EPS EVALUATION AT ECMWF

A. Lanzinger and B. Strauss

ECMWF

Abstract: An overview of operational evaluation of products of the ECMWF Ensemble Prediction System (EPS) is given. In addition to daily monitoring of ensemble spread, clustering and probability products, some objective verification techniques are applied routinely. These include for instance reliability diagrams to evaluate reliability (or calibration) of probability products, as well as "Talagrand" diagrams to evaluate the spread of ensemble forecasts. Some results and interpretation of these statistics will be discussed.

1. INTRODUCTION

A set of 33 ensemble forecasts, including the unperturbed "control" forecast, in T63 spectral resolution, is run at ECMWF on a daily basis since 1 May 1994. The use of EPS products as medium-range forecast guidance is becoming more widespread as experience builds up among forecasters and at ECMWF. Systematic evaluation contributes to this process by assessing the quality and usefulness of direct and derived output products.

2. DAILY MONITORING

The principal EPS output products, like 500 and 1000 hPa height field forecasts and derived cluster and probability maps, are monitored operationally. Emphasis is put on the consistency of EPS spread or the range of suggested flow scenarii with the variety of solutions in different "deterministic" global forecast models.

Internal consistency with previous days' EPS runs is also monitored with interest. A change of the entire ensemble or a large majority of forecasts from one flow configuration to a synoptically distinctly different situation in consecutive forecasts, valid for the same day, is clearly an undesirable behaviour of the EPS. This sort of behaviour has been observed on several occasions in the "early days" of ensemble forecasting. However, some changes in the EPS, like e g the concentration of initial perturbations to the northern hemisphere or the increase of the horizontal resolution of perturbations from T21 to T42, have significantly reduced this "flock of sheep" behaviour. This is shown in Fig 1 which depicts the ratio of the RMS difference between consecutive ensemble forecast means valid on the same day and the previous forecast's spread around the ensemble mean, for periods in early 1993, when perturbations were derived globally (and only three ensemble forecasts were run each week), and 1995, respectively.

3. OBJECTIVE EVALUATION

3.1 Evaluation of probability forecast 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. A range of methods to verify probabilistic forecasts exists, and descriptions of techniques and applications are given, e g, in Murphy and Winkler (1992), Stanski et al (1989) and Wilks (1995).

An important property of probabilistic forecasts is their "reliability". 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 Fig 2. 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 occurrences of the forecasted 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 Fig 2 is the reliability curve for the t+144 hours 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. A summary of reliability statistics of ECMWF probability products will be presented later.

A widely used measure of accuracy of probabilistic forecasts is the Brier score (Brier, 1950):

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

f_i is the forecast probability and o_i equals 1 if the event occurred and 0 if it did not. BS, here in the form of the "half Brier score", is simply the mean square error of the probabilistic forecast. BS does not take into account how close the observed or forecast 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.

It can be shown (see e g Hsu and Murphy, 1986) that BS is dependent of the observed frequency of occurrence of the event in the sample ("sample climatology"), which means that Brier scores with different underlying sample climatologies cannot be directly compared to each other. A possible reference is the Brier score of a trivial forecast which uses only observed climatological frequency as forecast probability. This

reference can be simply obtained:

$$BS_{lcl} = \bar{o}(1-\bar{o}) + (\mu-\bar{o})^2 \qquad (2)$$

The index lcl denotes long term climate, $\bar{o}$ is the sample climatology and μ is an observed long term climatological value of the frequency of the event. The first term on the RHS of equation 2 is the Brier score of a constant forecast with sample climatology $\bar{o}$. With the above reference a skill score may be defined as

$$SS_{BS} = \frac{BS_{lcl} - BS}{BS_{lcl}} \qquad (3)$$

This skill score is 1 for a perfect forecast, 0 for a probabilistic forecast which is no more accurate than a trivial forecast using climatology, and negative for even worse forecasts.

In the reliability diagram in Fig 2 Brier score and skill score are displayed in the title for complementary information. BS = 0.123 compared to a climate score BS_{lcl} = 0.184, yielding a skill score of 0.325.

Furthermore, 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, or relative concentration of forecasts at high and low probabilities, is needed to yield good skill. 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 achive high skill, as defined here, for such events. Further discussion on characteristics and interpretation of reliability diagrams and scores can be found, e g, in Stanski et al (1989).

Figures 3 to 6 show some examples of reliability statistics for a selection of ECMWF probability products. In Figure 3 the evolution of reliability with forecast range, for 850 hPa temperature cold anomalies of more than four degrees, verified angainst analyses, for Spring 1995, is depicted. While reliability remains quite good over the full 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. The 10-day scores show that there is still significant advantage of the EPS over a climate estimate of probabilities. Things are a bit different for warm anomalies, for the same period and forecast ranges (Fig 4). Here,

reliability decreases visibly towards day 10, with a particular increase in over-forecasting of the 0.4 - 0.7 probability range. The reliability curves for day-5 warm anomaly forecasts for different seasons in Fig 5 show quite good results for all seasons, but especially the winter curve reveals significant under-forecasting of low probabilities. Since most forecasts in this sample lie in this range this can be interpreted as being - at least partly - due to a cold model bias. This is also comfirmed by over-forecasting of cold anomalies over the entire range of probabilities above 0.1 in winter 1994/95 (not shown), without "compensating" under-forecasting of very low probabilities.

