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

LAURENCE J. WILSON

*Recherche en Prévision Numérique, Dorval, Quebec, Canada*

19 July 1999 and 22 December 1999

## 1. Introduction

In their paper, Buizza et al. (1999, hereinafter referred to as BHLG) present and discuss various attributes of the quantitative precipitation forecast (QPF) performance of the European Centre for Medium-Range Weather Forecasts ensemble system. The verification tools used are signal detection theory measures, reliability diagrams, the Brier score and skill score, the threat score, and the root-mean-square error. The verification data consisted of short-range (0–24 h) forecasts from the full-resolution model. In an appendix, BHLG justify the use of model data as ''observations'' on the basis that point precipitation observations are representative of smaller scales than model QPF output. Comparative summary results are presented for four 3-month seasons, for a model grid domain covering Europe. Variability in the performance is assessed for two smaller domains, as a function of time, and as a function of projection time. The variability in performance is also assessed as a function of the threshold chosen for probability estimation from the ensemble output. An interesting feature of the paper is the presentation of several case studies, which facilitate the synoptic interpretation of the relative operating characteristic (ROC) values.

My comments on this paper relate mostly to the ROC curve and its use and interpretation. The ROC is relatively new to meteorology, having been brought into the field as a verification tool by Mason (1982). Its use has become more widespread since the advent of ensemble forecasting. In Murphy and Winkler's (1987) framework for probability forecast verification, the ROC fits into the ''likelihood-base rate'' factorization of the joint distribution of forecasts and observations, which implies a stratification of the joint distribution according to the observation. Specifically, the ROC curve and two associated summary measures, the area under the curve (AZ) and the separation of the two conditional distributions (DZ), seek to assess the ability of a diagnostic system to separate cases where a signal is present from cases where only noise is present. In the application to precipitation forecasting described in BHLG, the ''signal'' is the occurrence of 12-h precipitation amounts over a threshold, and the diagnostic system is the ensemble system, which forecasts a probability distribution of precipitation amounts at each forecast time. Signal detection theory is fully described in Swets and Pickett (1982), and its application to meteorology was explored by Mason (1982).

I have two specific comments, discussed in the following sections. The first relates to the method of computation of the ROC, and the second relates to the interpretation of the ROC area in terms of the skill of the forecast.

## 2. Computation of the ROC

Figure 1 of BHLG indicates the plotted points of the empirical ROC have been joined by straight lines. This practice is inconsistent with previous published results of Mason (1982) and Swets (1986), both of which indicate that, for a wide variety of applications, the ROC is linear in terms of the standard normal deviates corresponding to the hit rate and false alarm rate, and are curved in linear probability space. Joining the points of the ROC by straight lines thus will lead to an underestimate of AZ, the magnitude of which will depend on the distribution of points across the range of the ROC. In BHLG, where ROCs are estimated for four thresholds of precipitation occurrence from plots of hit rate versus false alarm rate, the tendency is for the points to cluster toward the lower-left corner as the threshold increases, resulting in greater underestimation of AZ for higher thresholds than for lower thresholds. Assuming that all the ROC areas reported in BHLG were estimated as demonstrated in their Fig. 1, the differences in ROC areas for the different thresholds (e.g., Figs. 1a, 5, 6a, 8, and 9a,b) would be expected to be mainly due to

*Corresponding author address:* Dr. Laurence J. Wilson, Recherche en Prévision Numérique, Meteorological Service of Canada, 2121 Route Transcanadienne, Suite 500, Dorval, PQ H9P 1J3, Canada.
E-mail: lawrence.wilson@ec.gc.ca

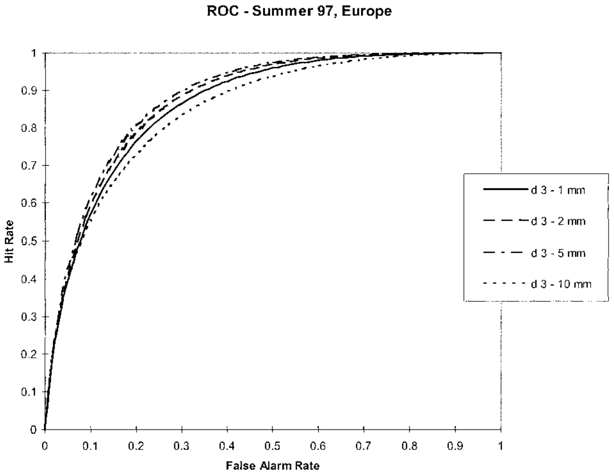

FIG. 1. Relative operating characteristic curves for day 3 European area 12-h precipitation forecasts, summer 1997, for thresholds 1, 2, 5, and 10 mm.

