2.8 SPATIAL VERIFICATION USING THE RELATIVE OPERATING CHARACTERISTIC CURVE

Laurence J. Wilson and W. R. Burrows
Meteorological Service of Canada, Dorval, Québec

1. INTRODUCTION

The Relative Operating Characteristic Curve (ROC) measures the ability of a probabilistic or categorical forecasting system to discriminate between situations preceding the occurrence and the non-occurrence of an event of interest. It has been applied, for example, to the assessment of the ability of ensemble forecast systems to discriminate between the occurrence and non-occurrence of precipitation accumulations over specific thresholds. The ROC can be applied to any set of probabilistic forecasts of a dichotomous variable from any source.

Data used for the preparation of the ROC for verification of short range weather element forecasts is usually in the form of timeseries of observations and their corresponding forecasts. The forecasts may be probabilistic or categorical and the observations are binary, taking the value of 1 or 0 according to whether the event occurred or not in the valid period. The ROC measures the extent to which observations of the event correspond to relatively high probability forecasts and observations of the non-event correspond to relatively low probability forecasts. In other words the ROC and its associated statistics measure the separation of the two conditional distributions of forecast probabilities, conditional on the occurrence and non-occurrence of the event. Timing errors will in general appear as a lower ability to discriminate occurrences from non-occurrences to the extent that the offset in time results in smaller separation of the two conditional forecast distributions.

Suppose instead that the verification dataset were to be made up of gridded forecasts at a particular time, and corresponding observations. Could the ROC give a quantitative measure of the ability of the forecast to spatially discriminate the occurrence or non-occurrence of an event of interest? Would not the spatial mismatch between forecast and observation also be reflected in the measures of the separation of the two conditional distributions?

The spatial verification problem is further illustrated by Figure 1, which shows high resolution lightning forecasts at 45 to 48 h range and the corresponding observations. It is immediately clear to the eye that this forecast is quite good in a broad context: The large frontal band of thunderstorms has been well-forecast and even the smaller cell north of L. Superior has only been missed by 100 km or so. What is not clear from the map is to what extent the details of the lightning patterns have been accurately forecast. Is there any meaningful information in the details of the forecast at all? And, at what maximum resolution is there some skill in the forecast? Given that we have access to high resolution forecasts and observations, it might now be possible to answer these questions.

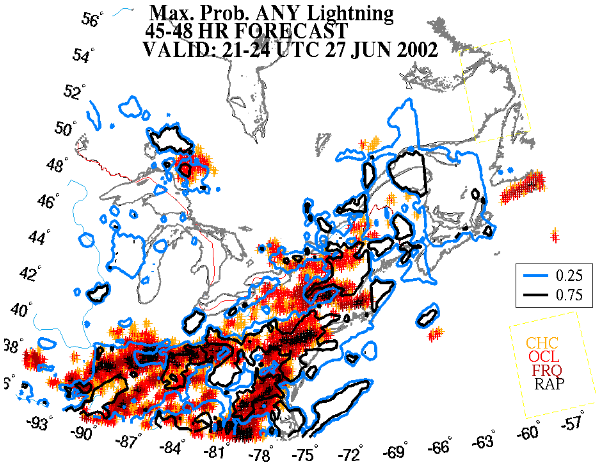

Figure 1. Example of probability of lightning forecast (contours) along with its verifying observations (coloured crosses). Darker colours indicate more frequent lightning occurrence.

This paper describes some experiments in the use of the ROC for spatial verification, to try to answer these questions. Forecasts are of the probability of occurrence of lightning (thunderstorms) over six hour periods, and the observations come from the Canadian lightning detection network. The data characteristics and processing methods are described in the next section, followed by a description and discussion of the results obtained so far.

2. DATA AND EXPERIMENT

The ROC verification experiments were carried out using forecasts of cloud to ground lightning occurrence, and corresponding observations from the Canadian Lightning Detection Network (CLDN) The basic predictand is the occurrence of lightning within 22 to 24 km of each grid point of the

Canadian operational GEM model, within a three hour period.