Finally, reliability curves for precipitation accumulated from t+120 to t+144 above 1 mm, for all seasons, are displayed in Fig 6. Verification here is against SYNOP observations from about 200 stations all over Europe. All curves indicate over-forecasting of high probabilities, least so in winter. Since T63 control forecast verification reveals no significant positive model bias, this behaviour can mainly be attributed to too small ensemble spread, which will be discussed in section 3.3.

Reliability, Brier score and other possible evaluation measures of probabilitic forecasts, like e g likelihood and relative operating characteristics (ROC), are discussed in another presentation during this workshop by L Wilson (1996).

3.2 **Evaluation of cluster reliability**

The population of EPS clusters can be interpreted as forecast probabilities of the flow situation indicated by the cluster mean field to occur. Thus, reliability of the clusters can be evaluated in a similar way as reliability for probabilistic forecasts of certain events.

However, the problem is not so well defined. The first difficulty is how to verify the clusters. A simple approach takes the cluster with the best anomaly correlation (or smallest RMS error) of cluster mean against analysis as the "true" solution, however excluding cases where the best cluster had less than 60 % anomaly correlation. (This - somewhat arbitrary - selection should exclude clusters with little correspondence with the verifying analysis to be regarded as the true solution.) In some cases there might be a discrepancy between the best cluster in objective score terms and from a subjective synoptic point of view. An evaluation of some test cases showed that generally the "ranking" by objective scores makes synoptic sense.

A perhaps more serious point of concern is of statistical nature. Large clusters tend to have smoother mean fields than smaller ones and thus are statistically favoured in terms of RMS error as well as anomaly correlation in medium range predictions.

Fig 7 shows a cluster reliability curve for the period September 1994 to October 1995, with cluster mean scores evaluated at forecast steps t+120, t+144 and t+168 hours. In the x-axis cluster population is plotted instead of forecast probability, and category intervals are 3 members. This is to avoid "quantum leaps" or odd probability intervals, given overall populations of 33 ensemble members. The curve shows quite good reliability. Despite a period of more than one year the sample sizes for the large clusters are still quite small, causing some zig-zag. The good reliability in the smaller clusters indicates that minority solutions in the EPS also verify at a well calibrated rate.

Obviously, this is not an evaluation of reliability of the forecast of certain flow situations. For this purpose fixed clustering, i e clustering into a predefined set of flow patterns, has to be used and verified.

3.3 **Evaluation of ensemble spread**

One of the principal questions of EPS is whether the forecast spread produced by initial perturbations is sufficient to cover the uncertainties in the forecast, due to uncertainties in the initial conditions but also model uncertainties. In other words, are the observed or analysed values falling into the range of predicted values?

One method to investigate this question is to count the number of occurrences of the observed values in intervals defined by the single ensemble members (see Fig 8), for one particular forecast step. From a theoretical point of view the frequencies in all intervals, including the extreme intervals outside the entire range of ensemble values, should be the same for large enough samples, if the chance for the observed value to lie in any of the intervals is the same. Fig 9 shows such an evaluation in graphical form, the so called "Talagrand diagram" (this method was suggested by O. Talagrand, personal communication), for the t+144 hours 500 hPa geopotential height forecasts of winter 1994/95, over the northern hemisphere. The distribution is by no means uniform and shows an excess in the extreme ends, clearly indicating that the ensemble spread is too small to cover all uncertainties. Evaluation for other seasons, domains and forecast ranges (except the very short range) all indicate this behaviour.

In light of these results, is it necessary to increase the ensemble spread, e g by increasing the amplitude of initial perturbations? Cleary this would achieve a flatter Talagrand distribution. However, it should be stressed that this is not a sufficient condition for a good ensemble system, since the "ideal" distribution can also be achieved by evaluating a large enough sample of random (accounting for seasonal characteristics) ensemble forecasts. Just increasing ensemble spread towards a climatological distribution would take away necessary sharpness from the EPS (see section 3.1). As discussed by L Wilson (1996) during this workshop, the reasons for observed values being outside the predicted range are too small spread around the ensemble

mean as well as errors in the position of the ensemble mean. An improvement in this position is certainly preferable to a spread increase, however it is also the much more difficult option since it means in practice general analysis, model and ensemble sytem related improvements. These statements also relate to the "flock of sheep" behaviour, discussed in section 2.

4. CONCLUSIONS

Only some aspects of evaluation of the Ensemble Forecasting System are discussed in this paper. They refer mainly to verification of output products. Further extension of operational evaluation is planned and being developped.

Results on probability products for weather events show some degree of skill, however some aspects are still unsatisfactory. Reliability varies for different parameters and thresholds, and has also seasonal dependency. A common feature in practically all reliability statistics is the over-forecasting of high probabilities. This can only partly be explained by systematic model errors and only for certain parameters at certain seasons, the more prominent reason being a lack of ensemble spread in conjunction with ensemble mean errors.

