Comments on "Probabilistic Predictions of Precipitation Using the ECMWF Ensemble Prediction System"

JOSIP JURAS

Geophysical Institute, Faculty of Science, University of Zagreb, Zagreb, Croatia

20 August 1999 and 14 December 1999

Buizza et al. (1999) have used a large number of various verification methods in their recent work on the Ensemble Prediction System (EPS) of the European Centre for Medium-Range Weather Forecasts. Most of the measures, like Brier score (BS), Brier skill score, and threat score have been in use for a long time, unlike the relative operating characteristic (ROC) and normalized distance between conditional distribution, which have found a wider application only in the last decade. Most of these measures have been developed for the verification of forecasts at a single site. We do not yet have enough experience with their application for the evaluation of EPS products. For example, the area under a ROC curve of 0.50 indicates a useless forecast system when this measure is used for verification of forecasts for a single site. Buizza et al. (1999) suggest an area of 0.70 as a limit of useful EPS product. The questions arise first as to why there is a difference in the limits that distinguish the useful from the useless forecasts and second if the limit of 0.7 can be accepted as a universal limit for the verification of the EPS products. It is my opinion that this limit varies and depends on the variability of climatological frequencies within the area. In the considered large area, which includes all of Europe, small parts of North Africa, and east Asia, the frequencies of precipitation are very different in various places. They range from only 3 days per year with the precipitation amount ≥ 1 mm in Cairo (Egypt) to 202 days in Bergen (Norway). If we aggregate probabilistic forecasts for places with different frequencies of an event in one set, we will obtain an artificially skilful set of forecasts, although forecasts are always identical and equal to the climatological frequencies of the event at the corresponding place.

This fact is illustrated by one example. Table 1 shows data for one hypothetical set of climatological forecasts for grid points that have different climatalogical frequencies of an event. They range from 0.05 to 0.45. The average frequency for the area is 0.20. The relative frequency of the points for each of the five categories are indicated in the second column. The forecasts show perfect reliability, which is a common assumption for climatological forecasts. From this data, one could compute the probability of detection (POD), the false alarm rate (FAR), and segments under the ROC for various probability thresholds. The total ROC-area for this example is 0.714. The resolution term in Brier score is 0.016 and Brier skill score is 0.10. This set of climatological forecasts seems skillful, although it is not. The apparent value of forecasts is due to the fact the set consists of forecasts for points with different climatalogical frequencies. For regions that are homogeneous with respect to the frequency of an event, limits between the successful and the unsuccessful forecasts according to the area under the ROC would be very close to 0.5 as it is the case for single-point forecasts during a climatic homogeneous season. On the basis of these considerations, the conclusion proposed by Buizza et al. (1999), "that the predictive skill of the system is lower for small regions" (Irish and Alpine regions) seems to be inadequate. It is assumed that for these relatively homogeneous regions, with respect to the frequencies of precipitation, a value of 0.70 for the area under the ROC is too high as a limit for a useful prediction. I strongly support the intention by the authors to provide

TABLE 1. Fictitious example of climatological precipitation forecasts. Here, n_k indicates relative frequencies of points with climatological frequency o_k of an event. The average climatological frequency ($\langle o \rangle$) and resolution term in Brier score (BS_{res}) are the sum of the values in the column marked $n_k o_k$ and $n_k(o_k - \langle o \rangle)^2$, respectively.

o_k	n_k	$n_k o_k$	$n_k(o_k - \langle o \rangle)^2$	POD	FAR	ΔROC
0.05	0.25	0.013	0.006	0.938	0.703	0.288
0.15	0.30	0.045	0.0008	0.713	0.384	0.263
0.25	0.25	0.063	0.0006	0.400	0.150	0.130
0.35	0.10	0.035	0.002	0.225	0.069	0.025
0.45	0.10	0.045	0.006	—	—	0.008
	$\langle o \rangle = 0.200$		$BS_{res} = 0.016$		ROC = 0.714	

Corresponding author address: Dr. Josip Juras, Geophysical Institute, Faculty of Science, University of Zagreb, Horvatovac bb, 10000 Zagreb, Croatia.
E-mail: juras@olimp.irb.hr

for the ''development of verification system that can verify EPS forecasts against detailed rain gauge data.'' The appropriate single measure for the verification of EPS products is the median (not the mean) of the distribution of Brier skill scores for individual observation stations (or grid points).

Figures 2–4 in Buizza et al. (1999) provide us with an excellent possibility to compare the subjective judgement of the accuracy of nine EPS forecasts with the corresponding values of ROC area, the values of which vary in the wide range from 0.50 to 0.93. However, based on Figs. 1c,d and Fig. 18 one could get a wrong impression that forecasts show a low accuracy. The horizontal line of ''no resolution'' and the corresponding orthogonal line are misplaced. The lines must pass through the value of the average frequency of the event on both the abscissa and ordinate. According to their Table 1 it seems that these values should be in the region of 0.20 for the threshold of 1 mm $(12\ h)^{-1}$ and 0.02 for 10 mm $(12\ h)^{-1}$. Correctly drawn, the attributes diagrams would make the forecasts look better.

REFERENCES

Buizza, R., A. Hollingsworth, F. Lalaurette, and A. Ghelli, 1999: Probabilistic predictions of precipitation using the ECMWF Ensemble Prediction System. *Wea. Forecasting,* **14,** 168–189.