasting System*, Proceedings, 11th AMS Conference on Numerical Weather Prediction,, Norfolk, Virginia, USA, 19-23 August 1996, J79-J82.

Ziehmann, C, 1998: *Comparison of the ECMWF Ensemble with an Ensemble Consisting of Four Operational Models*, Book of Abstracts, Seventh International Meeting on Statistical Climatology, Whistler, British Columbia, Canada, May 1998, 147.