This conclusion can also be drawn from examinations of the spread of 500 hPa height forecasts in relation to the occurrence of observed values (Talagrand diagrams). Furthermore, the "flock of sheep" behaviour in some cases, i e too high inconsistency of consecutive ensemble forecast relative to their spread of solutions, hints at this problem. There is, however, evidence that improvements in this respect have been achieved in the relatively short time of the existence of EPS.

ACKNOWLEDGEMENTS

Thanks to Anders Persson for his contributions on EPS inconsistency and to Pascal Mailier for his work to compute probability verification statistics.

REFERENCES

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

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

Murphy A. H., R. L. Winkler, 1992: Diagnostic verification of probability forecasts. International Journal of Forecasting 7, 435-455.

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.

Wilks D. S., 1995: Statistical Methods in Atmospheric Sciences. New York, Academic Press, 464 pp.

Wilson L.J., 1996: Verification of Weather Element Forecasts from an Ensemble Prediction System. Proceedings ECMWF Workshop on Meteorological Operational Systems, Nov 1995.

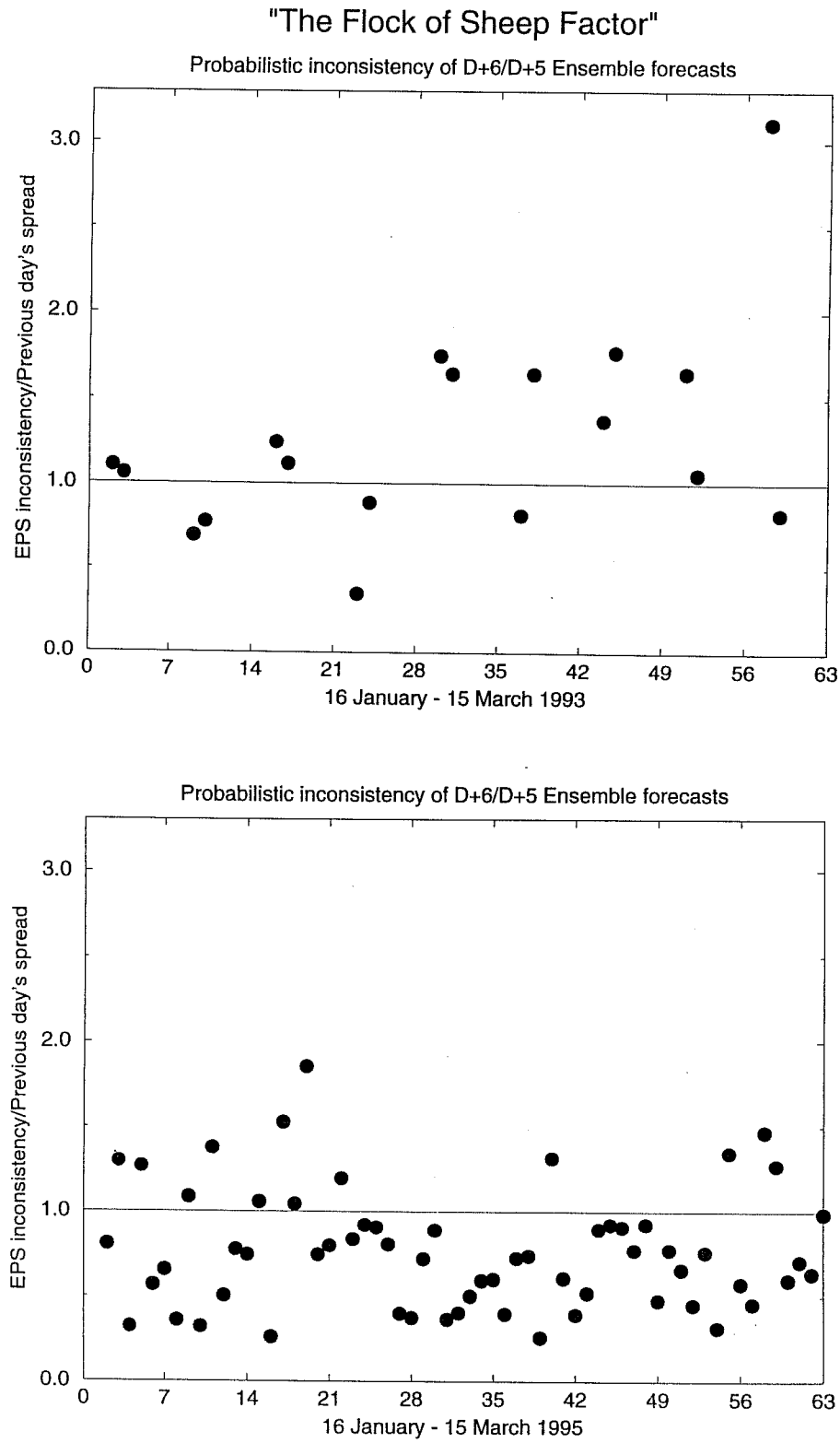

Fig 1: Time series of the "flock of sheep" factor (see text) for the periods 16 January - 15 March 1993 (top) and 1995 (bottom). Courtesy of Anders Persson.

