Statistical post-processing on EPS

C. J. Kok

D. H. P. Vogelezang

KNMI (Royal Dutch Meteorological Institute), De Bilt

Verschenen in:
Joint Report on EPS Experts and Medium-Range Users meetings 1999/2000.
ECMWF

1. Introduction

Statistical post-processing of model output is used at KNMI for almost 20 years already for many predictands and forecast times. Since last year we have also included EPS information in the statistical guidance, both for deterministic as well as for probabilistic quantities. This is done not only for quantities that cannot directly be obtained from the available output but also to see whether information from direct model output quantities can be improved.

In this report we will describe a deterministic and a probabilistic experiment; the latter will be discussed more elaborately. In both experiments not only EPS output is used but also output from the operational model of ECMWF.

In this paper the results of EPS are compared with results obtained by statistically post-processing model output not only originating from EPS but also from the operational model. In order to be as fair as possible to EPS we have looked at a predictand which is directly available from EPS. We have chosen the 2meter temperature (T2m) and we will compare this with the statistically obtained T2m, deterministically and probabilistically. From a forecasters perspective it might have been more appropriate to have used the maximum or minimum temperature but they were not available from EPS at the start of these experiments.

In section 2 a summary of the design of the experiments is given including a description of the data sets used. In sections 3 and 4 the results of the deterministic and probabilistic experiments are described and discussed. Finally, in the last section the conclusions based on these two experiments are summarised.

2. Design of the two experiments

In the two experiments only "summer" seasons are considered, with summers defined here as the period from April to October inclusive. The predictand is the two-meter temperature at midnight and noon at station De Bilt. The forecast range is day 3 till day 10. In both the deterministic and the probabilistic experiment three different forecasting systems are compared. Only one is directly derived from EPS. In the case of the deterministic experiment the ensemble mean temperature is taken and compared with the operational 2m temperature and also with temperatures based on a statistical procedure.

In the probabilistic experiment we take first of all the percentage of ensemble members exceeding certain thresholds as probabilities. This will be referred to as the direct model output (DMO) probabilities of EPS. The other two forecasting systems are based on the statistically post-processing of model output, one entirely based on predictors derived from the operational model only and the other on predictors from the ensemble as well.

More specifically, 2 sets of predictors are used in the statistical post-processing. The first one is a set of local and large scale predictors from the operational ECMWF model. The local set consists of a large number of DMO quantities on several levels including quantities derived from them like thicknesses, advection terms, time derivatives. This set of potential predictors was only available until day 6. The large scale predictors are all derived from 500hPa fields on an area that comprises a large part of

Western Europe and the Atlantic Ocean and can be divided into two categories. In the first category the forecast 500hPa fields are projected on the first three empirical orthogonal functions (EOFs) and these EOF coefficients are used as predictors. The EOFs have been calculated on a data set of about 40 years. (Kruizinga, 1979). A second source of information derived from the large scale flow are so-called analogs. In this analog method we take the forecast large scale 500hPa flow and try to find analogous flow patterns in the past 50 years and use as predictors in the statistical scheme the observed weather at De Bilt (in this case only temperature information is used) belonging to those analogous situations. These "large scale" predictors are available until day 10.

The second set of predictors comes from EPS but only from the 2meter temperatures of the ensemble. It is available for the whole forecast range. The quantities used are the ensemble mean temperature, the standard deviation and the percentage of ensemble members which exceed certain thresholds. These thresholds are the anomalous temperatures of -8, -6, ..., +8 degrees Celsius.

The first statistical forecasting system is based on both the above-mentioned sets of potential predictors. This system will be referred to as the "full" guidance and will be denoted in the plots by MOS. The second system (no-EPS), only used in the probabilistic experiment, is based on predictors which are derived from the operational model only and is included in the comparison to assess what skill can already be obtained without the use of EPS information. The five predictands in this experiments are the probabilities of anomalous temperatures being below 2° and below 4° below normal, exceeding normal, 2° and 4° above normal, respectively. The two sets of (five) statistical forecast equations have been derived on the summers (as defined above) of '96 and '97 and applied to the summer of '98 and compared with the results of EPS.

