OBJECTIVE LIMITS ON FORECASTING SKILL OF RARE EVENTS

Harold E. Brooks

NOAA/ERL/National Severe Storms Laboratory

Norman, Oklahoma

Michael Kay

NOAA/NWS/Storm Prediction Center

Norman, Oklahoma

1. Introduction

Forecasting rare, severe weather events is challenging. Equally challenging, however, is the problem of developing meaningful verification procedures that can be beneficial from the standpoint of forecasters, forecast users, and the forecasting organization. A wide range of difficulties arises in this context, from collecting good observations of the phenomena in question to conveying information from the verification in a meaningful way.

In this paper, we will focus on one particular problem associated with the verification of rare event forecasts: development of appropriate baselines for skill given that forecast difficulty varies from situation to situation. Efforts to identify "no-skill" baselines date back to Gilbert (1884) and have focused primarily on the use of climatological (either sample or long-term) data. Recently, there have been efforts to include meteorological information to stratify the forecasts into "easy" or "hard" forecasts. For example, Brooks et al. (1996) looked at forecasts of freezing rain for standard observing sites and used only observations of winter precipitation in the data set in order to eliminate a large number of correct forecasts of null events. (We note that this was a strong constraint. Weaker constraints, such as only considering observations when temperatures were less than ~5 °C, would have allowed more correct null forecasts into the data set.) Such efforts attempt to limit the credit given to forecasters for making easy, correct forecasts, particularly when the observations are dominated by non-events of the element of interest. To quote Peirce (1884), "The value of the expert work must be measured by the excess which is obtained over the man who knows nothing of the subject."

For the verification of severe thunderstorm forecasts, particularly in the form of guidance products such as the the convective outlook and watch products from the Storm Prediction Center (SPC), the problem takes on additional complexity. Unlike the freezing rain forecast problem, almost all of the observations come in from volunteer spotters, so that there is no regular temporal and spatial order to the observations. Further, outlook and watch products are issued with the explicit expectation that there will be "false alarms" (parts of the forecast for which there are no events) and "missed detections" (events which are not included in the forecast). Thus, the expected range of values of the probability of detection (POD) or false alarm rate (FAR), for example, does not run from 0 to 1 in practice. Here, we will discuss the concept of a "practically" perfect (PP) forecast and apply it to artificial and real data sets.

By "practically" perfect, we mean a forecast that is consistent with a forecaster would make given perfect knowledge of the events beforehand. If, as in the case of outlooks and watches, there are explicit or implicit limits on the size of the product (e.g., watches are rarely smaller than 10,000 km^2) or if the forecaster has the goal of having a minimum number of reports within a forecast area before a product should be issued, then there will be false alarms and missed detections associated with the PP forecast. The PP forecast can then be used to estimate the minimum and maximum scores that a forecaster could reasonably obtain. In general, that range will be much smaller than the absolute minimum and maximum, but will provide a range over which forecast performance can be judged. Note that such a concept is not limited to any particular score that can be derived from a set of verification data, nor is it limited to dichotmous yes/no forecasts. It can be applied to any forecast measure and to probabilistic forecasts easily.

2. Methodology

In order to develop the PP forecast, we will begin with the reports of events, as recorded at the SPC. Reports of severe weather are recorded on a grid with each grid box representing an area 40 x 40 km. For now, we will consider all severe weather reports as equal and look at only whether a box has had an event or not. (The methodology could be extended to consider intensity and number of reports, but we will limit the procedure to the simplest case, for now.) The PP forecast is then created by smoothing the events using nonparametric density estimation with a two-dimensional Gaussian kernel (Silverman 1986).

