VERIFICATION OF PROBABILISTIC FORECASTS

Andreas Lanzinger
ZAMG (Austrian Meteorological Service)
Salzburg, Austria

Summary: A brief overview of verification methods for probabilistic forecasts is given. They are routinely used at ECMWF to monitor the quality of probability products from the Ensemble Prediction System (EPS). A comparison of verification results for the periods January to October 1996 and 1997 is presented. It demonstrates the marked improvement of the new EPS, introduced in December 1996, over the old system with fewer perturbed forecasts and lower model resolution. The basic ideas of a new method of scoring the quality of the ensemble distribution, which involves fitting distribution functions to individual sets of ensemble forecast values, are also presented.

1. INTRODUCTION

The number of available probabilistic forecast products from Ensemble Prediction Systems (EPS) is increasing steadily. The assessment of their quality is of importance to the forecast centres providing these products and the users alike. For this purpose a number of more or less standardised verification tools exists. At ECMWF, some of them are used routinely to monitor changes in the quality of operational probability forecast products, to provide information to the member state users, and also to test EPS experiment behaviour.

2. VERIFICATION OF PROBABILITIES

2.1 Overview of methods

According to the attributes or desired qualities of the forecast different verification measures and methods are applied. The *accuracy* of the forecast probabilities, p, of an event to occur is usually measured by the (half) Brier score, BS.

$$BS = 1/N \sum_{i=1}^{N} (p_i - o_i)^2 \qquad (1)$$

o_i is 1 if the event occurred and 0 for non-occurrence. 0 is a perfect Brier score, 1 is the worst possible. Since BS is dependent on the frequency of occurrence of the event in the sample (sample climatology) BS values from different verifications, e. g. for different parameters or event thresholds, can generally not be compared directly with each other.

A skill score, BSS, can be defined which expresses the relative improvement of BS over the score BS_{cl} achieved by a trivial probability forecast, e.g., a constant climatological probability forecast.

$$BSS = 1 - BS / BS_{cl} \qquad (2)$$

The Ranked Probability Score, RPS, measures the accuracy of probability forecasts of ordered multi-category events:

$$RPS = 1/(J-1) \sum_{j=1}^{J} \left[\sum_{m=1}^{j} p_m - \sum_{m=1}^{j} o_m \right]^2 \qquad (3)$$

J is the number of categories. In this form, RPS values range from 0 (perfect) to 1 (worst possible). The RPS credits forecasts which concentrate high probabilities in categories around the observed value. As above, a skill score can be defined for the RPS.

Beside accuracy an important attribute is „*reliability*“. Good reliability means high correspondence between forecast probabilities and observed frequency of an event. This quality is usually examined by means of „reliability diagrams“. Several examples are shown in this paper.

Good reliability is not *per se* sufficient for good skill. Only if the forecasts exhibit „*sharpness*“ as well can high skill scores be achieved. Sharp forecasts have relative high frequencies of probabilities 0 and 1. Generally, short term predictions are sharper (more confident) than longer range ones. Reliability diagrams should always contain complementary information on sharpness.

While reliability diagrams stratify the sample by forecast probabilities, the so called ROC (Relative Operating Characteristics) statistics evaluate the performance from the viewpoint of occurrence or non-occurrence of an event. The question asked here is: What is the characteristic signal from the probability forecasts given the event occurred / did not occur? Or: How well can the forecast discriminate cases of occurrence and non-occurrence? Again, ROC statistics are best depicted as curves, examples can be found below. The area under the curve is a summary measure, its ideal value is 1. As a crude thumb rule, skilful forecasts have an area greater than 0,7.

The *discrimination* of the distributions of forecast probabilities conditional on occurrence and non-occurrence can be measured by the separation of the means of these two distributions. Their curves are sometimes referred to as „likelihood diagrams", which can be displayed as complimentary information with ROC curves. Marked separation of the two distributions show good discrimination ability of the probability forecasts.

A much more detailed discussion and description of these verification methods can be found, e.g., in Stanski et. al. (1989) and Hsu and Murphy (1986).

2.2 Verification results from ECMWF

In the following a few results from direct model output verification of ECMWF probability products from the EPS are presented. The statistics show comparisons between the period January to October 1996 and January to October 1997. The expected improvement of the new EPS with 50 perturbed forecasts and TL159 model resolution (introduced in December 1996) against the old EPS with 32 members and T63 resolution is evident throughout the results.

Fig. 1 shows reliability diagrams and ROC curves for probabilities of 850 hPa temperature (T850) anomalies greater than +4 K at 120 hours forecast time. The verification is against T850 analysis over Europe. Overall, reliability is clearly better for the new EPS: the curve is closer to the diagonal (perfect reliability), at least for probability values from 0.4 upwards. There is only a small improvement in the ROC curve, though. For negative T850 anomalies at the same forecast range (Fig. 2) the ROC curves show much clearer improvement. Reliability is also better, except for probabilities around 0.5. In Fig. 3 the marked improvement of the new EPS in the long medium range (9 days) for positive T850 anomalies is demonstrated.

The Brier skill score for positive and negative T850 anomalies was better throughout the forecast range in the '97 period as compared to '96 (Fig. 4).

The verification of probabilities of precipitation also demonstrates the benefits of the new EPS. In Fig. 5, results for 24-hour precipitation greater than 1 mm are shown for the same period and domain as before. Verification here is against precipitation fields accumulated in the first 24 hours of the EPS control forecast. Since this estimate can sometimes differ from the actual observed distribution even on the resolved scale the verification is not a very strict one from the viewpoint of the end user. Still, the improvements in reliability and ROC are evident, especially for the thresholds 5 mm per 24 hours (Fig. 6) and 10 mm per 24 hours (not shown).