Forecasts are made for three-hour windows out to 48 h. The method is MOS-CART (Model output statistics (Glahn and Lowry, 1972) - Classification and Regression Trees (Breiman et al., 1984)). Predictors come from the GEM model, and the prediction trees are based on two years of data. Further details of the forecast technique and its verification can be found in Burrows et al. (2003).

The basic dataset for this experiment consists of forecasts for 3 h intervals out to 48 h (8 projections), twice per day for 5 summer months of 2003, at all the grid points of the GEM model for which lightning data is available (about the southern 2/3 of Canada). Not all the data has been processed as of this writing, therefore the results shown below are selective and representative of work in progress. Further results will be shown at the conference. As is typical of meteorological datasets, the nominal sample size is large, but the proximity of the data points in space and time limits the degrees of freedom contained in the dataset. Largely for convenience, but also because of differences in the climatology of convective processes, the dataset was divided into east and west domains for the evaluation.

The first step of the assessment was to calculate the ROC for single cases, using forecasts from all the grid points in the domain as the sample, then to plot the ROC area values as a function of time. This would give an idea of the variability of the ROC over a fixed domain from one day to the next. Each ROC was fitted using the normal-normal model, which assumes that the conditional distributions are transformable to normal by means of a monotonic transformation (Swets, 1986). The advantage of this approach over the empirical ROC is that the areas obtained are more consistent from one case to another (Wilson, 2000). It would be expected that a relatively high ROC would be associated with a forecast map where the areas of high probability match the locations of the storm locations better than cases where they don't.

The second step was to use both the ROC and the reliability table to determine the maximum resolution at which the spatial details of the forecast were meaningful. To examine this, we redefined the predictand by steps, steadily increasing the radial distance over which we searched for lightning occurrences. The basic verification was point-wise, forecasts at each point were matched with the occurrence (1) or non-occurrence (0) of lightning within 22 km. Radii of 25, 50, 100, 250 and 500 km were then used to redefine the predictand. For each radius, the probabilities at all points within that radius of each grid point were averaged, and matched to the occurrence or non-occurrence of any lightning anywhere within the circular domain defined by the radius. We also tested the maximum probability within the domain, which should effectively allow "near-misses" to count. That is, if a peak probability occurs 50 km away from a lightning observation, it will count as a match for radii greater than 50 km, but not less. The sample is created by stepping through all the grid points of the domain for each case, using each as the center of the selected domain. As the domains increase in size, the overlap in the data increases, so we considered it advisable to carry out this part of the investigation using at least a month of data.

As the spatial criterion for matching lightning occurrences with peaks in the probability distribution is relaxed, one should see higher values of the ROC area, and greater reliability, in the sense that the forecast is becoming less demanding. In the limit, with very large radii, spatial discrimination is effectively removed all together, and one is left with a measure of the forecast's ability to discriminate precipitation days from non-precipitation days.

3. RESULTS

Figure 2 shows the ROC area calculated for the eastern Canadian domain, as a daily time-series for one month, July 2003. It can be seen that the values of the ROC vary over quite a large range, from a high of 0.88 on the 9th to a low of 0.65 on the 11th.

Figures 3 and 4 are the forecast-observation maps which correspond to the maximum and minimum ROC area respectively. Although there is quite a lot of small scale variation in both the forecasts and the observations, there is nevertheless a clear tendency for the higher probabilities (inside the red contours) to correspond to lightning areas, with relatively few "false alarms" in Fig. 3. In Fig. 4, by contrast, the areas of high probability forecasts do not bear any clear resemblance to the observed lightning areas. The ROC area for this example, 0.65, is well below the practical minimum acceptable limit for useful discriminant ability.

.