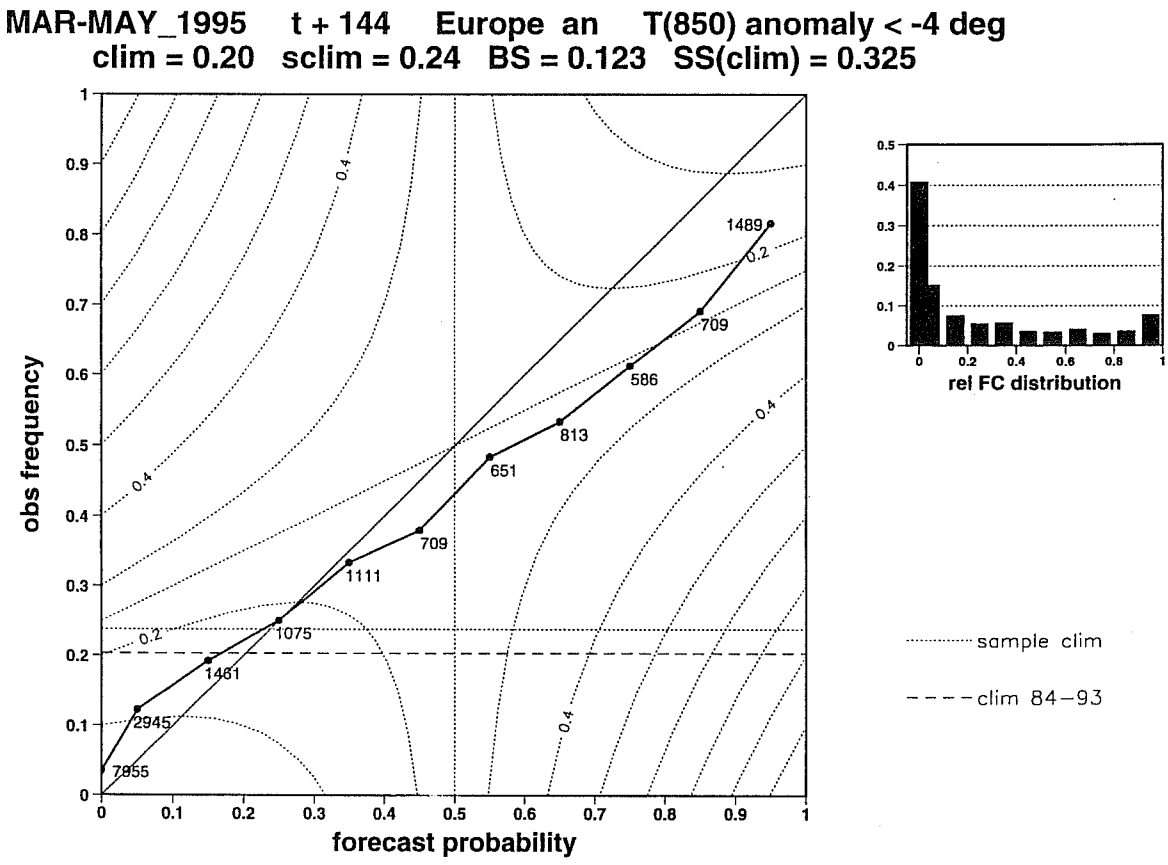

Fig 2: Reliability diagram for t+144 hours EPS probability forecasts of 850 hPa temperature cold anomalies of more than 4 degrees for spring 1995. Verification against analysis over the European area. Numbers next to reliability points indicate the absolute number of cases (forecasts) in the probability 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).