Monthly summary values of reliability (weighted distances of the reliability points from the diagonal) for different precipitation thresholds are depicted in Fig. 7. With the introduction of the new EPS a marked improvement (i.e. drop in reliability value) is noticeable.

3 VERIFICATION OF ENSEMBLE DISTRIBUTION

EPS tries to sample the probability distribution function (PDF) by a relatively small number of perturbed forecasts. For operational ECMWF probability products point probabilities are estimated by the number of ensemble members (incl. control) indicating the occurrence of the defined event over the total number of members. Instead of just utilising these point probabilities PDF's can be constructed from the ensemble values by fitting theoretical distribution functions for single point realisations. These functions can then be verified with respect to individual observations, as proposed by Wilson et. al. (1997). The reader is referred to this paper for a more detailed discussion of the method. Here, only principal ideas are presented.

Fig. 8 shows an example of a distribution fitted to the histogram of actual ensemble values for an individual location (grid point) and forecast step. The functions for the fit are chosen according to the shape of the climatological distribution of the respective parameter. E. g., normal distribution is used for temperature and Gamma distribution for precipitation. The climatological distribution is depicted in Fig. 8 together with the fitted one.

The scoring method is shown in graphical form in Fig. 9. Given the observed value, in this hypothetical case a temperature of -3° C, the score is the probability of a pre-defined interval around this value derived from the fitted PDF (shaded area in the graph).

$$P\,(X_{\mathrm{obs}}\,|\,X_{\mathrm{eps}}) = \int_{X_l}^{X_u} f\,(X_{\mathrm{eps}})\,dX \qquad (4)$$

The selection of the interval can be adjusted to the purposes of the verification. It can vary with observed value, e. g. for precipitation intervals should increase with precipitation rates. The larger the area under the PDF the better the score. The depiction shows that high scores are achieved by accurate placing of the ensemble and small spread at the same time. The ideal score is 1.

A skill score can be computed with respect to a standard distribution, e. g. in the form

$$\text{skill score} = \frac{P\,(X_{\mathrm{obs}}\,|\,X_{\mathrm{eps}}) - P\,(X_{\mathrm{obs}}\,|\,X_{\mathrm{std}})}{1 - P\,(X_{\mathrm{obs}}\,|\,X_{\mathrm{std}})} \qquad (5)$$

The standard distribution can be climatology or persistency, i. e. the climatological distribution of the parameter conditioned on the initial value (control analysis) of the forecast.

4 CONCLUSIONS

A complete set of verification tools is needed to assess the quality of probabilistic forecasts comprehensively. The most important attributes of these forecasts are accuracy (measured by BS, RPS), reliability combined with sharpness (reliability diagrams) and discrimination ability (ROC curves). Verification of one attribute on its own only demonstrates the quality of one aspect of the forecast. Good reliability, e. g., is only one component of a skilful probability forecast. It shows statistical consistency and is therefore a necessary requirement. Perfectly reliable forecasts can still lack skill, as it is the case for a constant sample climatology forecast.

Verification results from January to October 1996 and 1997 demonstrate improvements of the EPS introduced in December 1996. Some of the statistics are presented in this paper. The new operational system has higher model resolution and more ensemble members than the previous one. Both these enhancements have been shown to be beneficial in several case studies and other statistics. Better behaviour of probability forecasts for T850 anomalies are at least partly due to reduced model biases. The improvement of skill of precipitation probability forecasts especially for higher thresholds point to more realistic representation of spatial structures of precipitation.

REFERENCES

Hsu, W., A. H. Murphy, 1986: The Attributes Diagram. A Geometrical Framework for Assessing the Quality of Probability Forecasts. Int. Journ. Forec. 2, pp 285-293.

Stanski, H. R., L. J. Wilson, W. R. Burrows, 1989: Survey of Common Verification Methods in Meteorology. WMO, WWW Technical Rep. No. 8.

Wilson, L. J., W. R. Burrows, A. Lanzinger, 1997: A strategy for Verification of Weather Element Forecasts from an Ensemble Prediction System. Submitted to Mon. Wea. Rev. Sept 1997.