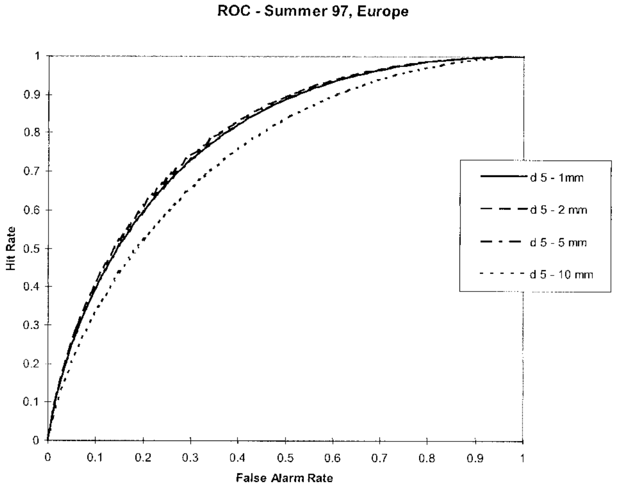

FIG. 2. Relative operating characteristic curves for day 5 European area 12-h precipitation forecasts, summer 1997, for thresholds 1, 2, 5, and 10 mm.

variations in the degree of underestimation of the true ROC rather than to real differences in the ability of the ensembles to discriminate between occurrences and nonoccurences of precipitation amounts above the threshold. BHLG have used Stanski et al. (1989) to guide their computation of the ROC. As one of the authors of Stanski et al. (1989), I can admit that the ROC example (Fig. 4.2) in that publication does show points connected by straight lines, which could be misleading. In that case, however, the points are well distributed across the range of the ROC, which would minimize the magnitude of the error.

To help me investigate this issue, Dr. Buizza has graciously supplied me the data that was used to construct his Fig. 1, along with other data used to obtain results reported in BHLG. I recalculated the ROCs for the four thresholds 1, 2, 5, and 10 mm, for both 3 and 5 days, using a program obtained from I. Mason (1987, personal communication) that is based on the procedures outlined in Swets and Pickett (1982). This program calculates the standard normal deviates corresponding to the empirical hit rates and false alarm rates, then fits a straight line by the least squares method. Tests of goodness of fit are also computed, such as the Student's *t*-value for significance tests on the slope and the correlation between the two variables. For ease of plotting in linear probability space, the fitted ROC is transformed back to probability values at 100 equally spaced points. It should be noted that the only assumption that is made by fitting the ROC in this way is that the underlying distributions before occurrence and nonoccurrence of the events can be transformed to normal by a monotonic transformation; the underlying distributions do not themselves need to be Gaussian (Swets 1986).

The fitted ROCs are shown in Fig. 1 for day 3 forecasts and in Fig. 2 for day 5 forecasts, for the European area, for summer, 1997. For 1 and 2 mm, the fit was extremely good; the correlation was 1.0 to three figures. For 5 and 10 mm, the correlation dropped slightly, but was never lower than 0.98. This confirms the previous experience of Swets (1986) and Harvey et al. (1992) that empirical ROCs are very close to linear in normal deviate space. The level of significance, however, dropped more quickly. At 10 mm, the fit was significant only to the 99% confidence level, while at 20 mm, a fit could not be obtained. This is a reflection of the tendency of the points to cluster near the lower-left corner of the ROC for the higher thresholds, forming a weaker basis by which to determine the full ROC. Beyond 10 mm, there were not enough occurrences of the event to determine the distribution before occurrences with sufficient reliability.

I also recomputed ROC curves for the data from summer 1996, again for day 3 and day 5. Precipitation occurrences were less common in 1996 than in 1997 at all thresholds (as forecast by the full-resolution model); thus, the 1996 data presented greater problems to obtain an acceptable ROC. For the 10-mm threshold, the *t*-test results indicated significance only to the 80% level. Once again, the difficulty is that lower ''observed'' frequencies of the event prevent a reliable description of the conditional distribution of probability forecasts preceding its occurrence, regardless of how frequently the model attempts to predict it.

All the results obtained are shown in Table 1, in comparison with the results obtained by BHLG. The figures in the table support the following comments.