3. Deterministic results

The verification results of the three forecasting systems on the summer of '98 will be presented in this section. The results for the operational 2meter temperatures (OPER), the ensemble mean 2meter temperatures (ENSM) and for the statistical Model Output Statistics temperatures (MOS) are expressed as a function of forecast time in terms of bias and standard deviation in Fig. 1 (top and bottom panel respectively).

The bias of the ensemble mean temperatures shows a large daily cycle: it alternates from somewhere around -0.4 at noon to circa +0.4 at midnight. Presumably this is due to a systematic underestimation of the more extreme temperatures. The amplitude of this flip-flop decreases with increasing lead time. The bias of the operational model is small and does not show a daily cycle at all. The bias of the MOS equations, on the other hand, shows a similar behaviour as the ensemble mean but with opposite sign. Its amplitude, however, is much smaller. This daily cycle is probably due to an overcorrection due to the somewhat larger errors in the dependent data. Note that the ensemble mean was one of the leading predictors in the MOS equations.

The standard deviation of the operational model is much larger than the standard deviation of the ensemble mean temperatures. This is already true at day 3. The MOS temperatures give the smallest values although the results differ only a little from those for the ensemble mean. But given also the smaller biases one can say that over the entire forecast range the MOS equations are better than the deterministic forecasts of the

operational model and the ensemble mean. This is in agreement with results obtained for other predictands and seasons (not shown) over the last couple of years.