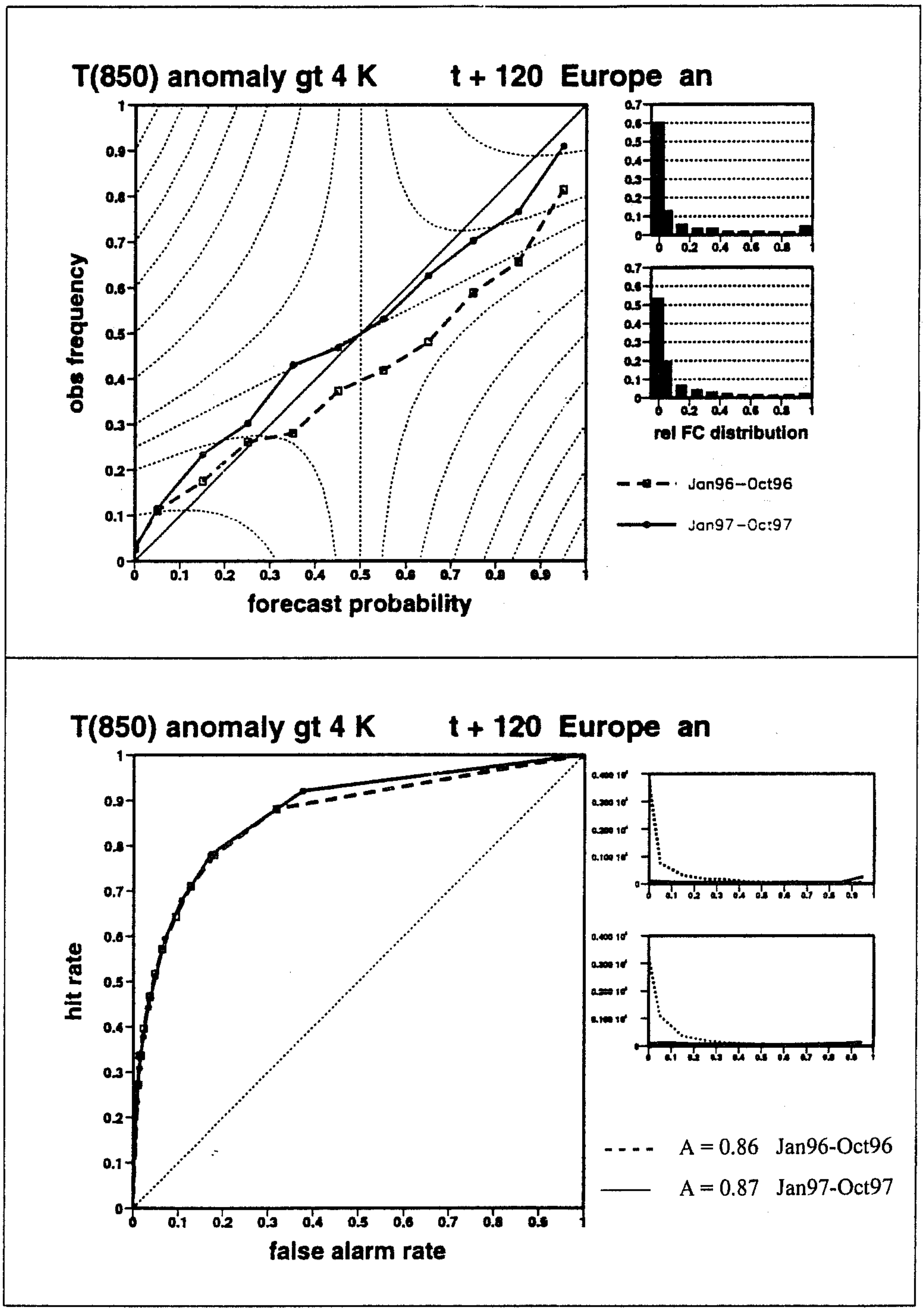

Fig. 1: Reliability diagram (top) and ROC curves (bottom) for T850 anomalies greater than +4 K at forecast step t+120 hours over Europe. Comparison of the periods Jan-Oct 1996 (dashed curves) and Jan-Oct 1997 (solid curves). Histograms on the right hand side of the reliability diagram are relative frequencies of forecast probabilities (sharpness) for the 1996 (top) and 1997 (bottom) periods, respectively. The insets on the right hand side of the ROC curves are distributions of forecast probabilities conditional on occurrence (solid) and non-occurrence (dotted) of the event. Top: Jan-Oct 1996; Bottom: Jan-Oct 1997. The area under the respective curves (A) is given at the bottom right.