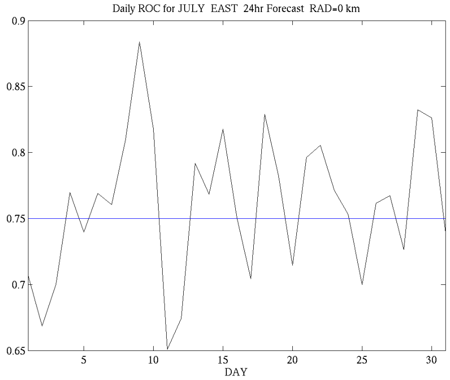

Figure 2. Daily ROC values calculated over the Eastern Canada grid for the month of July, 2003. A horizontal line is drawn at 0.75, indicating an approximate lower limit to usable discriminant power.

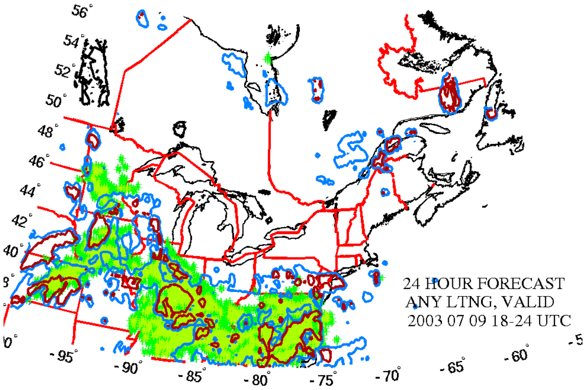

Figure 3. Lightning probability forecasts (blue contours=30% and dark red contours=75%) and corresponding observations (green and yellow shading) for 24h forecasts from 9 July, 2003. ROC area for this case was 0.88

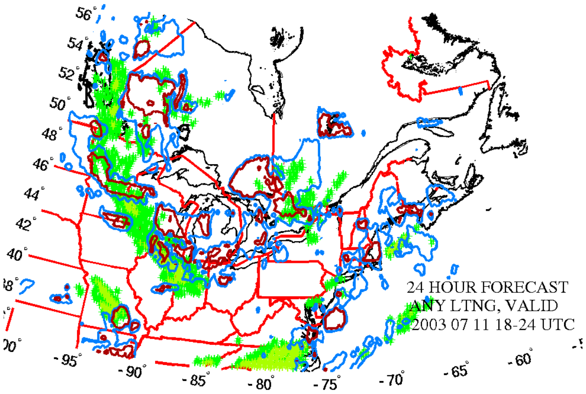

Figure 4. Same as Fig. 3, but for 11 July. ROC area was 0.65

Figure 5 shows the daily ROC areas for the month of July, again for the eastern region, but this time, with a 500 km radius. This means any occurrence of lightning within 500 km would be matched to the probability forecast either averaged over the whole area (dashed) or the maximum forecast probability anywhere in the area (solid). Comparing Fig. 5 to Fig. 2, there is evidence that relaxing the spatial criterion for the forecasts results generally in modest increases in discriminating ability, as expected. The best days do not change much, but discrimination on the “poorer” cases improves by a greater amount. Therefore, the overall day-to-day variation decreases. Perhaps most of the skill comes from an ability to discriminate active lightning days from inactive days. This could be checked by comparing the daily lightning frequencies with the ROC areas.

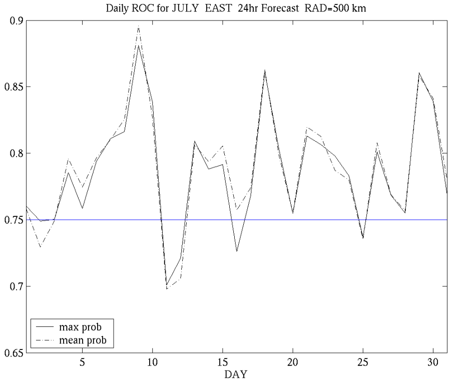

Figure 5. Daily ROC areas for Eastern Canada, for the month of July, 2003, for probability of lightning occurrence over circular areas of 500 km radius. Maximum probability in the domain (solid) and mean probability within the domain (dashed)

The curve for the mean probability is very similar to the curve for the maximum. Since the ROC is invariant with respect to bias in probability values, this probably means that the standard deviation of the probability distribution changes little from one day to another, and therefore the difference between mean and maximum probability varies little from one day to another.

