VALIDATION OF THE ECMWF ENSEMBLE PREDICTION SYSTEM

Bernard Strauss, Andreas Lanzinger
European Centre for Medium Range Weather Forecasts
Shinfield Park, Reading, UK

1. INTRODUCTION

Probabilistic forecasting has been an issue for many years in operational meteorology, and various techniques have been proposed or used to verify probabilistic forecasts (Stanski et al., 1989). However, the recent introduction of ensemble forecasting poses a new challenge, not only for the verification of ensemble forecasts, but also for the validation of an ensemble system. While the validation of a forecast model is reasonably well understood (although there are still problems and pitfalls to be avoided), the validation of an ensemble system requires new concepts which are not yet so well defined.

What should be expected from a perfect ensemble forecast? This fundamental question remains to be clarified. What a perfect deterministic forecast model should provide is clear: it should return a forecast field which is identical to the verifying analysis (at the resolution of the system, and assuming that the analysis is "correct"). What a perfect ensemble forecast should provide can be seen from different aspects. The simplest expectation, most natural but most radical, is that the ensemble should provide the "best" statistical estimation of the truth, without any reference to a control forecast. However, a probably more realistic approach is to refer to the skill of a control forecast, and to state that the ensemble should provide an estimation of the truth with a dispersion similar to the dispersion of that control forecast.

2. OVERALL STATISTICAL PROPERTIES

A valid ensemble must comply with two types of statistical properties. The first type deals with the own characteristics of the ensemble forecasts, regardless of their skill. Basically, the statistical distribution of the ensemble must be the same as the distribution of the analysis. The second type, more stringent, deals with the distribution of the errors of the ensemble forecasts.

In this paragraph, we explore a condition of the first type, following a method suggested by O. Talagrand (personal communication). The distributions of the ensemble members should be compatible with the distribution of the analysis, and this can be investigated by counting the number of occurrences of the observed values in the intervals defined by the single ensemble members, at each grid point for the ensemble forecast to be verified. Given that on the long run the ensemble members should all behave with the same characteristics as the real atmosphere, the frequency of occurrence of all intervals should be the same, including the two extreme intervals outside the entire range of ensemble values. Figure 1 shows such an evaluation in graphical form, for the forecast of geopotential at 500 hPa over Europe at day 1, 2, 3, 4, 6 and 10. The distribution shows an excess in the extreme classes at all ranges, more pronounced in the shorter ranges. The middle interval (number 17) is over-represented at day 1 and 2, due to the symmetric distribution of the initial perturbations.

With an ensemble of 32 members, each class should occur in 100/(32+1) = 3% of the cases (dotted line in figure 1). This applies in particular to the two extreme classes, which means that, even with a "perfect" 32 member ensemble, the analysis will not be captured by the ensemble at about 6% of the grid points.