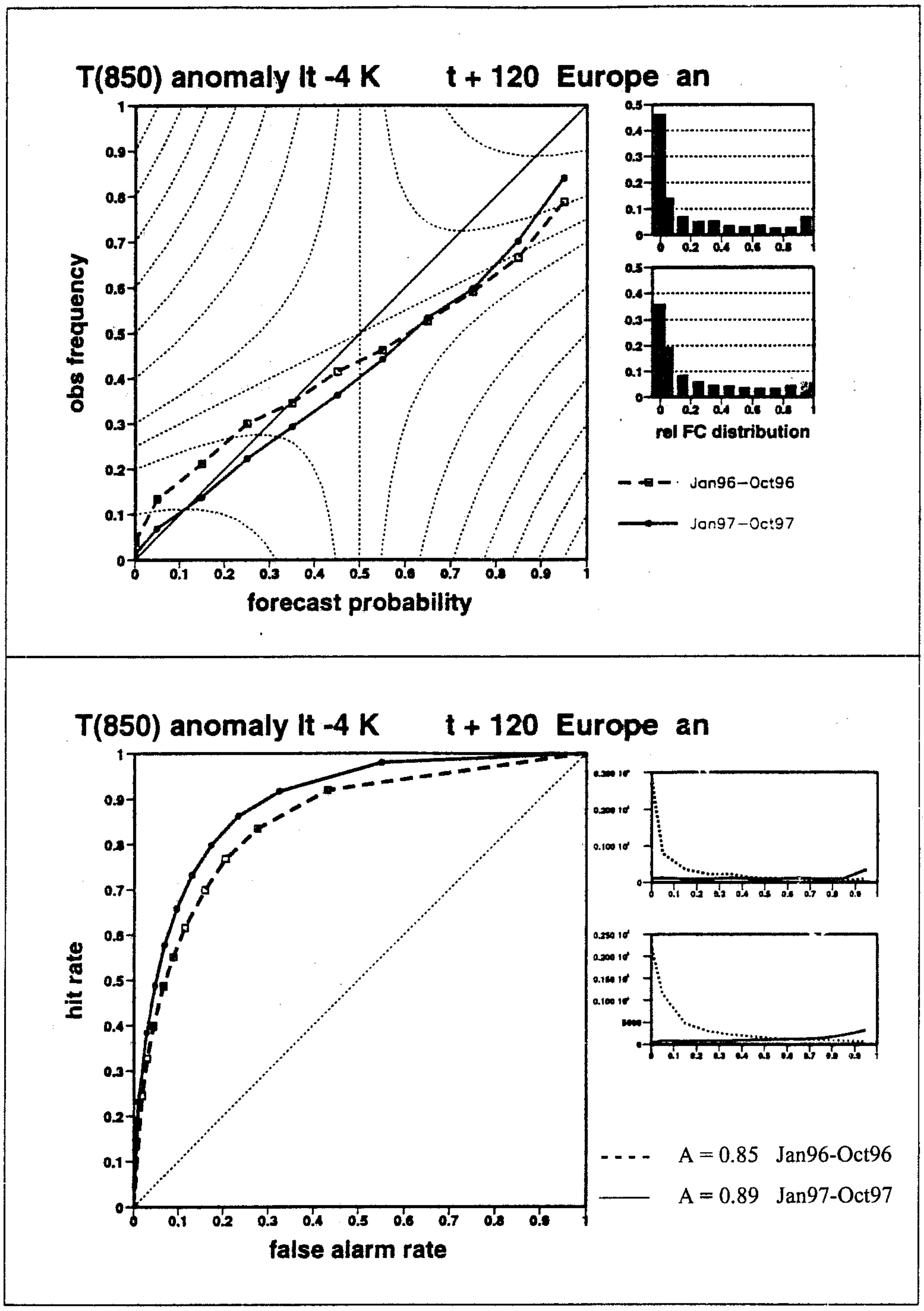

Fig. 2: As Fig. 1, but for T850 anomalies less than -4 K.

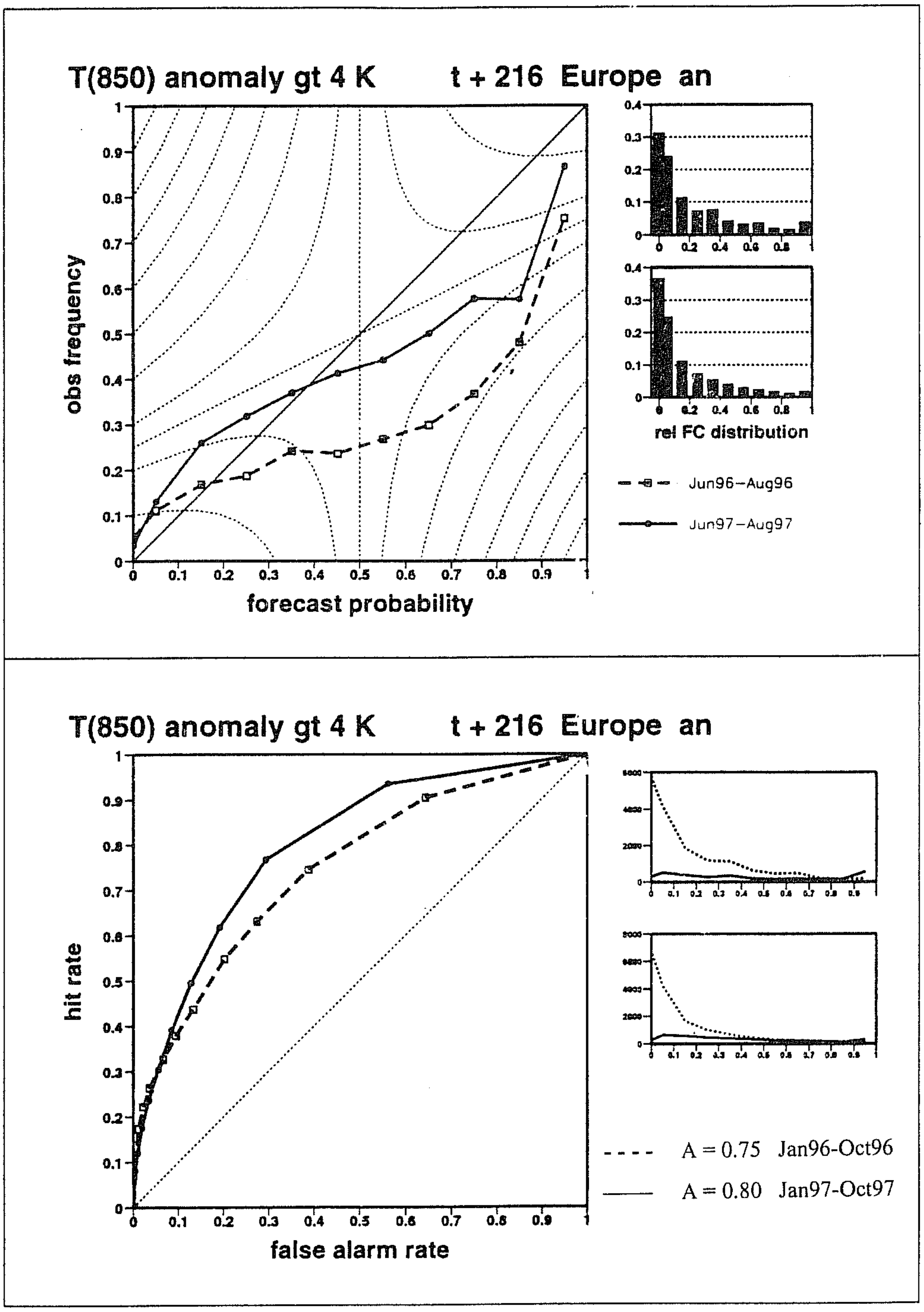

Fig. 3: As Fig. 1, but for forecast step t+216 hours (9 days).