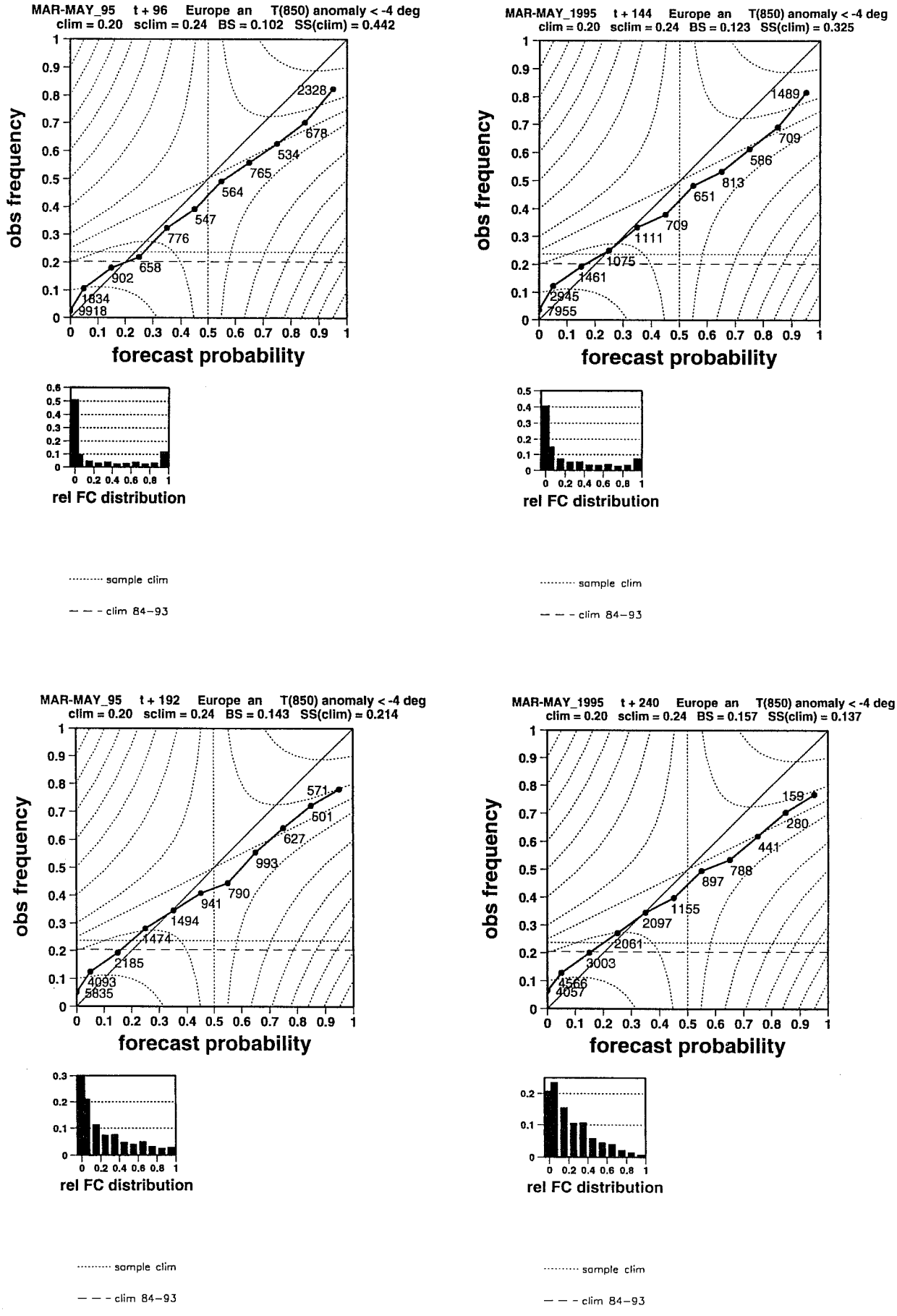

Fig 3: Reliability diagrams for 850 hPa temperature cold anomalies of more than 4 degrees for spring 1995 over Europe. Top left: t+96; top right: t+144; bottom left: t+192; bottom right t+240 hours forecast range.