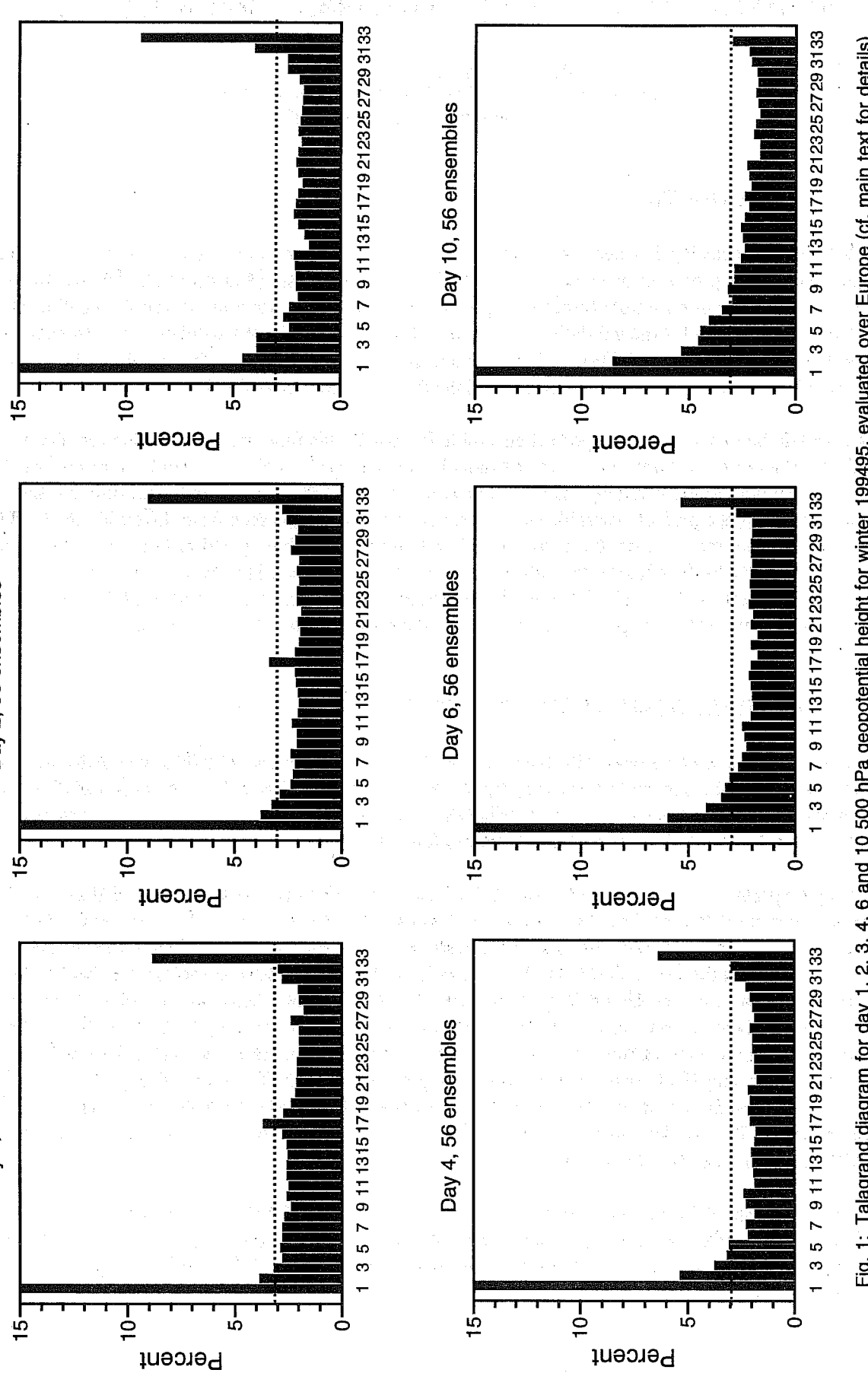

Fig. 1: Talagrand diagram for day 1, 2, 3, 4, 6 and 10 500 hPa geopotential height for winter 199495, evaluated over Europe (cf. main text for details)

It should be noted that the property measured by the Talagrand diagram, like all properties of the first type, is a necessary condition if the ensemble is to be skilful, but it is by no means sufficient. Figure 2 shows a Talagrand diagram over 19 days in September 1994, for an ensemble formed of fields taken at random among analyses in the September months over the past 10 years. The diagram is close to perfect, but obviously this "climate ensemble" yields no skill whatsoever above climatology.

3. SKILL OF THE ENSEMBLE

As suggested in the introduction, there is no well established way of looking at the skill of an ensemble forecast. A first possibility is to consider the ensemble as a collection of individual deterministic forecasts, and to look at the skill of each of them. This is a tempting and apparently natural approach. A second approach is to consider the ensemble as what it truly is, that is, a statistical estimation of the state of the real atmosphere.

In the first approach, one can for example compute the skill of the best and of the worst ensemble members and compare it to the skill of the control forecast. This is shown in figure 3 for each day of February 1995, at day 1 and 6 over Europe. The measure of skill is the root mean square error (RMSE). It can be seen that, in the very short range, the skill of the best ensemble member is practically the same as the skill of control. This should be expected in the linear range with the present system, as the control forecast is computed from the best available analysis. In the medium range, the RMSE of the best ensemble member is around 0.7 the RMSE of control. Expressed in terms of forecast range, the best ensemble member over Europe is roughly as skilful as an average day 4 control forecast. It should also be noted that there are large variations from day to day.

The mean values relative to control for day 1 to day 10 are shown in figure 4. The ratio for the best member seems to tend to an asymptotic value between 0.6 and 0.7. This value is compatible with what can be expected from an ensemble based on a control forecast with a spread around that control forecast similar to its RMSE. A simple example proposed by P. Courtier (personal communication) illustrates this point. Let us consider an ensemble in a one-dimensional space, with two members $\alpha+\sigma$ and $\alpha-\sigma$, where α is a centred symmetric random variable of standard deviation σ and density d, the "analysis" being always equal to 0. The best of the two ensemble members defines a random variable B which is centred. The variance of B is then

$$\sigma_B^{\,2} = \int_{-\infty}^{0} (x+\alpha)^2 d(x)\,dx + \int_{0}^{\infty} (x-\alpha)^2 d(x)\,dx$$

Assuming d normal, this yields

$$\sigma_B^{\,2} = 2\sigma^2 (1 - 2/\sqrt{2\pi})$$

$$\sigma_B \sim \frac{2}{3}\sigma$$

It may seem disappointing that one should not expect the best ensemble member to be much more skilful than the control on average. However, this rather suggests that the first approach is not the correct one, and that an ensemble forecast should not be seen as a collection of conventional forecasts. This has serious implications for the forecasters, because it means that questions such as "which ensemble member should I go for in today's forecast?" are simply not appropriate, and lead nowhere.