The variation in ROC values as a function of domain radius is shown in Fig. 6 for the western region, for the month of July, 2003. In this case, the maximum probability over each circular domain of the given radius was matched with the occurrence of lightning anywhere in the domain. The domains were stepped through all the grid points of the domain as the center point, then pooled over the whole month of July for computation of the ROC. For Fig. 6 and the figures that follow, a 6 hour time window was used. That is, the maximum of the two three-hour probabilities was chosen as the forecast and matched with occurrences in the six hour period. This is essentially the temporal analogue to pooling over the spatial domains. In the temporal case and probably in the spatial case, it would be more strictly correct to take the sum of the probabilities over the component periods (or points) and subtract the joint probabilities. The latter are difficult to estimate, but one could perhaps do so using the long term climatology stratified by time of day and month. This is one of many subjects for future investigation.

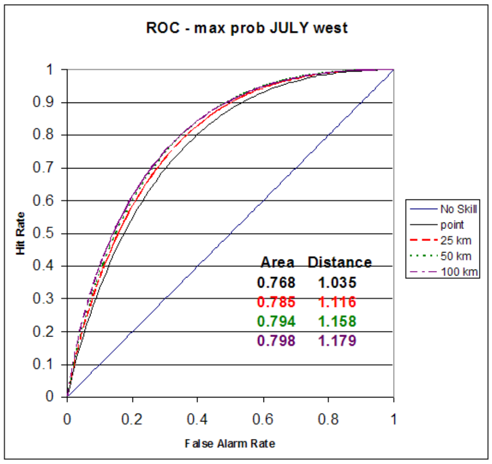

Figure 6. ROC curves for July lightning probability forecasts, for point (solid), 25 km (dash), 50 km (dot), and 100 km radius (dash-dot). See text for further explanation.

In Fig. 6, the curves are fitted with the normal-normal model, and areas and separation distances are shown in the inset table. The distances are in terms of the standard deviation of the conditional forecast distribution for non-occurrences. It is that distribution which is better estimated from the data. The figure shows that the discriminant ability of the forecast improves only slightly on average as the spatial matching is relaxed from 0 to 100 km. The areas are slightly above the lower limit of useful skill, but as indicated before, the principal source of discriminant ability probably comes from the method's ability to distinguish lightning days from non-lightning days in the sample and over larger scales, rather than from the ability to correctly determine the location of the lightning on a specific day to within 100 km. The fact that the ROC area for a 100 km radius is only slightly higher than for the individual point forecasts supports this hypothesis.

Although this paper is mainly about application of the ROC to spatial forecasts, made possible by access to both forecasts and observations at high resolution, we turn now briefly to verification of the same July set of forecasts using the reliability diagram. While the ROC relates to the likelihood-base rate factorization of the joint distribution of observations and forecasts, the reliability table refers to the other factorization, called calibration-refinement by Murphy and Winkler (1987).

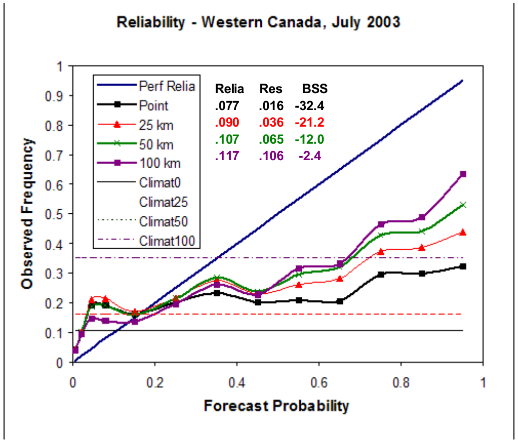

Figure 7. Reliability table for July probabilistic lightning forecasts, 24-h projection, for western Canada, for occurrences within domains of radius 0, 25, 50 and 100 km. Horizontal lines show the sample climatological frequency of occurrence for each radius. The reliability (relia), resolution (res) components of the Brier score, and the Brier Skill Score (BSS) are shown for each radius.