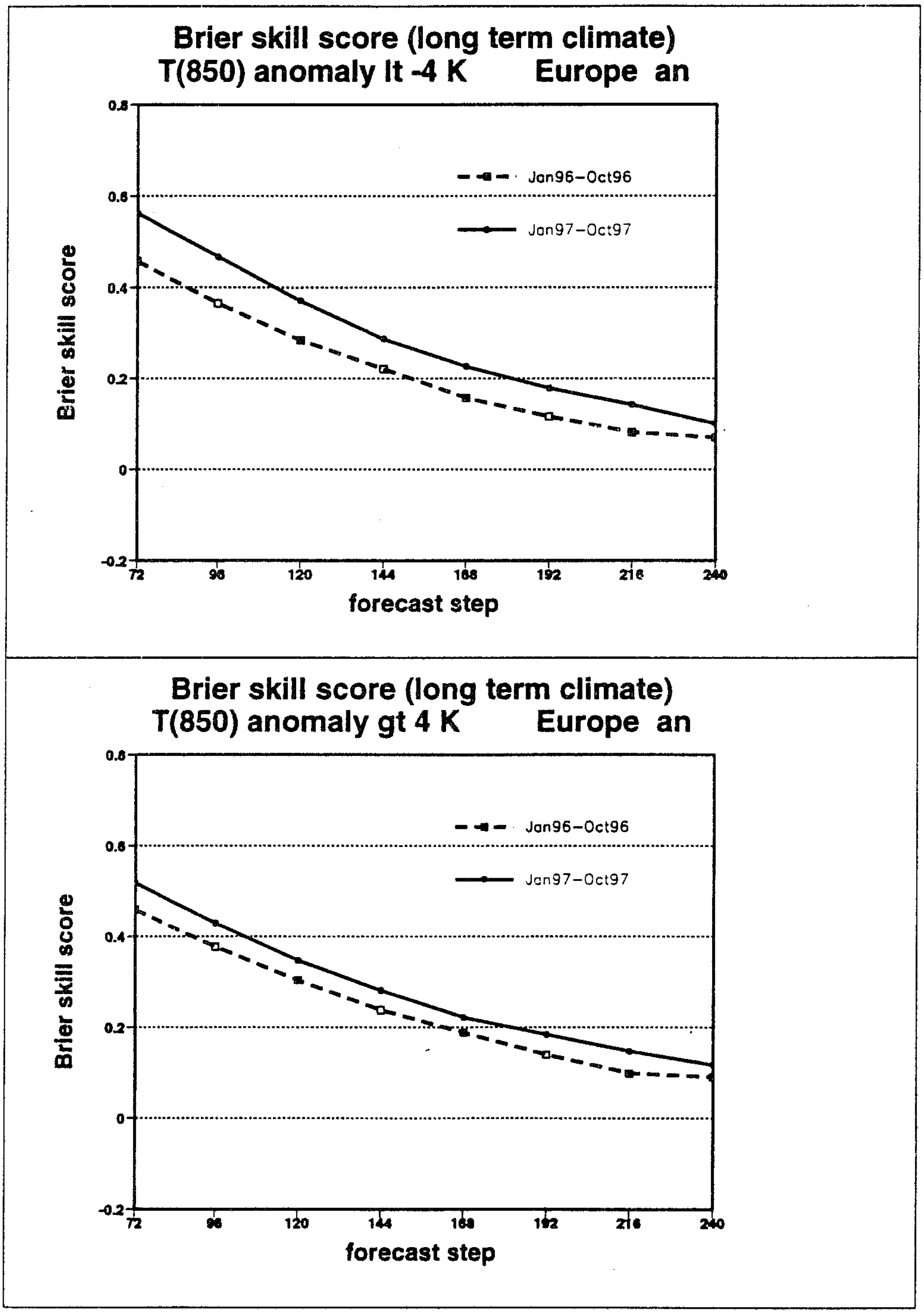

Fig. 4: Brier skill scores against forecast step for T850 anomalies less than -4 K (top) and greater than +4 K (bottom) over Europe. Comparison for periods as in Fig. 1.