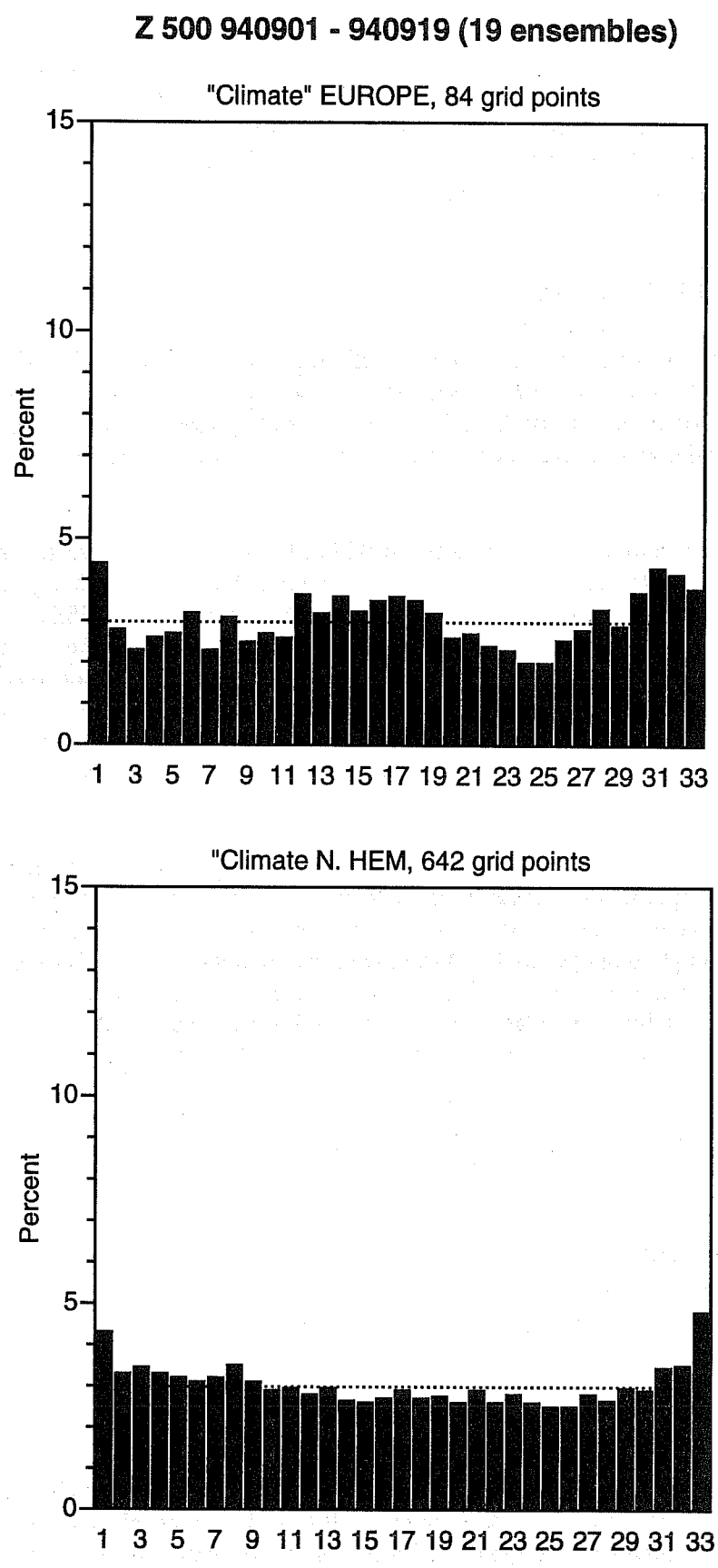

Fig. 2: Talagrand diagram for 19 days in September 1994, for a "climate ensemble" (ensemble formed of real analyses taken at random in previous September months)