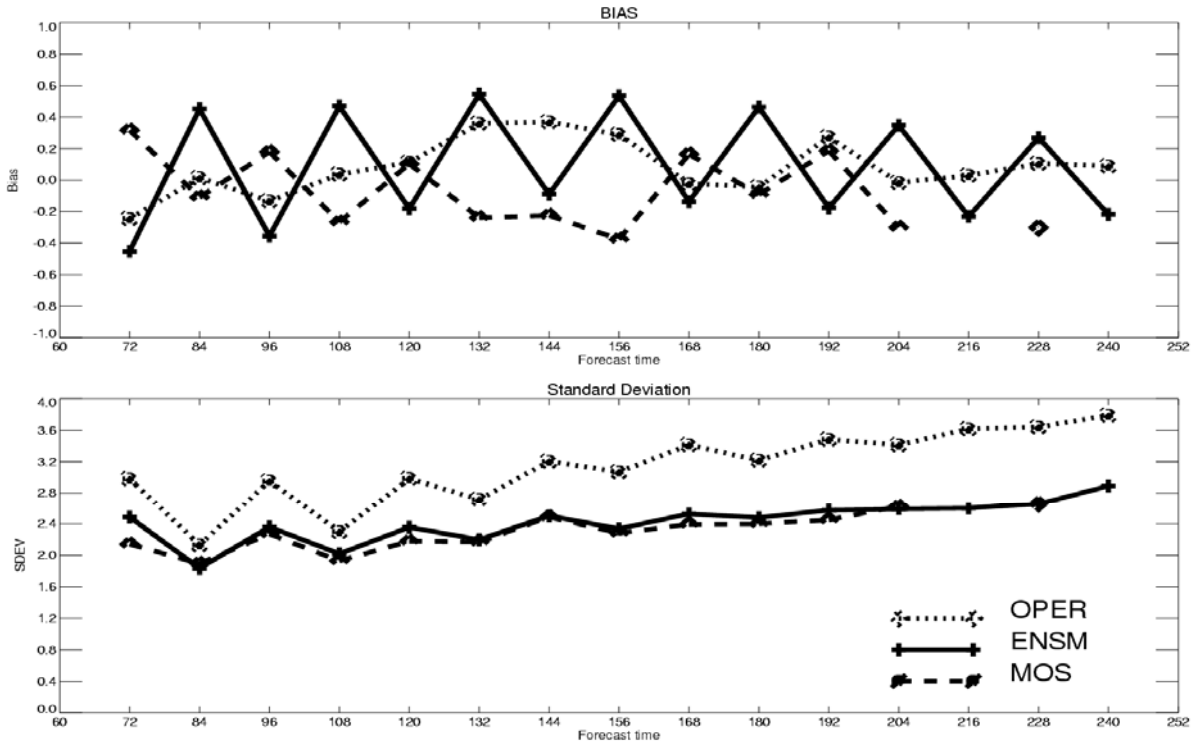


Fig. 1 Verification results for the operational (OPER), the ensemble mean (ENSM) and the Model Output Statistics (MOS) 2meter temperatures in terms of mean error (top) and standard deviation (bottom).

4. Probabilistic results

As was described in chapter 2 we have calculated probabilities of temperatures exceeding 5 thresholds. These thresholds are -4, -2, 0, +2 and +4 degrees with respect to the climatological mean. Only temperatures at noon are considered. We have compared three different forecasting techniques on the summer of last year. One is the DMO probabilities of EPS, the other two are statistical in nature: one including also EPS predictors (referred to as MOS) and one using operational model predictors only (no-EPS). A description of the different sets of potential predictors is given in chapter 2. The rationale behind taking only 5 thresholds (or 6 classes), which only crudely represents the probability distribution, is that experiments showed that on a limited data set of only 6 months it was already extremely difficult to derive stable statistical relations and also to proove significant skill of either of the three forecasting systems. Deviations of more than 6 degrees from normal for instance simply happen on too few occasions.

We will compare the three probabilistic forecasting systems for the categories defined by the above-mentioned five thresholds. On the one hand we take the probabilities directly from EPS simply by counting the number of ensemble members that exceed these thresholds. This will be called the DMO of EPS. On the other hand we have a statistical MOS system which calculates the probabilities on the basis of the same set of predictors we have seen before in the deterministic experiment. Additionally we also derived MOS equations on the basis of predictors of the operational model only in order to see what skill can be obtained without using EPS information. We have to stress that in this case only limited information is available beyond day 6.

But before comparing these three sets of probability forecasts let's look at how these forecasts may be presented and especially how the general characteristics of the EPS and MOS probabilities may differ from one another. In Fig. 2 a consistency plot is given for consecutive forecasts verifying on the 17th of June 1999. The forecasts are from 4 to 10 days in advance for DMO EPS probabilities. The bars consist of probabilities for 6 classes (derived from the probabilities exceeding the 5 thresholds) plotted on top of each other and ranging from the probability of temperatures below 4 degrees below the normal temperature value at the bottom to the probability of over 4 degrees above normal at the top. Categories below the climatological mean are plotted in blue below the zero line, and likewise, the ones above climatology in red above the zero line. The climatological mean 12UTC temperature is about 18.2 degrees around this time of year.

The consistency plot for the MOS probabilities (not shown) shows much less abrupt changes from day to day. The higher consistency of the MOS predictions is not surprising off course, since due to the nature of MOS extreme statements are penalised severely if they are wrong. Note that the MOS equations are not purely based on EPS but have increasingly more information from the operational model with decreasing forecast time.

The verification on the summer of last year has been performed for the five thresholds separately (Fig. 3) and is presented in terms of the ROC area and the Brier Score (BS). The ROC area is calculated as the surface under the curve of the Hit Rate as a function of the False Alarm Rate. These two scores have been calculated by constructing contingency tables by taking a number of probability thresholds above which the event is said to have occurred. These thresholds are taken every 1 percent. In the discussion of the experiments more emphasis is put on the results in terms of the Brier Score (the mean square difference between the forecast and observed probabilities).