The reliability table shows that none of the forecasts are reliable. There is a tendency to underforecast the lower probabilities (below 15%) and a tendency to overforecast the higher probabilities at all radii. The point forecasts show practically 0 resolution (the curve is nearly horizontal), but the resolution does tend to increase as the matching radius is increased. At 100 km, the higher probabilities are still substantially overforecast, but there is a tendency for the reliability curve to parallel the 45 degree line, which means that these forecasts could be calibrated with a simple constant offset in the probabilities. Despite this tendency, the actual reliability component of the Brier Score tends to rise slowly (worse because of the negative orientation). Most likely this is due the effect of changes in the forecast frequency distribution as the radius increases. (Fig. 8). The resolution component rises faster, which more than offsets the negative effect of the decreasing reliability, and thus the Brier Skill Score goes up. (We have used the Brier Skill Score because changes in the uncertainty component of the Brier Score, due only to differences in the climatological frequency of occurrence of lightning in the sample, make it difficult to compare Brier scores.) This is perhaps the most positive result so far: With calibration, these forecasts, interpreted at 100 km resolution, might produce useful guidance. In reality, the ability to predict with some skill the occurrence of lightning within 100 km, 24 h in advance would be significant.

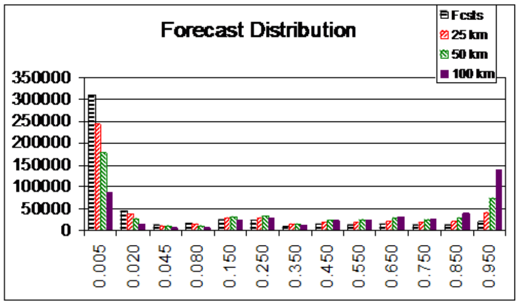

Figure 8. Distribution of 18-24-h probability forecasts of lightning occurrence, western Canada, July, 2003, for matching radii of 0, 25, 50 and 100 km.

4. Discussion

This paper describes some work in progress to evaluate the maximum resolution at which high resolution statistical lightning forecasts contain useful skill. More investigation is needed before we will be able to answer that question. In addition to testing the ROC on different combinations of spatial and temporal resolutions, we will evaluate changes as a function of projection time, and evaluate different stratifications of the data for computation of the ROC. It might be worthwhile to investigate ways of estimating the joint probabilities of lightning occurrence, both for adjacent projections and for adjacent grid points, in order to account for the lack of independence in the dataset and to enable the estimation of probabilities over larger spatial and temporal domains.

If the results answer the questions posed at the end of the introduction with some clarity, then the next step would be to go back and redevelop the statistical equations for the highest resolution at which skilful forecasts can be obtained, using smoothed or otherwise processed model predictors. This way, we can be sure to extract the maximum usable resolution from the operational models for statistical weather element prediction.

5. References

Breiman, L., J. H. Friedman, R. A. Olshen, and C. J. Stone, 1984: *Classification and Regression Trees*. CRC Press, ISBN 0412048418, 358 pp.

Burrows, W. R., Colin Price, and L. J. Wilson, 2003: Statistical models for 1-2 day warm season lightning prediction for Canada and the northern United States. Proceedings, 12th International Conference on Atmospheric Electricity, Versailles, France, 21-24.

Glahn, H. R. and D. A. Lowry, 1972: The use of model output statistics (MOS) in objective weather forecasting. *J. Appl. Meteor.*, **11**, 1203-1211.

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

Swets, J. A., 1986: Form of empirical ROC's in discrimination and diagnostic tasks: Implications for theory and measurement of performance. *Psychological Bulletin*, **99**, 181-198.

Wilson, L. J., 2000: Comments on "Probabilistic predictions of precipitation using the ECMWF ensemble prediction system". *Wea. Forecasting*, **15**, 361-364.