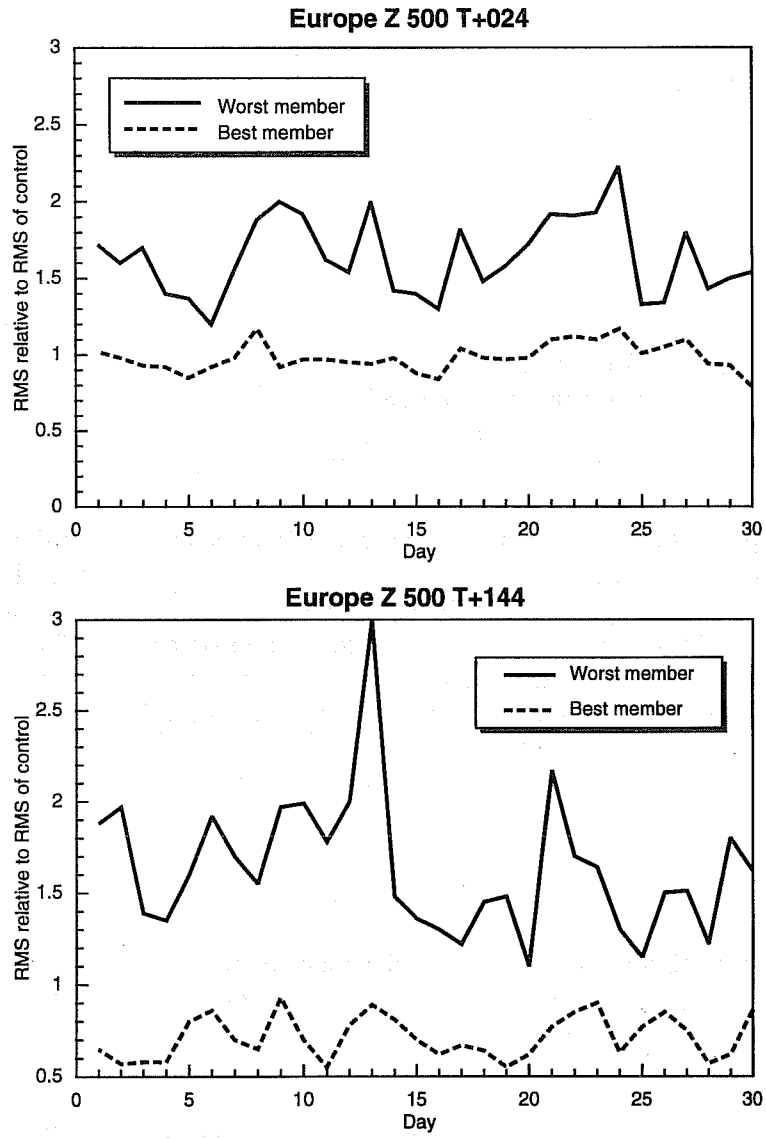

Fig. 3: Time series of RMS error of the best and of the worst ensemble members relative to the RMS error of control, day 1 and day 6, 500 hPa height over Europe, February 1995.

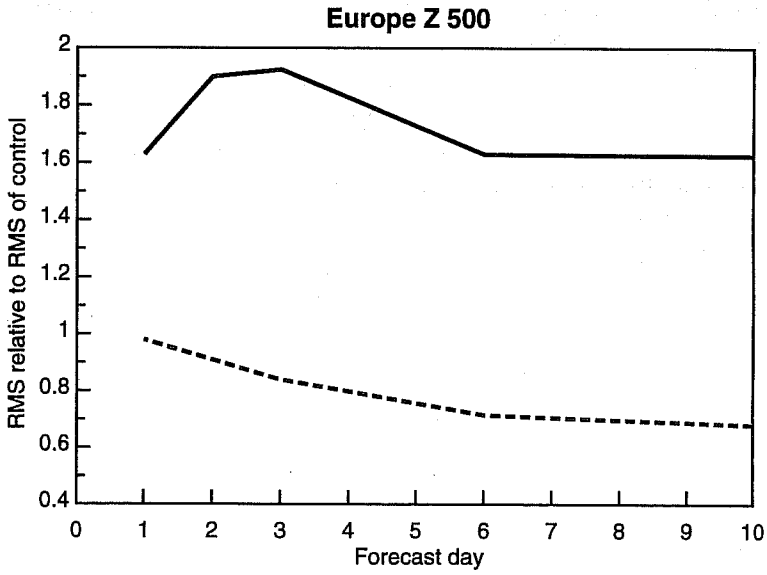

Fig. 4: Mean ratio of the best and worst RMS relative to control, day 1 to 10, 500 hPa height over Europe, February 1995.