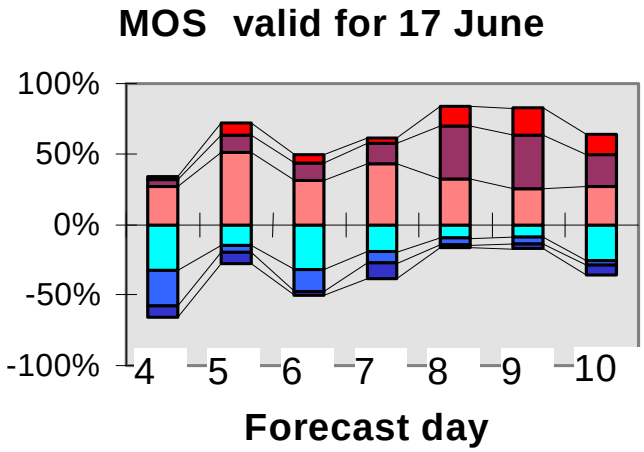


Fig. 2 Consistency plots for DMO EPS probabilities (top) and guidance MOS probabilities for 4 to 10 days in advance and valid for the 17th of June 1999.

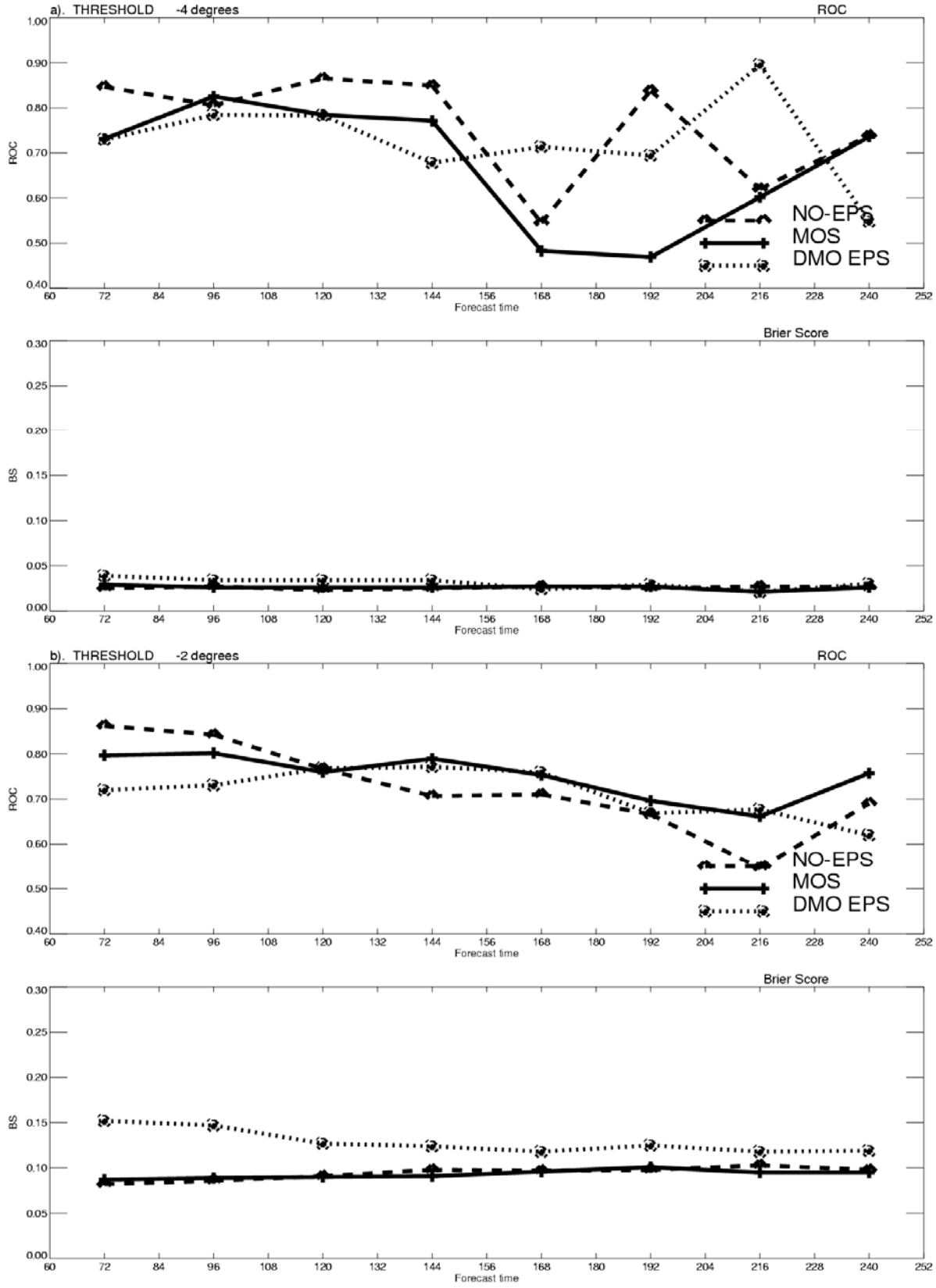
In Fig. 3a the verification results are shown for the probabilities of an anomalous temperature of less than 4 degrees below the climatological mean. The observed frequency was less than 3 percent (only 5 cases). The BS is off course very small for all three forecasting systems with a minor advantage for the two statistical schemes. It hardly changes with forecast time. The ROC area shows a very noisy behaviour. This is probably due to the small number of cases for this category.