Specifically, at each grid point in the domain, the PP forecast value, f, is given by

$$f = \sum_{n=1}^{N} \frac{1}{2\pi\sigma^2} \exp\left(\frac{1}{2}\left[\frac{d_n}{\sigma}\right]^2\right)$$

where d_n is the distance from the forecast grid point to the n-th location that had an event occur, N is the total number of grid points with events, and σ is a weighting function that can be interpreted as the confidence one has in the location of the forecast event. Increasing σ is is equivalent to increasing the uncertainty associated with the forecast as one would do with increasing lead time of the forecast. That is, in the context of severe weather forecasting, very small σ can be thought of as being associated with the warning stage, while larger σ is associated with the watch or convective outlook stages. The technique is similar to ones commonly used in objective

analysis, with the values at each grid point being limited to 1 or 0, depending on whether an event occurred there or not. The technique can be extended to include consideration of the number and intensity of events simply by redefining the “event” of interest, but that is beyond the scope of this paper.

The field of f gives an artificial forecast that is as accurate as could be expected for a forecaster knowing the locations of events with a confidence level associated with σ; it gives the probability that an event occurs in a given grid box. To illustrate the effects of the variations in σ, see <URL: http://www.nssl.noaa.gov/~brooks/prague5.gif>.By setting a threshold probability, the probabilistic forecasts can be converted to a dichotomous yes/no forecast of the event. [This makes for easier comparison to SPC products which are of a dichotomous nature (in the case of a watch) or take on a small number of values (convective outlook).] After that, a 2x2 contingency table can be developed for the forecasts and events and standard performance measures calculated. As the threshold probability increases for making a "yes" forecast, the POD decreases and the FAR increases. The CSI takes on a maximum value at some intermediate threshold. The value of the CSI at a threshold probability of 0 (always forecast yes) is equal to the areal coverage of the event. This represents one estimate of the lower bound on expected performance, which the forecaster could get by forecasting that the event would occur everywhere. A slightly greater lower bound can be found by noting that there is a large drop in CSI from a threshold probability of 1% to a threshold of 0%, and considering the value that CSI approaches as the threshold probability approaches zero. In this case, for $\sigma = 3$, the reasonable lower bound on CSI is ~0.12.

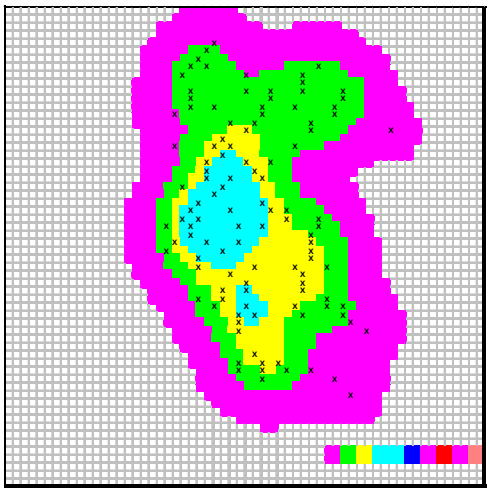

Fig. 1: Location (dots) of observed severe weather in 24 hour period beginning 1200 UTC 26 April 1991. PP forecast corresponding to σ=3 in color. Outside level 1-10% probability of severe weather. Successive levels: 11-20%, 21-30%, 31-40%.

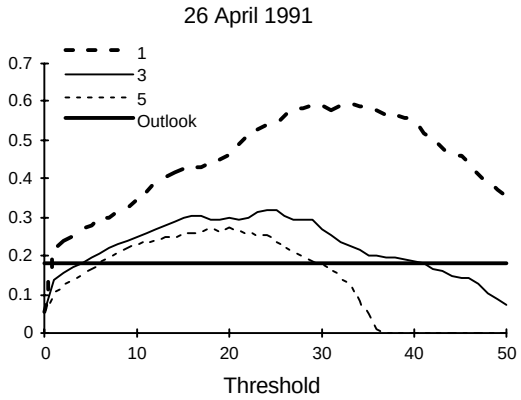

Fig. 2: CSI for 2x2 table for PP forecast for varying threshold probability values for σ=1, σ=3, and σ=3. CSI for actual convective outlook shown for reference.