The ensemble is a source of probabilistic (statistical) information, and what should be expected from an ensemble system is a statistical estimation of any quantity describing the real atmosphere. The answer to users' questions, including forecasters' questions, will always be of a statistical nature: "the probability of having more than 100 mm precipitation over the coming 7 days is 20%", or "the probability of the trough developing into a cut-off low at day 7 is 75%", etc... The challenge is to get from the ensemble a statistical information which is as sharp and as reliable as possible.

Going back to the question "which ensemble member should I go for in today's forecast?", it supposes that one or some ensemble members should be correct over the whole area of interest and at all forecast ranges. This could perhaps be expected from an extremely large ensemble, but with the present system it is much more realistic and pertinent to expect that some ensemble members will be accurate at some forecast day over some location, and other ensemble members will be accurate at some other time and place, etc..., so that the probabilistic information will indeed be reliable all the time.

Let us look at an example to illustrate the skill of an ensemble forecast in a statistical way. The scatter plots in figure 5 show a verification of various ensemble forecasts valid on 1 April 1995. The range of the predicted ensemble values is plotted against the analysed value for each grid point over Europe. The four scatter plots are for the ensembles at day 1, 6 and 10, and for a "climate ensemble" as used in figure 2, all verifying on the same day. Most of the day 1 grid point ensemble forecasts are highly accurate, i.e., they give a sharp and reliable estimation of the real value. However, at a few grid points the ensemble forecasts are not so sharp, even at day 1. The degradation from day 1 to day 6 and 10 is clearly seen. For that particular day, the ensemble at day 10 is at best marginally more skilful than the "climate ensemble".

4. VERIFICATION OF DERIVED PRODUCTS

One important aspect of the EPS is the forecast of probabilities of meteorological events, e.g. the probability of precipitation to occur in a defined period. Skilful medium-range prediction of probabilities of weather events certainly would be of great value to end users if used adequately in decision making processes.

The two main properties of probabilistic forecasts are reliability and sharpness. Reliability indicates the correspondence between forecast probability and the observed frequency of occurrence of an event. It is best depicted in graphical form, in a so-called reliability diagram, as shown in figure 6. The reliability curve is constructed by splitting the range of forecast probabilities into intervals (or a set of discrete values as usually in the case of subjective probability forecasts), counting the observed occurrences of the event in all forecast probability classes and plotting the relative frequency of occurrence in every class against the interval centre. For perfect reliability the points in this curve lie on the diagonal. Points below the diagonal indicate that probabilities were over-forecast, points above the diagonal mean under-forecasting. The example in figure 6 is the reliability curve for the day 6 forecast of the event 850 hPa temperature anomaly less than -4 degrees, verified against analyses over Europe, for spring 1995. It shows quite good reliability, however, high probabilities were predicted slightly too frequently, whereas low probabilities were under-forecast. In other words, the EPS was slightly too confident in predicting this event.

The histogram plotted next to the reliability diagram in figure 6 shows the relative distribution of forecasts in probability intervals, which reflects the sharpness of the forecast. A probabilistic forecast is sharp if it often predicts probabilities close to 0% or 100%. A reliable probabilistic system which has minimum sharpness will always predict a probability equal to the sample climate frequency of occurrence of the event (which is a non trivial forecast).