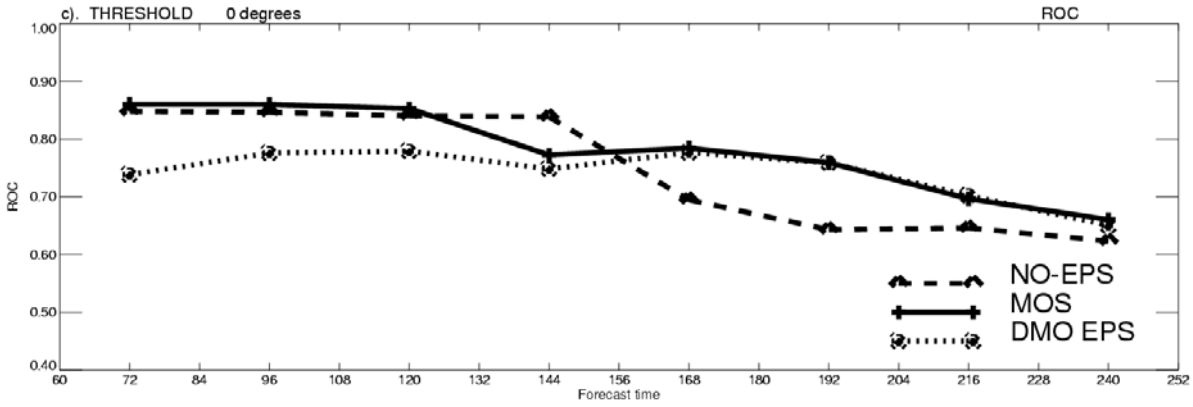
For the probability of temperatures below 2 degrees below normal (Fig. 3b) with an observed frequency of 11.5% the results are much more significant. The Brier Score of the full statistical scheme is much better than that of the DMO EPS over the entire forecast range. Also the no-EPS MOS equations seem to have more skill than the EPS probabilities. This is not the case, however, when the ROC area is considered. At day six the no-EPS guidance becomes worse than the ensemble probabilities. We also see that EPS improves with increasing forecast time indicating that the spread of the ensemble is not large enough until day 5 or 6. The small spread results in strong statements (very high or very low probabilities) for the event to occur and these will be severely penalised in terms of the BS if the event does not occur.