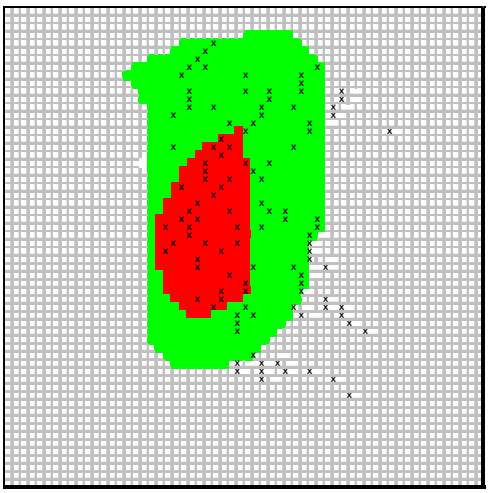

Fig. 3: Morning convective outlook and observed severe weather locations (dots) for 24 hour period beginning 1200 UTC 26 April 1991. Green shading indicates slight risk and red shading high risk.

As σ increases, the maximum in CSI generally decreases (Fig. 2). Assuming that a value of σ can be found which produces forecasts that "look" like real forecasts (e.g., similar areal coverage), the *practical* maximum value of CSI for that situation, given the nature of the forecast, can be obtained. Thus, the simple artificial forecast can be used to estimate the upper and lower bounds on the performance measure. For this example, for σ = 3, the CSI ranges from ~0.12 to 0.30 with the maximum at threshold probabilities between ~15-25%. For comparison, considering a convective outlook forecast of slight risk, or higher, to be a "yes" forecast, the value of CSI for the early convective outlook on 26 April 1991 (covering the same time period as the events shown in Fig. 1) was 0.18. This is ~33% of the way to the value of the PP forecast.The outlook area was shifted slightly to the west of the optimal location (Fig. 3).

3. Idealized Data

In order to investigate a wide variety of conditions, we can construct idealized event location data sets and look at the results of the PP forecasts associated with them. We have used two primary parameters in making the idealized data sets over a 60 x 60 domain--the density of coverage and whether the data are randomly distributed or if they are constrained to be more likely to occur near the diagonal and near the center of the domain. The idea behind the "constrained" data is to mimic severe weather outbreak situations.

As examples of the what can be learned from this effort, we present results from cases with dense coverage (~13% of the points in the domain, approximately the same effective coverage as for the 26 April 1991 outlook) and sparse coverage (~2%) for both random and constrained distributions.

Although the overall coverage in the dense constrained case is approximately the same as in the 26 April 1991, the points in the idealized case are more concentrated. As a result, the CSI saturates at a relatively low threshold value and stays at about the same level for a wider range of threshold values than for the real data (Fig. 4). Note that there is also a smaller difference between the CSI curves for σ =1 and 3 for the idealized case than for the real case. We interpret this as being an effect of the greater concentration of points. In the equivalent real world situation, where the forecaster has strong reason to believe that the events will be concentrated in a particular area, long lead-time (i.e., low confidence) forecasts can be almost as accurate as shorter lead-time forecasts. When there is no reason for a forecaster to constrain where events will occur, as in the random case, absolute performance is lowered. Just as significantly, however, the difference in peak performance between σ =1 and 3 increases. This makes intuitive sense for the parallel real world situation, resulting from higher confidence in location being possible at shorter lead time. Finally, the flatness of the CSI curve in the random case over thresholds from 0 to 20% illustrates the small range in forecast performance that are likely in situations where all that a forecaster knows is that severe weather will occur and doesn't have any reason to distinguish where it will occur. In such extremely difficult forecasts, the practical range of forecast performance is small.forecasts can be almost as accurate as shorter lead-time forecasts. When there is no reason for a forecaster to constrain where events will occur, as in the random case, absolute performance is lowered. Just as significantly, however, the difference in peak performance between σ =1 and 3 increases. This makes intuitive sense for the parallel real world situation, resulting from higher confidence in location being possible at shorter lead time. Finally, the flatness of the CSI curve in the random case over thresholds from 0 to 20% illustrates the small range in forecast performance that are likely in situations where all that a forecaster knows is that severe weather will occur and doesn't have any reason to distinguish where it will occur. In such extremely difficult forecasts, the practical range of forecast performance is small.