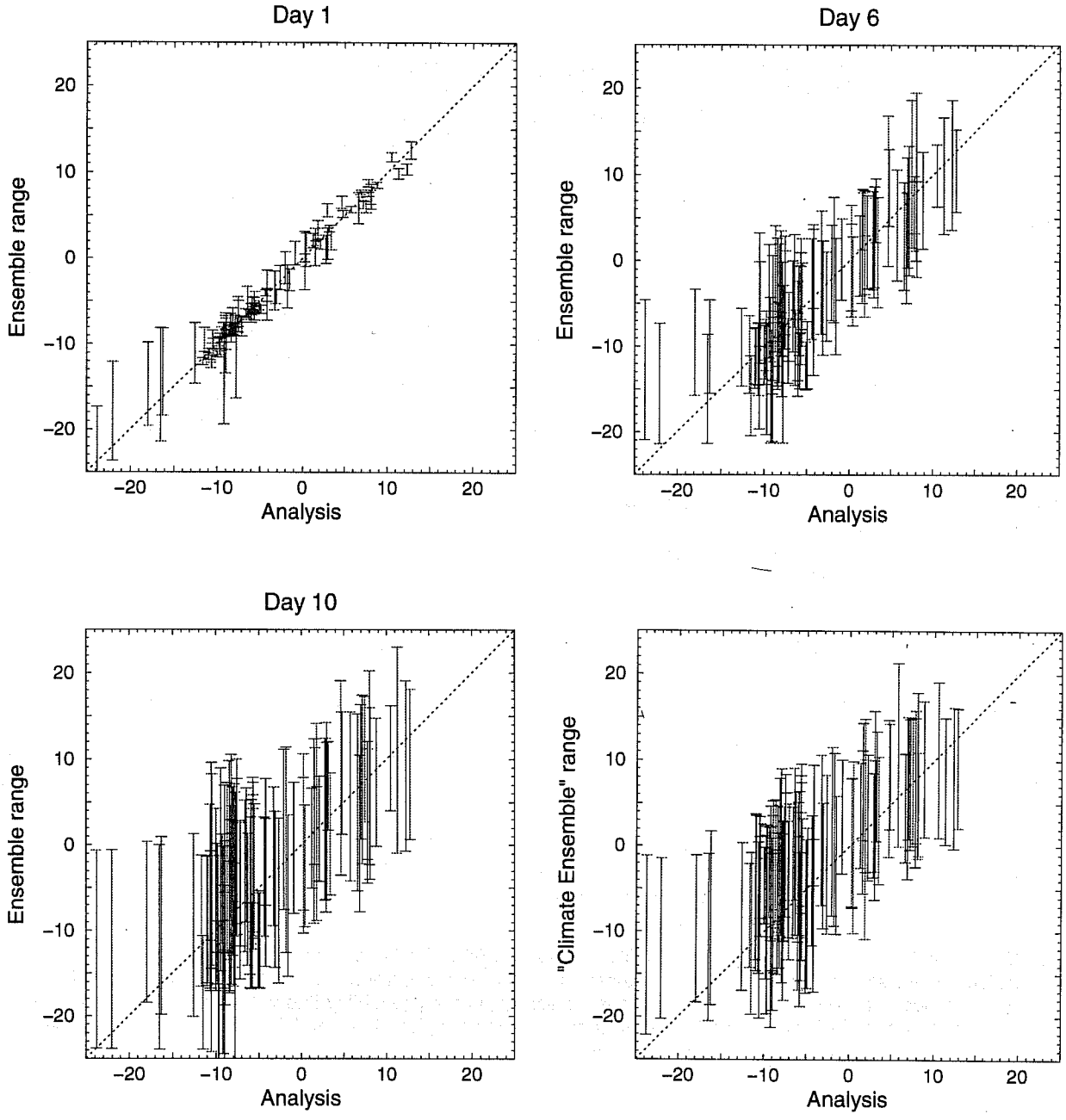

Fig. 5: Scatter plots of analysed values versus ensemble range, 850 hPa temperature over Europe valid on 1 April 1995, at day 1, 6 and 10 and for a "climate ensemble" as used in figure 2.