1) Values of the area AZ are generally higher than those obtained by BHLG. This is a consequence of fitting a smooth curve instead of approximating the curve with a set of straight lines. Compared to experience with ROC areas for precipitation these values are generally in the typical range (Swets 1988), if one

TABLE 1. Comparison of fitted and BHLG ROC areas and DZ for summer 1996 and 1997 data, for Europe, days 3 and 5. Differences are computed as (fitted − BHLG) in all cases.

| | Day 3 | | | | Day 5 | | | |
|---|---|---|---|---|---|---|---|---|
| Threshold (mm) | 1 | 2 | 5 | 10 | 1 | 2 | 5 | 10 |
| Fitted ROC area, summer 1997 | 0.866 | 0.876 | 0.884 | 0.851 | 0.787 | 0.793 | 0.787 | 0.745 |
| BHLG area, summer 1997 | 0.832 | 0.825 | 0.778 | 0.697 | 0.764 | 0.755 | 0.690 | 0.585 |
| Difference | 0.034 | 0.051 | 0.106 | 0.154 | 0.023 | 0.038 | 0.097 | 0.160 |
| Fitted ROC area, summer 1996 | 0.870 | 0.869 | 0.860 | 0.863 | 0.795 | 0.797 | 0.736 | 0.620 |
| BHLG area, summer 1996 | 0.820 | 0.794 | 0.694 | 0.593 | 0.759 | 0.732 | 0.633 | 0.520 |
| Difference | 0.050 | 0.075 | 0.166 | 0.270 | 0.036 | 0.065 | 0.103 | 0.100 |
| DZ, summer 1997 | 1.567 | 1.633 | 1.688 | 1.471 | 1.127 | 1.158 | 1.126 | 0.933 |
| BHLG DZ, summer 1997 | 1.220 | 1.210 | 0.887 | 0.454 | 0.677 | 0.653 | 0.381 | 0.124 |
| Difference | 0.347 | 0.423 | 0.801 | 1.017 | 0.450 | 0.505 | 0.745 | 0.809 |
| DZ, summer 1996 | 1.589 | 1.587 | 1.526 | 1.544 | 1.164 | 1.176 | 0.894 | 0.434 |
| BHLG DZ, summer 1996 | 1.190 | 1.060 | 0.551 | 0.179 | 0.711 | 0.628 | 0.245 | 0.026 |
| Difference | 0.399 | 0.527 | 0.975 | 1.365 | 0.453 | 0.548 | 0.649 | 0.408 |
| Fitted ROC area difference, 1997–1996 | −0.004 | 0.007 | 0.024 | −0.012 | −0.008 | −0.004 | 0.051 | 0.125 |
| BHLG area difference 1997–1996 | 0.012 | 0.031 | 0.084 | 0.104 | 0.005 | 0.023 | 0.057 | 0.065 |
| DZ difference 1997–1996 | −0.023 | 0.047 | 0.163 | −0.074 | −0.037 | −0.018 | 0.232 | 0.499 |
| BHLG DZ difference 1997–1996 | 0.030 | 0.150 | 0.336 | 0.275 | −0.034 | 0.025 | 0.136 | 0.098 |

considers the 10-mm category to represent a ‘‘storm.’’ However, the Swets (1988) values are all for short-range forecasts compared with observations, while the present results are for medium-range forecasts compared with short-range model output. One would expect AZ to be lower in the medium range, but higher when the forecasts are compared with output from a higher-resolution version of the same model. Any false alarms and missed events exhibited by both the full-resolution model and the ensemble model will not be accounted for in the present verification results.

2) The difference between the fitted area and the BHLG area increases with threshold. As noted above, this too is expected because the plotted points tend to cluster toward the lower left-hand corner of the curve for higher thresholds.
3) The separation distance (DZ) statistics are consistent with the AZ statistics. This is expected from the theory of the ROC: The greater the DZ value, the smaller the overlap of the two distributions, and the greater the signal detection ability of the system.
4) Area AZ changes little over the range of thresholds from 1 to 10 mm. In other words, the ROC appears to be nearly invariant under a change of threshold of the physical variable, for a given dataset. Changing the threshold amounts to sampling the ensemble probability distribution at different values of the random variable, and the set of probabilities generated will vary monotonically with the threshold for each forecast. The calculation of the ROC involves sampling the probability distribution for a given precipitation threshold over the whole sample at different probability thresholds (usually deciles), again monotonically. These procedures are essentially equivalent and it is therefore not surprising that the ROC varies little for different precipitation thresholds. Nevertheless, there is some indication from the results that the system does a poorer job of discriminating at the higher thresholds for the 1996 data especially. While at first this might seem inconsistent with the previous argument, the monotonic transformation of probabilities applies to the distribution over the whole sample. It is still possible that the transformation leads to a poorer separation of means after partitioning into the two conditional distributions.
5) There is much less evidence in these results to support a substantial impact of the system upgrade between 1996 and 1997, as claimed by BHLG. Of the eight ROC areas computed for summer 1996 and summer 1997 (four for each of day 3 and 5), only four of these show a positive trend between 1996 and 1997, and all changes are small with the exception of 5 and 10 mm for day 5. Bearing in mind that the 1996 day 5 result for 10 mm was not statistically significant, there is some evidence that the newer system more clearly detects the signal for occurrences of 5 or more and 10 or more mm of rain in 12 h. This might indicate that the higher-resolution ensemble system retains the sharpness of the forecast precipitation distribution further into the forecast period than did the older system, compared to the full-resolution model. The impact is clearly not substantial, however.