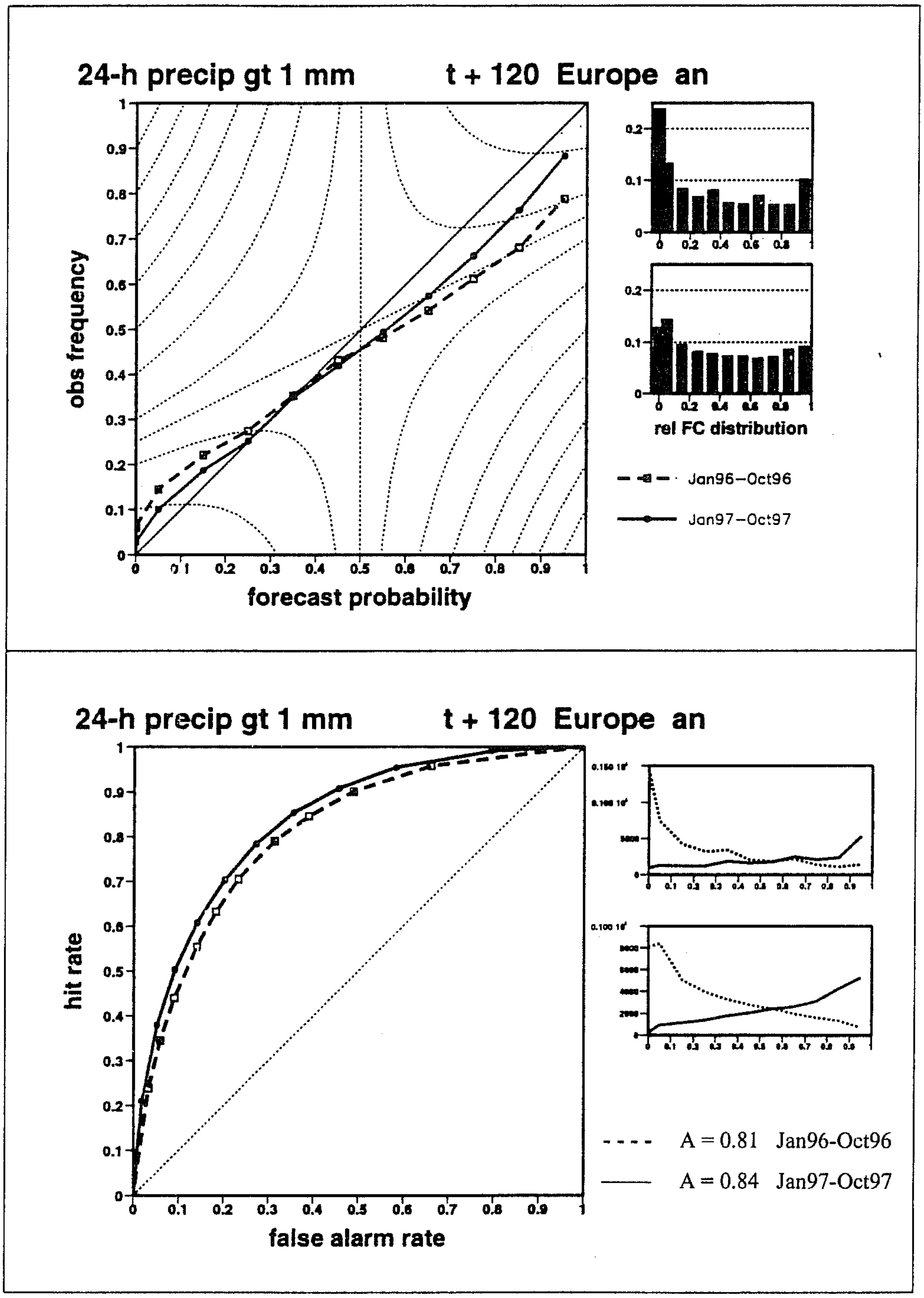

Fig. 5: As Fig. 1, but for probability of precipitation accumulated at forecast day 5 (t+96 to t+120 hours) greater than 1 mm.