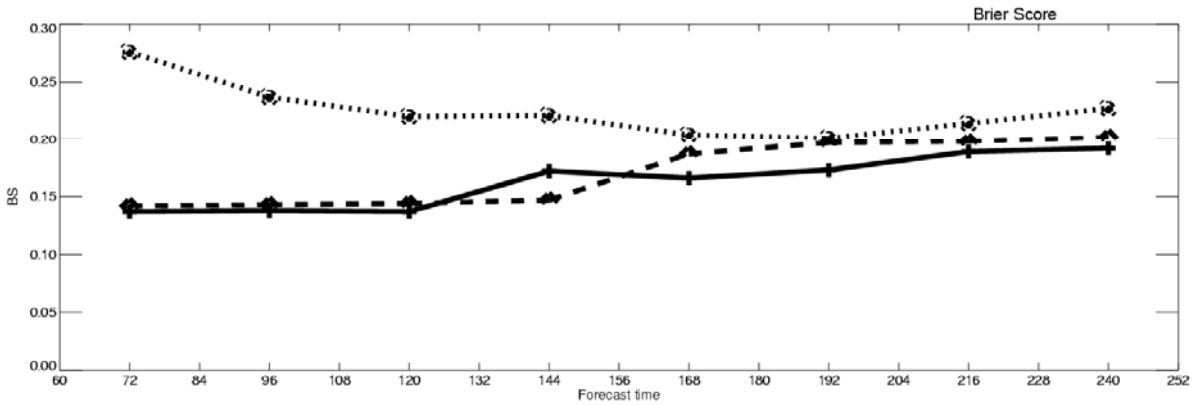
More or less the same results apply for above normal temperatures (Fig. 3c), comprising over 70 percent of the cases. Once again both statistical schemes perform better than EPS in terms of the BS. But now at day seven the no-EPS equations deteriorate and become worse than EPS only in terms of the ROC area.

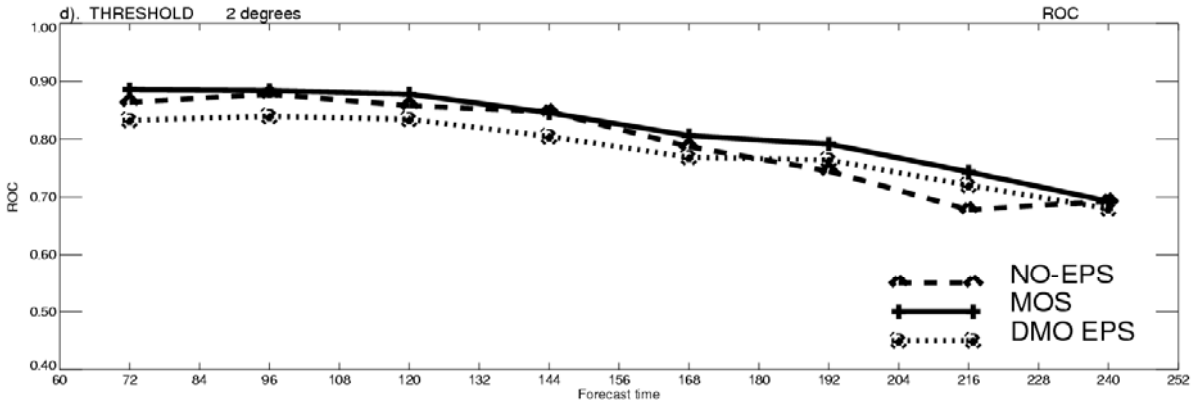
The higher skill of the MOS schemes is also demonstrated for the last two categories (Fig. 3d,e), in which probabilities of temperatures over 2 and 4 degrees above normal, respectively, are considered. The results for the highest temperature category may still be significant, in contrast to those for the coldest temperatures, since temperatures warmer than 4 degrees above normal occurred in almost 15 percent of the cases.

