COMPARATIVE VERIFICATION OF PRECIPITATION FROM EPS AND FROM THE OPERATIONAL T213 FORECAST

by Ove Åkesson
SMHI
S-601 76 Norrköping
Sweden

1. Introduction

With the introduction of ensemble forecasts probability forecasts of temperature, precipitation and wind could be calculated from the spread of the ensemble members given that the distribution be realistic. However, until now the spread generally has not been large enough, particularly in the early stages of the 10-day forecast but becomes more reasonable towards the end . It has been shown from other verification studies that the ensemble mean of 850 and 500 hPa temperature and geopotential is in general more accurate than the corresponding value of the T213-forecast from around forecast days 4 or 5 (D+4/5) onwards (Emmrich et al, 1993) and Juga (1995).

The T213 precipitation forecasts are known to be of very good quality from regular comparative verification summaries issued monthly and seasonally in bulletin form in the Swedish Meteorological Service. The EPS based on 32 T63 low resolution forecast members was shown to be detrimental for predictions of extreme weather events of wind and precipitation (Åkesson 1993). In this study EPS precipitation is being evaluated and compared to the T213 forecast values both in deterministic as well as probabalistic terms using all available forecasts during a two-month period from 8 verification stations. Murphy (1993) has pointed out that a forecast cannot unanimously be stated to be superior to another forecast by simply comparing one or a few scores, since forecast goodness is multi-dimensional.

2. Probabalistic verification

In this evaluation study the 1 mm/24 h probability limit has been used for 7 Swedish and one Norwegian synoptic stations in September and October of 1995. Other higher limits are also available but the sample was considered too small to get significant and reliable results. The Brier score has been used for the probabalistic verification and as a reference the "climatological" mean value over all stations during this period has been used. The improvement of the Brier score result over the reference was calculated for each time step and referred to as skill.

From the T213 deterministic forecast probabalistic distributions based on one year of statistics of ECMWF precipitation forecasts were assigned to the categorical values and these probabilities were then treated in the same way as the EPS-based probabilities and skill values calculated. The distributions could be compiled for each forecast length from D+3 out to D+7 forecasts, but there was no statistics available for longer lead times.

The probability distributions used for the deterministic values might also be questionable whether they are optimal. Ideally these distributions should be derived individually for each station and for different seasons, since local and seasonal variations are considerable, but this was not possible and instead average values for all stations and for one full year were compiled.

3. Deterministic verification

It might be argued that the uncertainties in the medium range forecasts should be dealt with in probabalistic terms only, but it was also considered interesting to evaluate categorical values for several classes of precipitation amounts. In this study the categorical values of the T213 forecast and the mean ensemble values were divided into 5 classes: 0-0.3 mm, 0.4-1 mm, 1.1-3 mm , 3.1-10 mm and more than 10 mm /24 hours from which contingency tables were set up and verification measures and skill scores derived.

4. Verification results

Reliability diagrams for forecast D+3, 5 and 7 for EPS and for the corresponding T213 forecasts are presented in Fig. 1. There is a significant difference in sharpness between the probability distributions of the EPS and the T213 forecasts indicating the lack of sufficient spread of the ensembles. From statistics it is not possible to assign a value of 90, 80 or even 70% probability of precipitation exceeding 1 mm/24 hours to a high categorical value (<15 mm/24h) for a day 6 or day 7 T213-forecast, whereas the EPS is much more confident or sharper in this sense. The results in terms of skill is significantly higher for the T213 forecast for D+3 to 7. The ensemble skill is negative already at D+5, indicating lack of prognostic value, since a simple constant climatological value would normally be more accurate. This is to some

```
T213 D+5 FC CLASS  1    2    3    4    5
------------------------------------------ OBS
CLASS=LE     0.3/  44   10    6    1    3

CLASS= 0.3- 1.0/  30    7    4    7    2

CLASS= 1.0- 3.0/  21   14   11   10    2

CLASS= 3.0-10.0/  22    1   16   11    2

CLASS=GT   10.0/   2    0    0    2    0
      ------------------------------------

      ------------------------------------
PERCENT CORRECT:  32. OUT OF A TOTAL NO OF:  228,   SKILL:   0.09
BIAS:  0.1,  RMSE:  4.3,  MAE:  2.5,  OBSMEAN:  2.0
```

```
EPS D+5  FC CLASS  1    2    3    4    5
------------------------------------------ OBS
CLASS=LE     0.3/   8    2    0    0    0

CLASS= 0.3- 1.0/  27    5    0    4    1

CLASS= 1.0- 3.0/  57   14   19    9    5

CLASS= 3.0-10.0/  27   11   18   18    3

CLASS=GT   10.0/   0    0    0    0    0
      ------------------------------------

      ------------------------------------
PERCENT CORRECT:  22. OUT OF A TOTAL NO OF:  228,   SKILL:   0.06
BIAS:  0.6,  RMSE:  3.9,  MAE:  2.5,  OBSMEAN:  2.0
```