Fig. 4: As in Fig. 2, except for idealized situation with "dense" coverage (~13% of a 60 x 60 domain). "Random" indicates event locations are randomly distributed, while "constrained" is for the same number of event locations with constraints applied to the locations so that they lie along the diagonal of the domain near the center. This distribution resembles severe weather outbreaks. Numbers in parentheses (1 or 3) indicate value of σ for that curve.

Fig. 5: As in Fig. 4, expect for "sparse" coverage (~2% of a 60 x 60 domain).

When the coverage is less dense, a somewhat different picture appears (Fig. 5). In these cases, the random distribution allows for a high-certainty PP forecast that has a better CSI over a narrow range than for the constrained case. This is apparently because of the ability of the PP forecast to avoid false alarms. As σ increases, however, there is very little, in terms of performance as measured by CSI, that the PP forecast does. For $\sigma = 3$, there are no probabilities of 10% or greater, so that all forecasts are for "no" event and the CSI=0. For the constrained case with $\sigma = 3$, there is a broader range where positive CSI exists, due to the clustering of points brought on by the constraint that, in effect, means that a particular point is more (less) likely to have an event if its neighboring points have (does not have) an event, than in the completely random case. Thus, having some knowledge of where events might occur allows for better absolute forecast performance at long lead time.

4. Concluding Thoughts

The PP forecast for artificial data mimics many of the intuitive aspects of real forecast situations. As reasons to believe that events will occur at a particular location (e.g., constraint on location, density of coverage of events, forecaster confidence) increase, the absolute forecast performance increases. Also, when there is reason to believe where events will occur, there is a wider range of possible forecast performance. In other words, forecasters can demonstrate skill depending on their ability to identify conditions that constrain the likely locations of events.

The PP forecast also provides a range of reasonably achievable forecast performance. By comparing an actual forecast to the range expected from the PP forecast, we can estimate how far along the continuum from no-skill to perfect the forecast was. Note that this evaluation does not depend on any particular measure; the PP forecast can be used as the basis to provide bounds on any method of evaluating forecasts. In particular, it would be highly desirable to develop reliable asymmetric scoring rules for evaluating 2x2 contingency tables. CSI has the disadvantage of penalizing forecasters equally for missed detections and for false alarms. In many situations, the perceived costs of those two errors may well be very different. Methods that consider those differences are currently under development (Briggs, personal communication).

We plan to evaluate a long series of outlook and watch products from the SPC using the PP forecast as a basis for evaluation. Primary obstacles include methods for transforming the PP forecast into the polygonal form of watch products and in determining an appropriate value of σ for the different products. If those obstacles can be overcome, however, the PP forecast holds promise as a method for providing a consistent basis for evaluating the complete range of severe weather guidance products.

Another possible application of the PP forecast is in using them to consider a probablistic interpretation of events in conjunction with probabilistic forecasts. While the PP forecast has an inherent probaiblistic interpretation, events are typically viewed as dichotomous. Frequently, however, there may be reason to put more or less confidence in spotter reports and the absence of a report of an event does not necessarily mean that the event did not occur. For instance, radar "detections" of severe weather in locations where there are no spotters could be interpreted as having some probability of associated severe weather greater than 0. Then, the joint probability distribution of forecasts and events could be evaluated in the manner suggested by Murphy and Winkler (1987).

5. Acknowledgments

Although any shortcomings and misinterpretations are entirely our own fault, we thank Matt Briggs, Barb Brown, Chuck Doswell, and Dan Wilks for their comments and suggestions on this work.

6. References

Brooks, H. E., J. V. Cortinas Jr., and R. H. Johns, 1996: Experimental winter hazards forecasting at the Storm Prediction Center: Verification results for probabilistic freezing rain forecasts. *Preprints*, 15th AMS Conference on Weather Analysis and Forecasting, Norfolk, Virginia, American Meteorological Society, 127-130.

Gilbert, G. K., 1884: Finley's tornado predictions. *Amer. Meteor. J.*, **1**, 166-172.

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

Peirce, C. S., 1884: The numerical measure of the success of predictions. *Science*, **4**, 453-454.

Silverman, B. W., 1986: *Density Estimation for Statistics and Data Analysis*. Chapman & Hall, 175 pp.