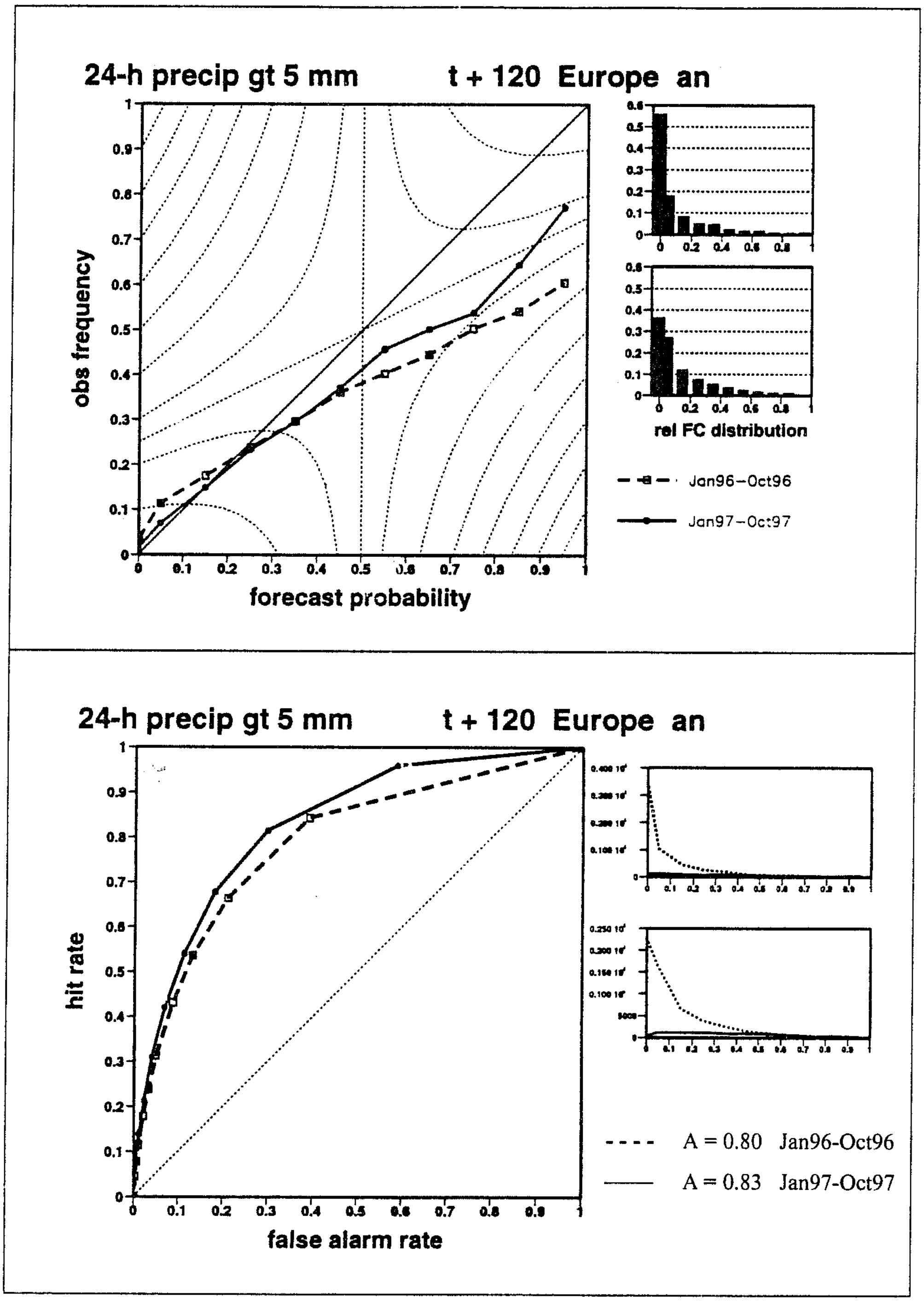

Fig. 6: As Fig. 5, but for 24-hour precipitation greater than 5 mm.

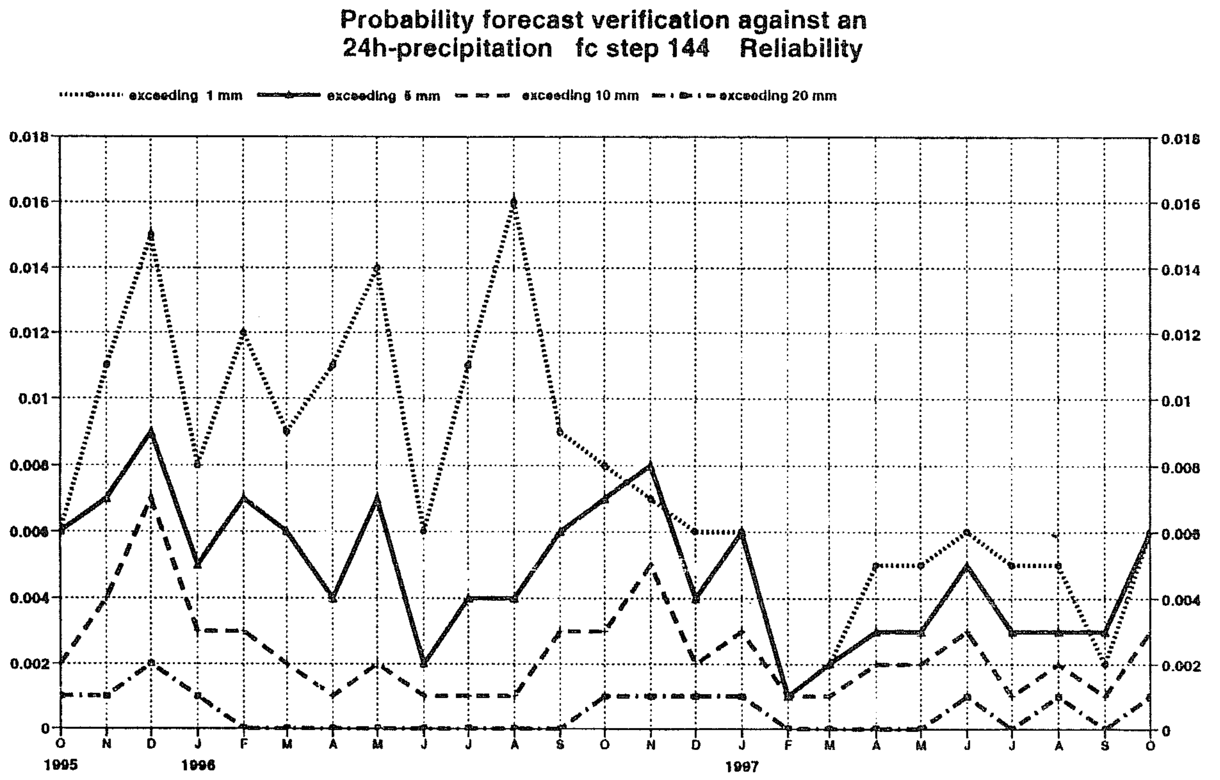

Fig. 7: Time series of monthly summary reliability values for probability of precipitation at forecast day 6 (t+120 to t+144 hours) greater than 1 mm (dotted), 5 mm (solid), 10 mm (dashed) and 20 mm (chain-dashed).