Table 1. Contingency tables at D+5 for T213 (left) and mean ensemble forecasts (right) for 8 synop stations during September and October 1995.

4. 1 Data correction

In Fig. 1 it is obvious that the ensemble bias of around 20% will penalize the results, whereas the bias is nearly negligible for the T213 forecast. Removing this bias will drastically improve the results for the ensemble probabilities as can be seen in Fig 1 (right). Removing the bias with a constant and then rotating the values around the observed frequency value leads to a higher resolution but also to a less sharp distribution in the tails of the diagram. This adjustment was carried out empirically due to the small data sample. However, the scores from this simple manipulation clearly exceed the scores obtained from the T213 forecasts for all time steps and with substantially higher sharpness than the T213 distributions. This manipulation or correction will ultimately lead to a more general and objective calibration procedure as soon as more data has been collected.

5. Conclusions

The results show that for precipitation it was preferable to use the deterministic T213 forecasts to the corresponding uncorrected EPS for forecast lengths between D+3 and D+7. The EPS forecasts had lost prognostic value already at D+3, whereas the T213 forecast remained skillful over climatology even out to D+7. However, after removing the bias and the over-confidence with a simple empirical correction procedure resulted in higher scores and higher sharpness for the EPS than for the T213 for all time steps between D+3 and D+7. The precipitation probabilities are therefore potentially very useful.

A comparison between the deterministic T213 forecasts with the corresponding ensemble mean forecasts revealed that the T213 forecasts were generally more accurate and realistic for various amounts of precipitation than EPS. This result might have been different and in favour of the EPS if the individual ensembles would have been used instead of the ensemble mean. This was not tested or further investigated, however. Contrary to temperature or geopotential, which are continuos variables, the use of ensemble mean of precipitation, which is non-continuos, should be avoided.

Despite the much lower resolution the EPS probability forecasts based on the T63-model were useful once the bias had been removed. An increase in model resolution of the ensemble members would most probably increase the usefulness even further and therefore it would seem preferable to enhance model resolution instead of increasing the number of ensemble members when higher computer speed is acquired in 1996.

1. REFERENCES

Åkesson O., 1993: Experimental use of ensemble forecasts at SMHI. *Proceedings from Fourth Workshop On Meteorological Systems*, 111-115.

Emmrich, P. and Balzer, K., 1993: Progress in Medium-Range Forecasting by means of Ensemble Forecasting. *Proceedings from Fourth Workshop On Meteorological Systems*, 173-177.

Juga, 1995. Use of ECMWF EPS products in the Finnish Meteorological Institute (FMI). *Expert meeting on Ensemble Prediction System*. 10-12.

Murphy A. H., 1993: What is a Good Forecast? An Essay on the Nature of Goodness in Weather Forecasting. *Wea. Forecasting*, 8, 281-293.

degree due to the over-confidence but also to the relatively large positive bias. The skill scores for the T213 forecasts, however, are all positive, albeit small at D+7.