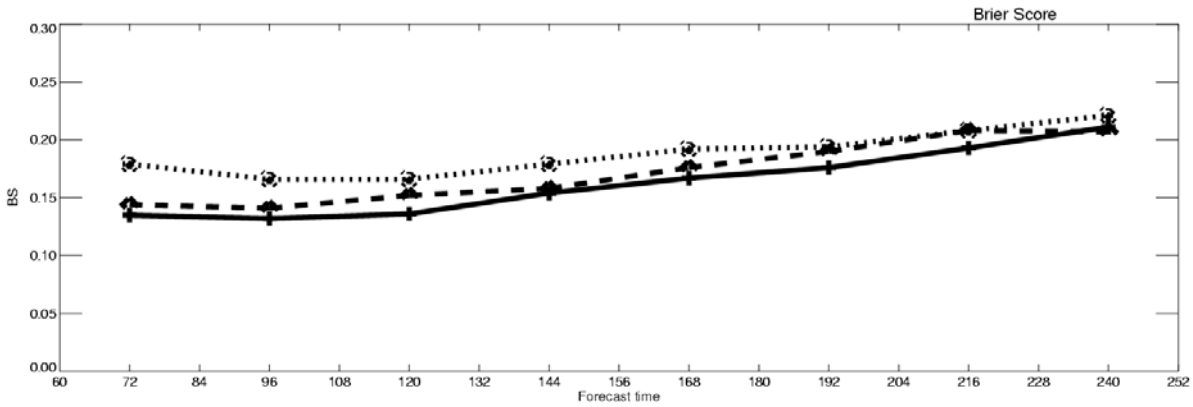
A final feature to be mentioned is the different behaviour of the two statistical schemes. In general the full scheme performs better than the no-EPS, especially in terms of the BS, indicating that EPS supplies important predictors. This seems to be obvious for lead times higher than day five, but it even seems to be true, though much less prominent, at days three and four.


a). THRESHOLD -4 degrees
ROC
ROC
Forecast time
NO-EPS
MOS
DMO EPS
Brier Score
BS
Forecast time
b). THRESHOLD -2 degrees
ROC
ROC
Forecast time
NO-EPS
MOS
DMO EPS
Brier Score
BS
Forecast time