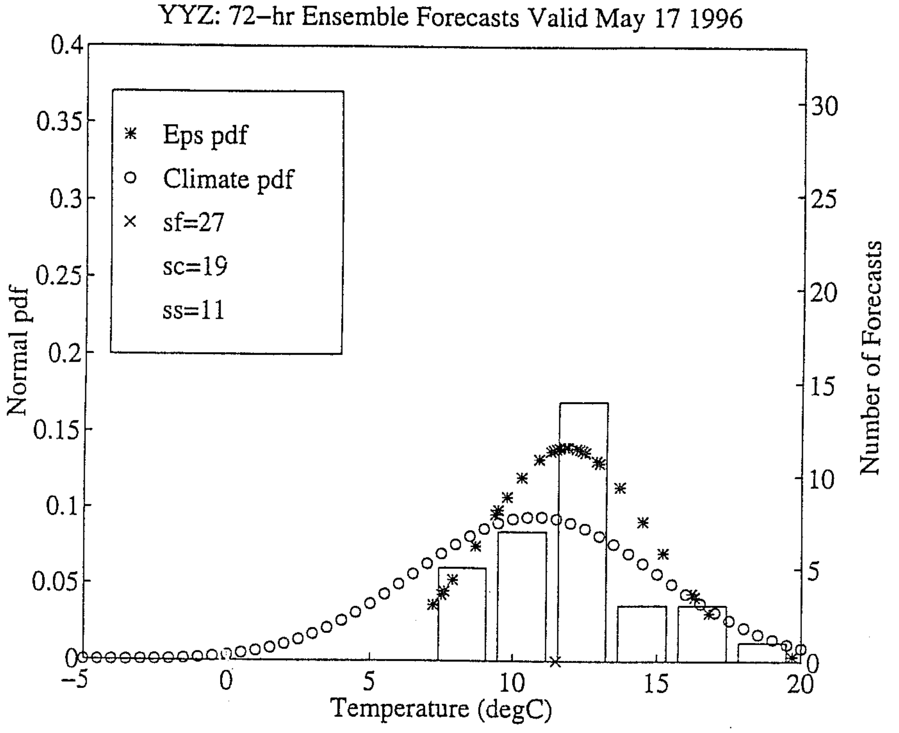

Fig. 8: Example histogram of an ensemble of temperature forecast values at one individual location with a fitted normal distribution (stars) and a climatological distribution (circles). The observed temperature is indicated by an X on the abscissa. In the legend, *sf* is the score for this ensemble forecast example, *sc* is the score for the climate distribution and *ss* is the skill score, each multiplied by 100. (From Wilson et. al., 1997).

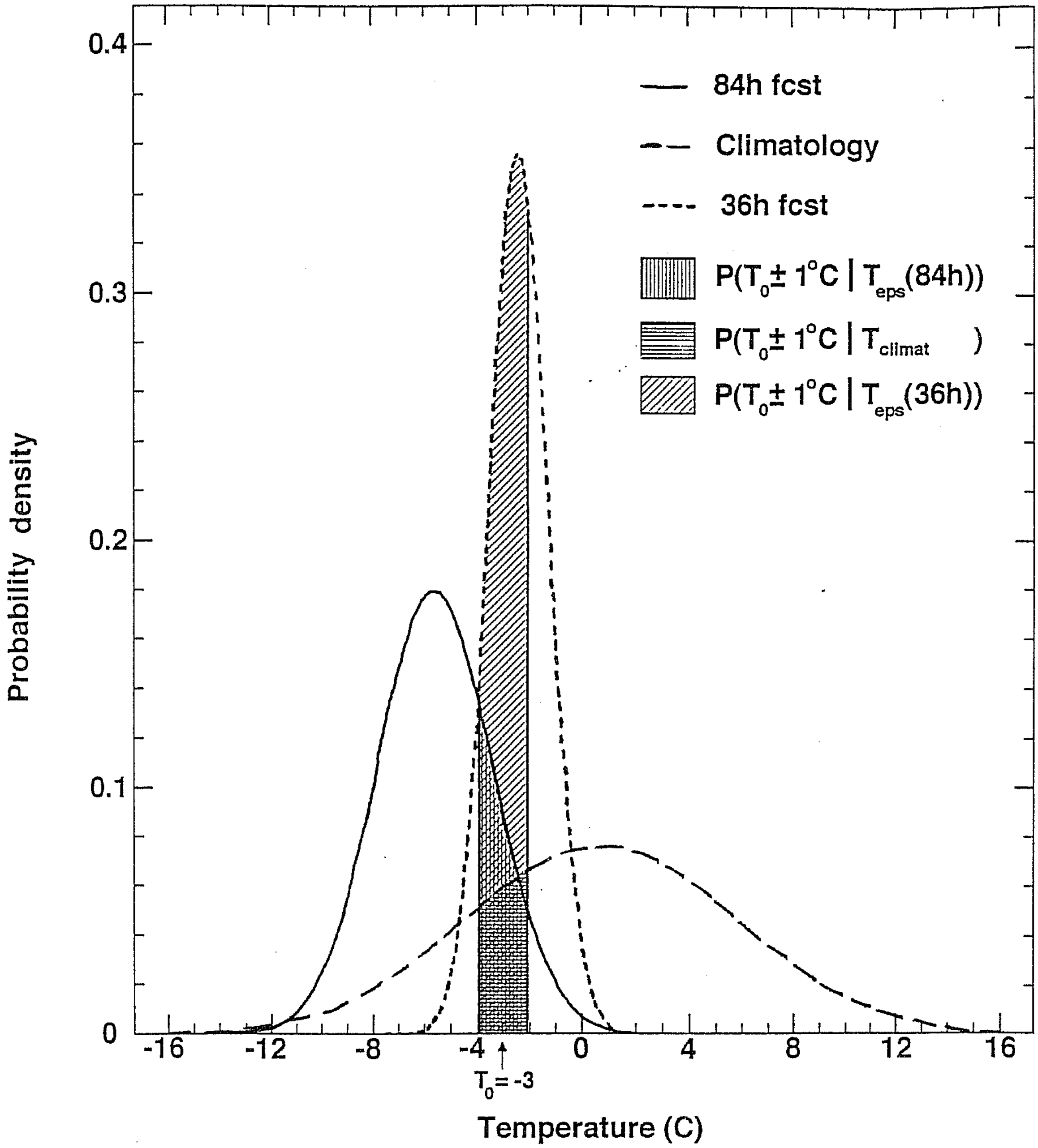

Fig. 9: Schematic illustration of the scoring method. Hypothetical distribution for a 36 hour ensemble forecast (dashed), an 84 hour forecast (solid) and the climatological distribution (long dashes). The hatched areas indicate the scores as denoted in the legend, given the window of +/- 1 degrees around T_o, the observed temperature. (From Wilson et. al., 1997).