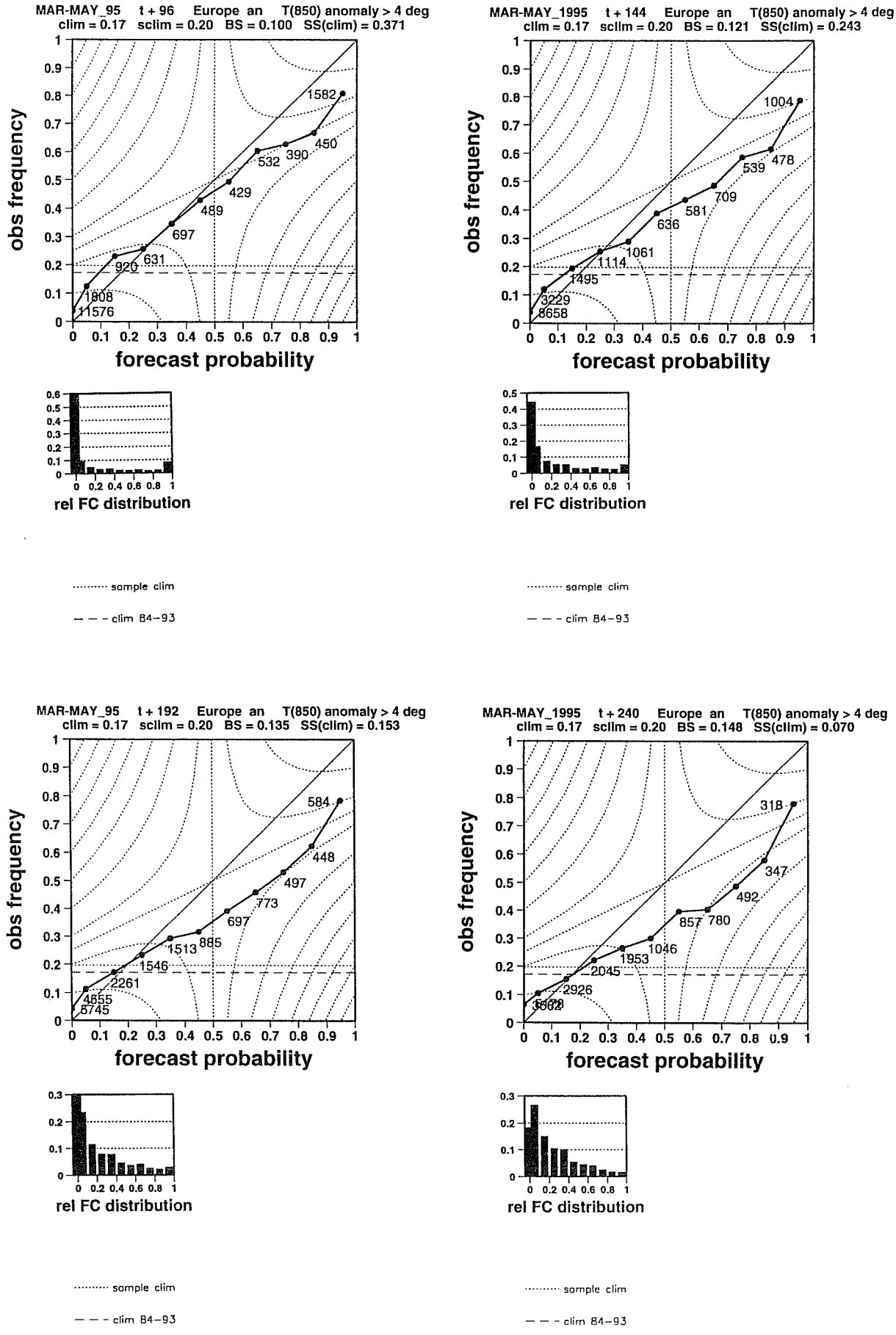

Fig 4: As Fig 3, but for 850 hPa temperature warm anomalies of more than 4 degrees.

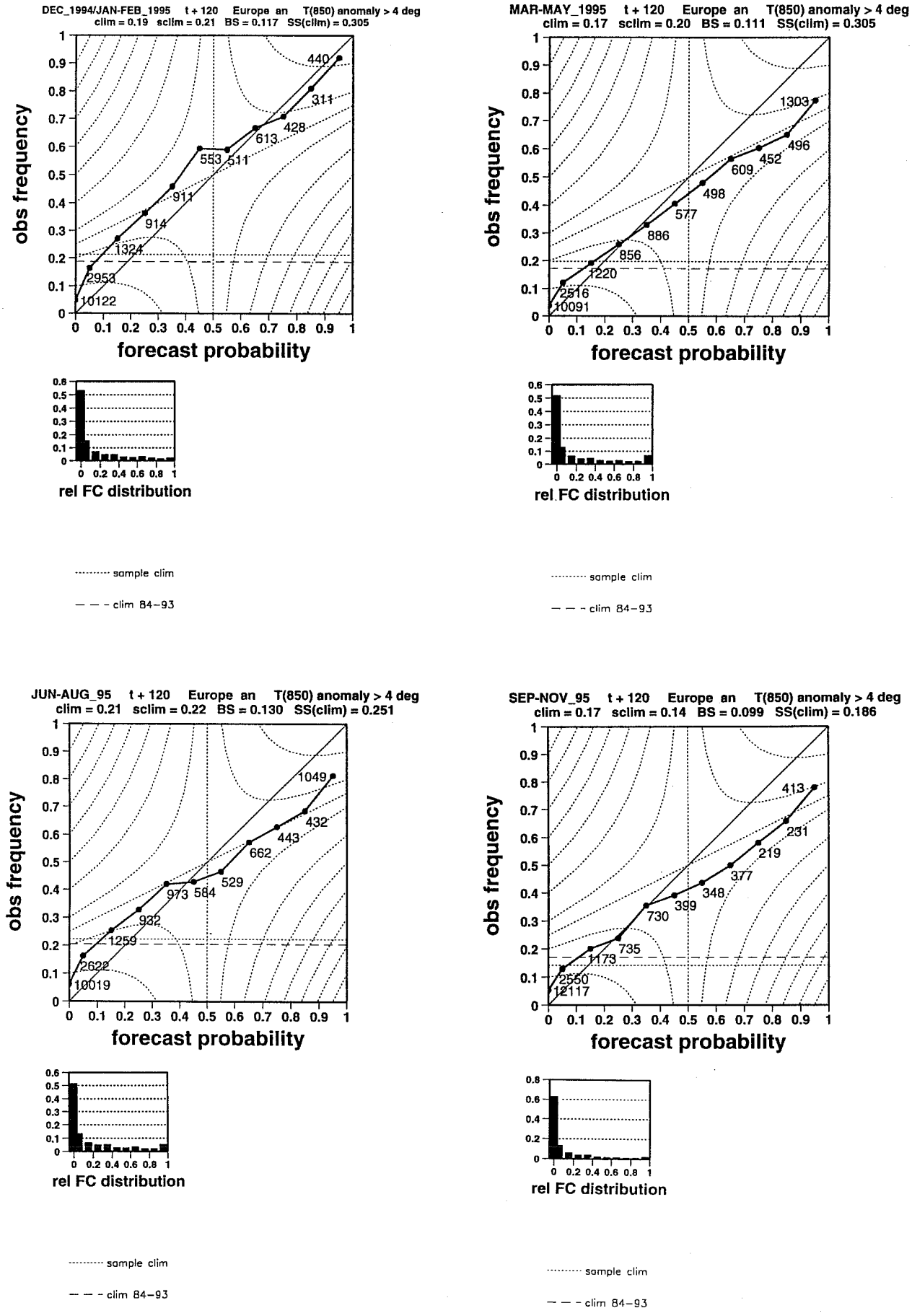

Fig 5: Reliability diagrams for 850 hPa temperature warm anomalies of more than 4 degrees for t+120 hours over Europe. Top left: winter 1994/95; top right: spring 1995; bottom left: summer 1995; bottom right: autumn 1995.