c). THRESHOLD 0 degrees
ROC
ROC
Forecast time
NO-EPS
MOS
DMO EPS


Brier Score
BS
Forecast time


d). THRESHOLD 2 degrees
ROC
ROC
Forecast time
NO-EPS
MOS
DMO EPS


Brier Score
BS
Forecast time

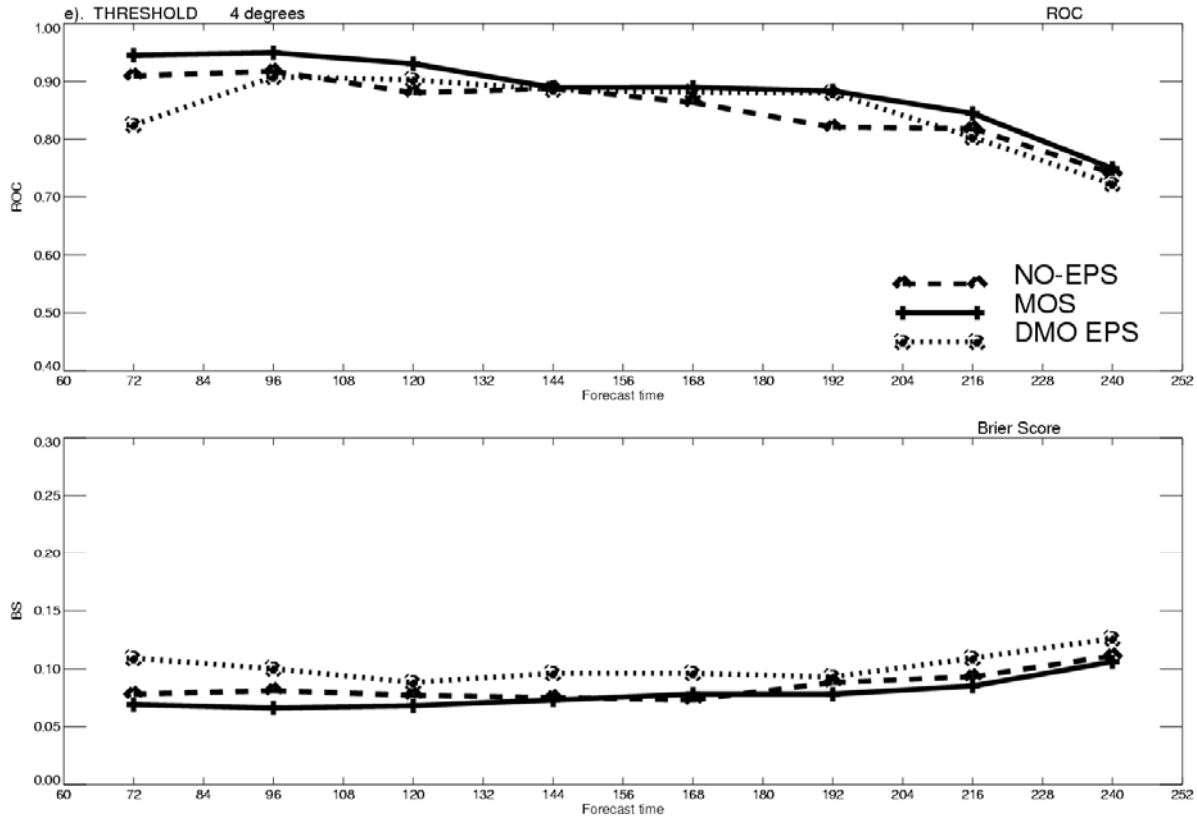


Fig.3. Verification results in terms of ROC area and Brier Score for probabilities of anomalous temperatures exceeding 5 thresholds: -4° below normal, -2° below normal, zero degrees, +2° and +4° above normal (Figs. 3a to 3e respectively). Three forecasting systems are shown: probabilities directly derived from EPS (DMO EPS, dotted lines), and two statistical schemes, one using predictors from EPS as well as from the operational model (MOS, thick lines) and one using predictors from the operational model only (no-EPS, dashed lines).

5. Summary

The skill of EPS 2meter temperatures, both in deterministic as well as in probabilistic sense, has been assessed for a single station in The Netherlands on the summer of 1998. The summer was defined here as the period ranging from April till October. The forecast range of day 3 till day 10 was considered. The results have been compared with the output of statistical temperature equations which were derived on an independent data set consisting of the two previous summers. The predictors that were included in the statistical scheme originated from the operational model (both small scale and large scale predictors were used) as well as from EPS.

The ensemble mean temperatures proved very hard to beat. Even at day 3 they were already considerably better than the operational forecasts. The statistical MOS equations, however, were slightly better still for all forecasts times. This is in agreement with experience over the last number of years not only for temperatures but also for other deterministic predictands.

Counting the number of ensemble members exceeding certain thresholds can be interpreted as the probability of its occurrence. In this paper we have called this DMO EPS probabilities. These probabilities have been compared with two statistical schemes in terms of Brier Score and ROC area. The "full" statistical method, in which predictors of both the operational model as well as EPS are used, provided by far the best probabilistic forecasts of summer anomalous midday temperatures. However, for one of the most extreme temperature classes with very few cases the difference was much smaller. Even the second statistical method with predictors from the operational model only performed better than the DMO EPS probabilities for lead times to at least around day six. For the highest forecast times this was not the case. Another interesting feature was the improvement of the statistical performance when EPS predictors were included. This is very conspicuous for longer lead times. Apparently there seems to be important information in the temperature plumes which is not available in the predictors in the way we have derived them from the operational run. Remember that these predictors are either local (grid point De Bilt) or derived from large scale 500hPa fields. Presumably predictors at the intermediate scale are provided by EPS.

But also the shorter lead times seemed to be benefitting from EPS predictors. Apparently even at day three there is information in EPS that can be used statistically to improve the guidance despite the fact that the probabilities of EPS itself are highly insufficient for the short forecast ranges.

References

Kruizinga, S. 1979. Objective classification of daily 500 mbar patterns. *Sixth Conference on Probability and Statistics in Atmospheric Sciences*. Banff, Alberta, Canada.