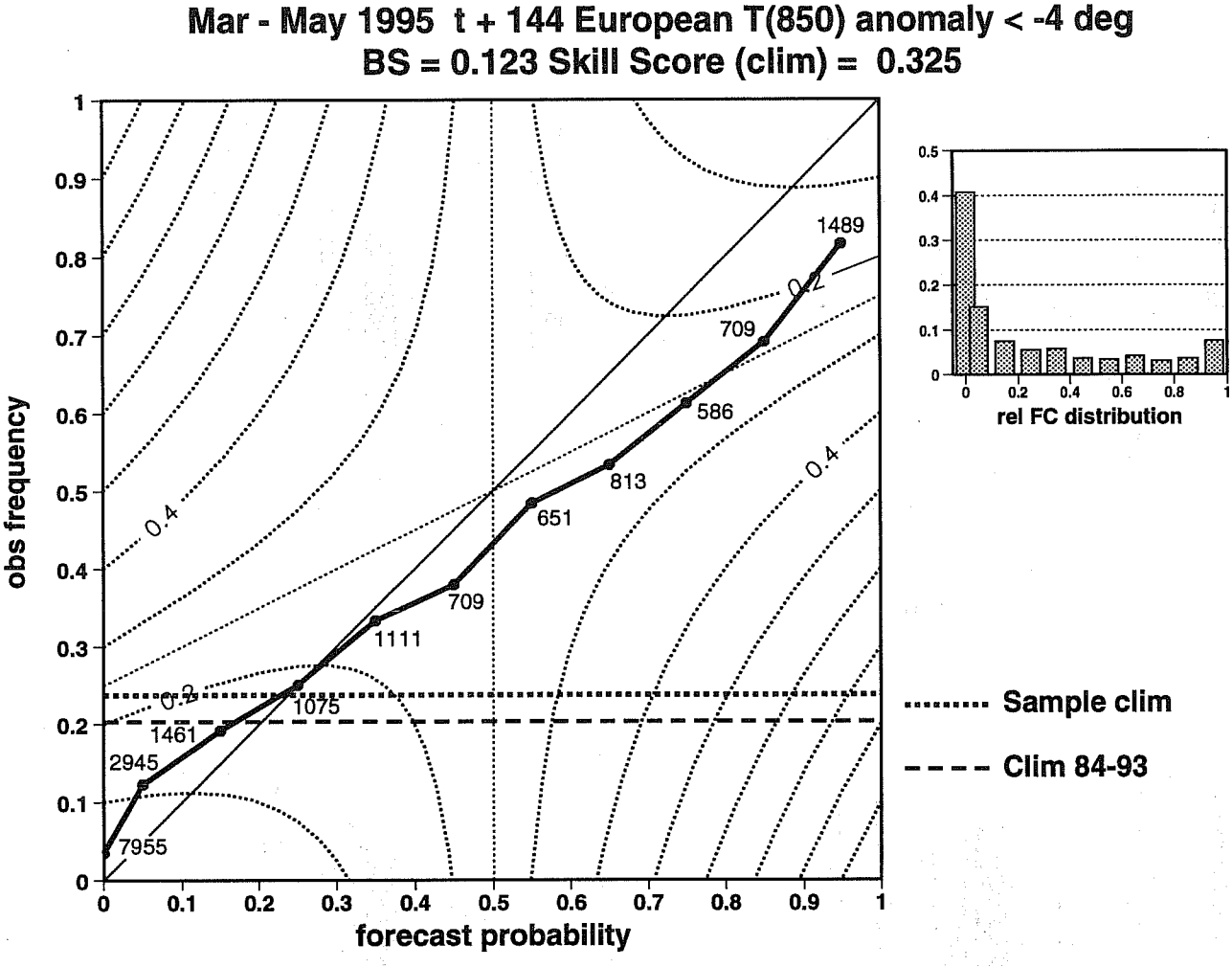

Fig. 6: Reliability diagram for day 6 EPS probability forecasts of 850 hPa temperature cold anomalies of more than 4 degrees for spring 1995. Verification against analysis over the European area. Numbers next to reliability points indicate the absolute number of cases (forecasts) in the probability interval. Horizontal lines denote the levels of sample (dotted) and long term (dashed) climatologies. The small histogram shows the relative distribution of forecasts in probability intervals (sharpness).