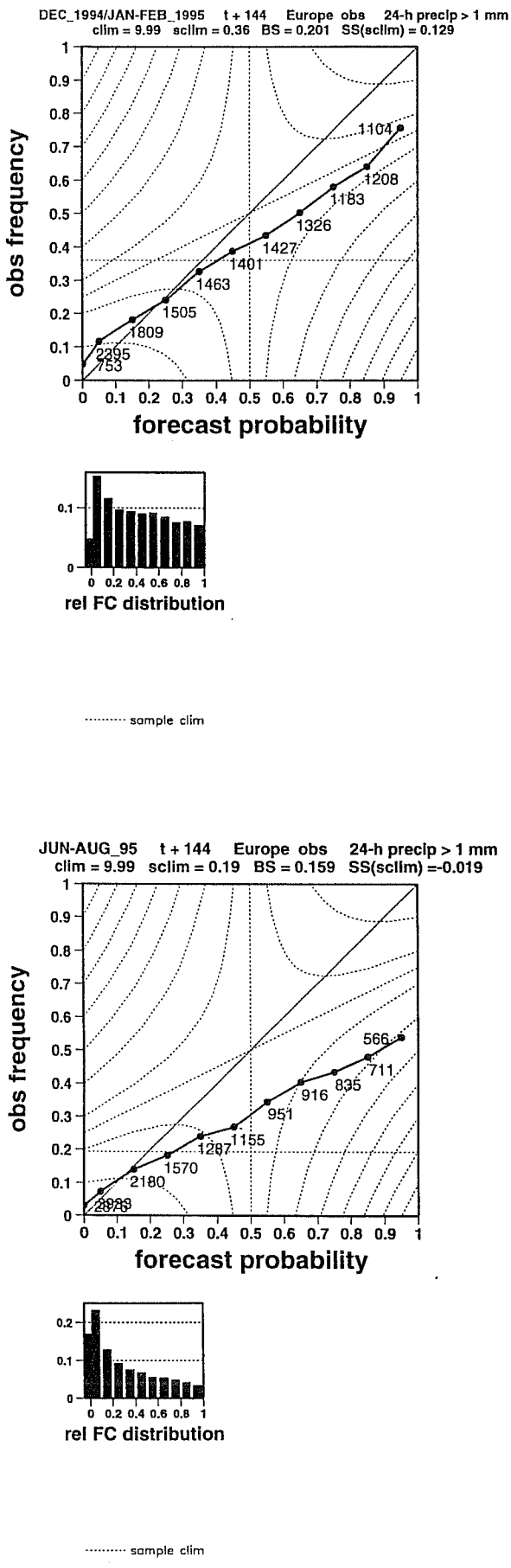

Fig 6: Reliability diagrams for precipitation, accumulated between t+120 and t+144 hours, greater than 1 mm. Verification against observations from about 200 SYNOP stations in Europe. Top left: winter 1994/95; top right: spring 1995; bottom left: summer 1995; bottom right: autumn 1995. Note that no long term climatology is yet available for precipitation observations and thus the skill score, SS(sclim), in the title text has sample climate reference.

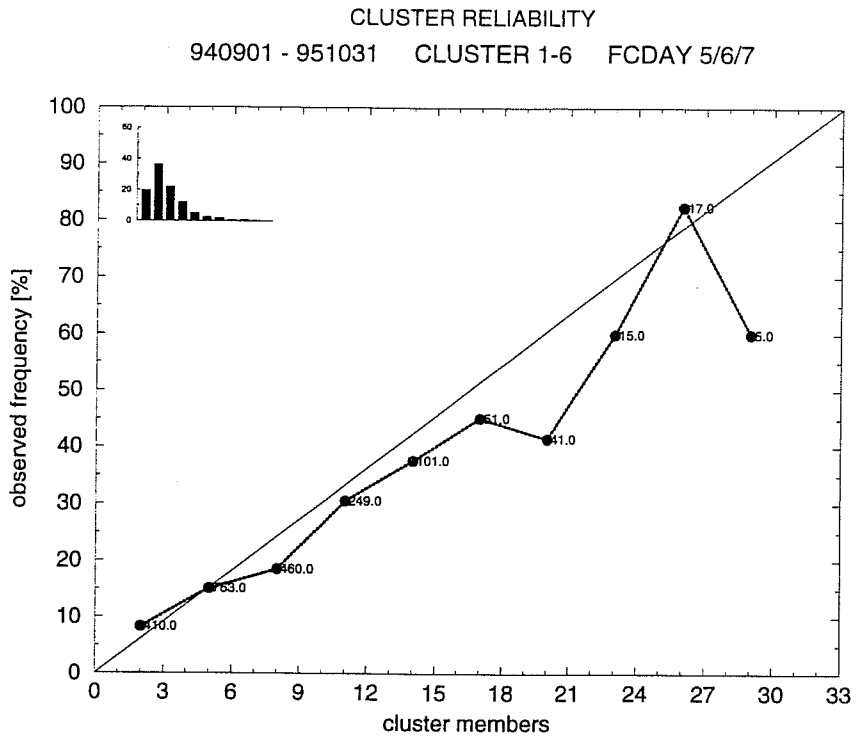

Fig 7: Reliability diagram for synoptic clusters (see text), for the period September 1994 to October 1995.