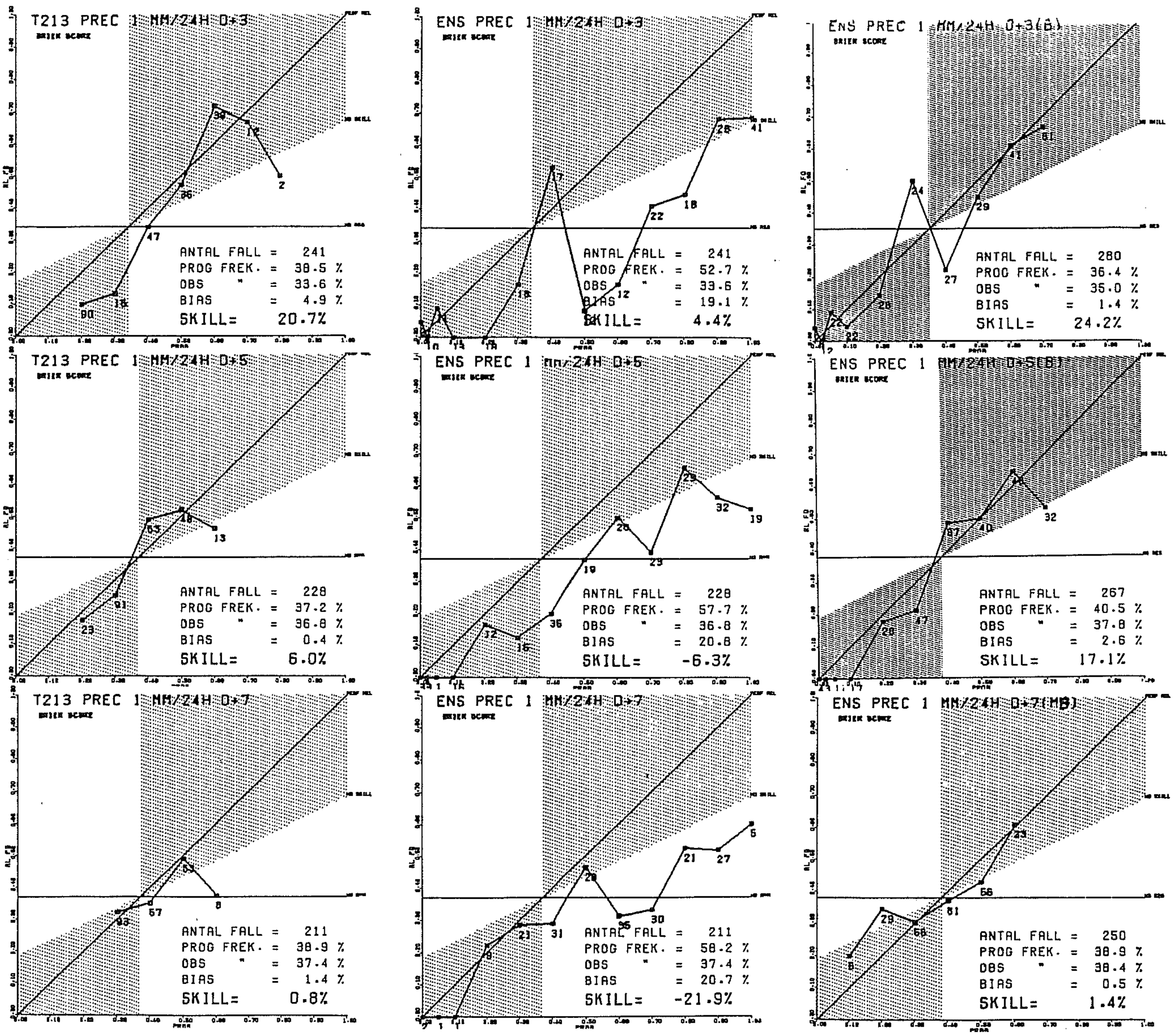

Fig. 1. Reliability diagrams for D+3, 5 and 7 for T213 24-hour accumulated precipitation forecasts (left) and the corresponding EPS forecasts (middle) for 8 Nordic synop stations during September and October 1995. In the right column bias-corrected and manipulated EPS forecasts are shown.

Verification results in terms of categorization of the deterministic values into 5 classes: 0-0.3 mm, 0.4-1 mm, 1.1-3 mm, 3.1-10 mm and more than 10 mm are shown in contingency tables for the T213 and the ensemble mean forecast in table 1 for D+5. Various diagnostic measures can be derived from these tables and here the percent correct and Heidke skill score, where the reference is based on chance, have been calculated. From Table 1 it is obvious that there is a large positive bias in the EPS for amounts smaller than 1 mm/24h, whereas the bias is negative for amounts exceeding 10 mm. The T213 forecasts are generally more realistic and nearly neutral for all categories although there is a certain overprediction of amounts smaller than 0.3 mm. The percent correct and the skill values are both in favour of the T213 forecast and this is the case also at D+3 and D+7.

Comparing the ensemble forecast mean with the T213 forecasts may however not be the correct thing to do; instead the individual ensemble forecasts listed in a similar table would probably have improved scores, but due to the large data volumes the ensemble data was unfortunately not saved.