The implications of these results for the discussion and conclusions presented in BHLG are substantial. First as suggested above, in Figs. 1a,b, 5, 6a,b, 8, and 9, revised curves should be higher and lie closer together. The magnitude of underestimation of AZ will be sensitive to the location of the last point on the empirical curve, which may vary greatly from case to case. As shown above the difference can be as much as 0.25 or so, enough to obscure differences due to variations in the signal detection ability of the forecast system.

Although the analysis shown above relates only to the summer data, I would expect similar effects to apply to the winter data, specifically in Figs. 15 and 16a,b.

Second, the comparison between 1996 and 1997 performance would also be affected (Figs. 10a and 10b in BHLG). Since precipitation occurrence was a rarer event in 1996 than in 1997 for the summer data, one would expect greater underestimation of the true ROC for the 1996 data, which would possibly lead to a false indication of improvement between the two years. Indeed, this is consistent with the results presented above, which show little change between 1996 and 1997, except for the 10-mm threshold. While a complete reanalysis would be needed to confirm this, the results suggest that the two curves in Fig. 10a would lie on top of each other, and both be higher, and the two curves in Fig. 10b would be much closer together, meaning that the gain in signal detection is less than the 3 days indicated in BHLG.

While it is often considered better to use nonparametric methods in verification, because no assumptions are involved, all the previous experience reported in the literature supports the validity of the normal–normal model in signal detection theory as providing a more accurate and stable estimate of the ROC than a linear nonparametric method. The results presented here conform well to the previous work in this field and lead to somewhat different conclusions from those contained in BHLG.

## 3. Interpretation of the area under the ROC in terms of forecast skill.

Stanski et al. (1989) suggest that a minimum threshold of 0.7 serves as a reasonable lower limit for meaningful skill of the forecast in the signal detection theory context. This value is mentioned several times in BHLG. This guide was intended to apply to realistic situations where model output is compared to observations. Although the choice of a specific threshold value aids in comparative assessment of the results, it is easy to forget that the results in this case are generated by comparing the ensemble forecasts with forecasts from a higher-resolution version of the same model. Inevitably, this would lead to artificially high values of all the verification scores, and it is unknown how the score values would ‘‘map’’ to the real world if observation data or at least output from a completely independent model were used for verification. Perhaps a higher threshold should be used to allow for the impact of using model output as verification data. The statement at the end of section 3 of BHLG (‘‘Concluding, these results . . .’’) may turn out to be true but for the wrong reasons. Furthermore, in this context, claims such as ‘‘gain in predictability up to three days,’’ as stated in the abstract, must surely be artificial. If it is atmospheric predictability that is referred to, how can this be judged when no atmospheric data are used in the verification?

## 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.

Harvey, L. O., K. R. Hammond, C. M. Lusk, and E. F. Mross, 1992: The application of signal detection theory to weather forecasting behavior. *Mon. Wea. Rev.,* **120,** 863–883.

Mason, I., 1982: A model for assessment of weather forecasts. *Aust. Meteor. Mag.,* **30,** 291–303.

Murphy, A., and R. L. Winkler, 1987: A general framework for forecast verification. *Mon. Wea. Rev.,* **115,** 1330–1338.

Stanski, H., L. J. Wilson, and W. R. Burrows, 1989: Survey of common verification methods in meteorology. WMO World Weather Watch Tech. Rep. 8, 114 pp.

Swets, J. A., 1986: Form of empirical ROCs in discrimination and diagnostic tasks: Implications for theory and measurement of performance. *Psychol. Bull.,* **99,** 181–198.

——, 1988: Measuring the accuracy of diagnostic systems. *Science,* **240,** 1285–1293.

——, and R. M. Pickett, 1982: *Evaluation of Diagnostic Systems—Methods from Signal Detection Theory.* Academic Press, 253 pp.