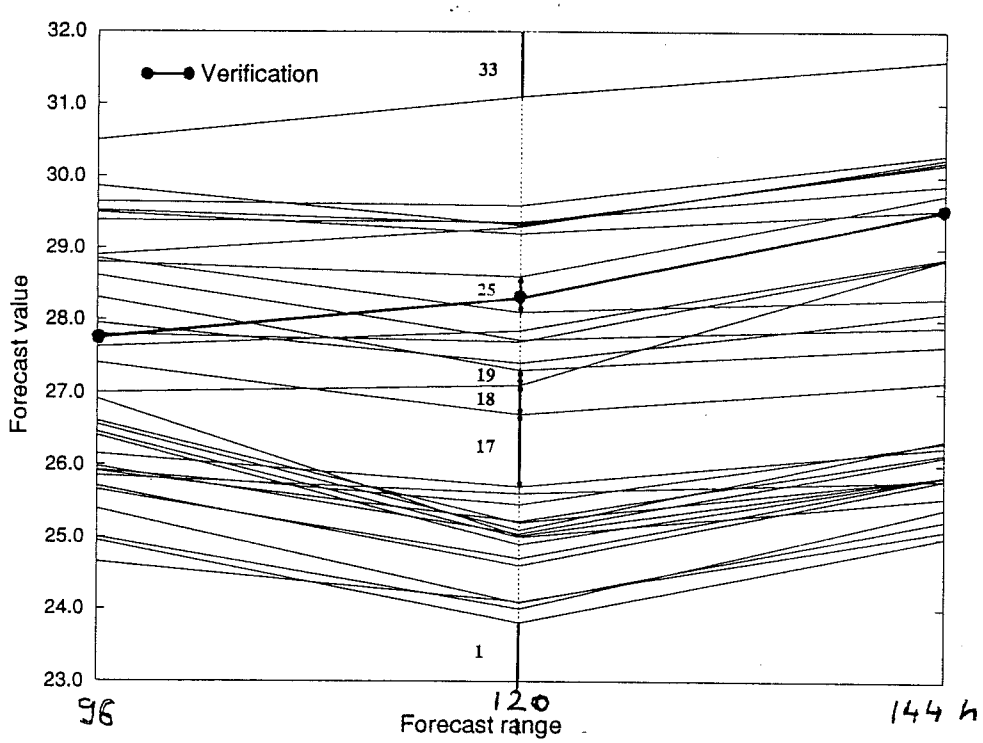

Fig 8: Schematic diagram on the definition of intervals for Talagrand statistics.

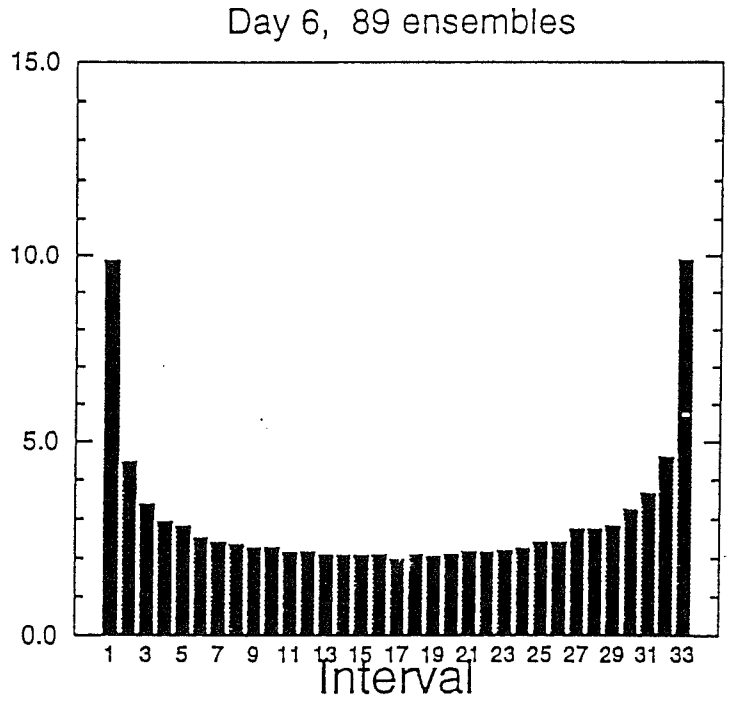

Fig 9: Talagrand diagram for t+144 hours 500 hPa height for winter 1994/95, evaluated over the northern hemisphere.