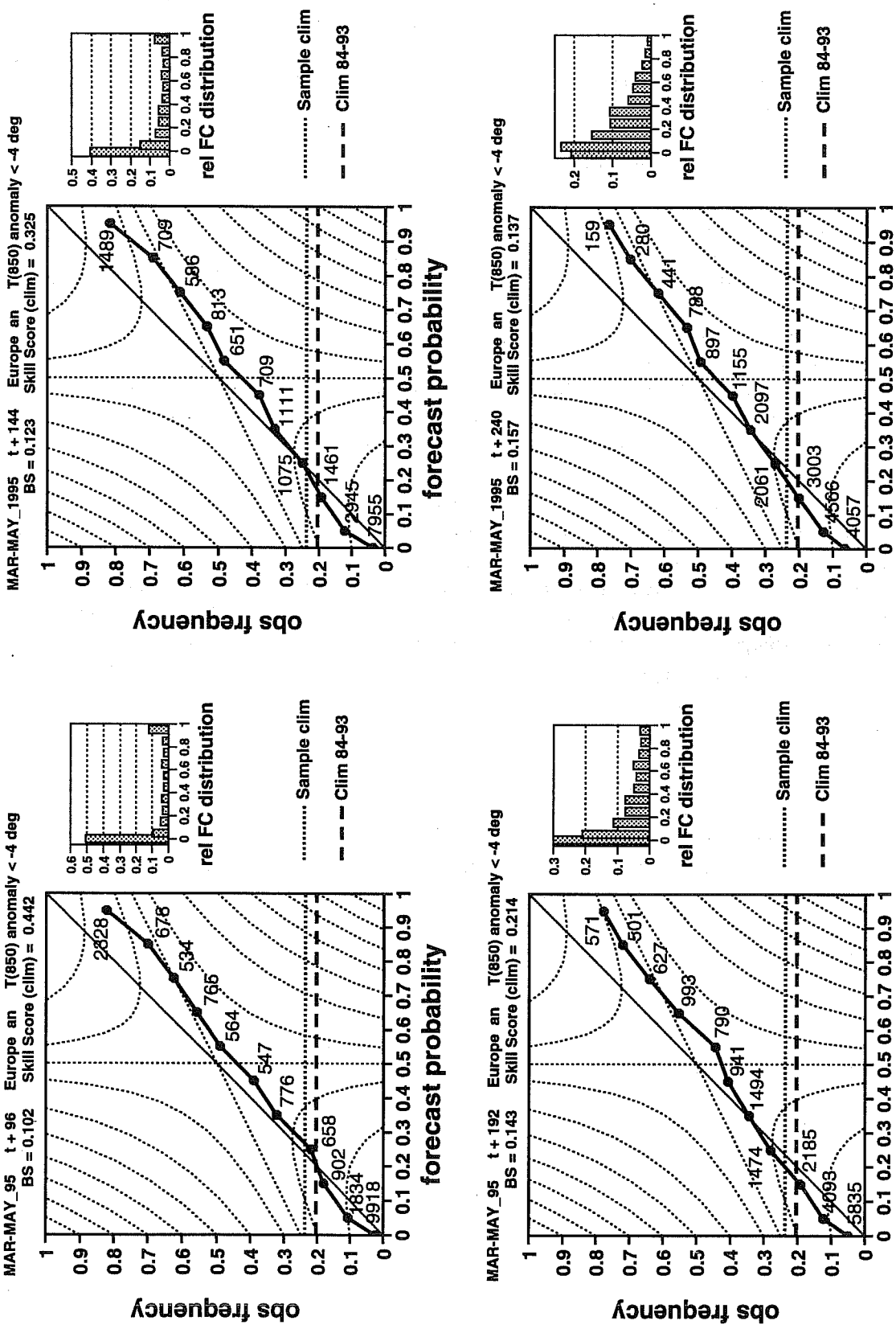

Fig. 7: Reliability diagrams for 850 hPa temperature cold anomalies of more than 4 degrees for spring 1995 over Europe. Top left: day 4; top right: day 6; bottom left: day 8; bottom right day 10 forecast range.

A widely used measure of accuracy of probabilistic forecasts, which takes account of both reliability and sharpness, is the (half) Brier score, BS (Brier, 1950). BS is the mean square error of the probabilistic forecast, whereby the observations have either the value 1.0 for occurrence of the event or 0.0 for non-occurrence. BS does not take into account how close the observed values are to the threshold of the defined event. The score is 0 for a perfect and 1 for the worst possible set of forecasts. Note that these extreme values can only be obtained by a categorical forecast, i e using exclusively probabilities 0 and 1. Since BS is strongly dependent on the sample climatology of the event, comparison between scores from different samples is not very meaningful.

To answer the question of skilfulness of a probabilistic forecast a score obtained by a reference forecast is needed for comparison. A possible reference is a forecast which uses only observed long term climatological frequency as forecast probability. A skill score can be computed which expresses the relative improvement of the forecast against the reference score. Such a skill score is 1 for a perfect forecast, 0 for a probabilistic forecast which is no more accurate than a trivial forecast using long term climatology, and negative for even worse forecasts. In the reliability diagram in figure 6 Brier score and skill score are indicated in the title. BS = 0.123 compares to a climate score BScl = 0.184, yielding a skill score of BSS = 0.325, i.e. a 33% improvement in accuracy against a climatology forecast.

In addition, isolines of Brier score are plotted in the diagram to help understand the relation between the position of points in the reliability curve and the contribution of the sub-samples to the overall score. The minima (good BS) are in the bottom left and top right corners, and relatively low values stretch along the diagonal. This illustrates that high reliability contributes towards good overall scores. However, since the overall score is the sum of sub-sample scores weighted by their relative frequency (depicted in the adjacent histogram), also some sharpness, i.e. relative concentration of forecasts at high and low probabilities, is needed to yield a good score. The degree to which sharpness can be realistically achieved depends on the climatological likelihood of the event. Forecasts for very rare events will naturally concentrate strongly in low probability classes, and generally it is quite hard to achieve high skill, as defined here, for such events.

In figure 7, the evolution of reliability with forecast range, for 850 hPa temperature cold anomalies of more than four degrees, verified against analyses, for spring 1995, is depicted. While reliability remains quite good over the entire forecast range, sharpness (see histograms) declines due to increasing ensemble spread. As discussed above, this leads to a reduction in skill score, despite almost identical reliability curves. Nevertheless, the 10-day scores show that there is still significant advantage of the EPS over a climate estimate of probabilities.

REFERENCES

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

Stanski H. R., L. J. Wilson, W. R. Burrows, 1989: Survey of common verification methods in meteorology. Second edition, published as WMO WWW Rep. No 8.