Objective Limits on Forecasting Skill of Rare Events

Nathan M. Hitchens* and Harold E. Brooks

NOAA/OAR/National Severe Storms Laboratory, Norman, Oklahoma

Michael P. Kay+

Cooperative Institute for Mesoscale Meteorological Studies, University of Oklahoma, and NOAA/NWS/Storm Prediction Center, Norman, Oklahoma

(Manuscript received 18 October 2012, in final form 13 December 2012)

ABSTRACT

A method for determining baselines of skill for the purpose of the verification of rare-event forecasts is described and examples are presented to illustrate the sensitivity to parameter choices. These ''practically perfect'' forecasts are designed to resemble a forecast that is consistent with that which a forecaster would make given perfect knowledge of the events beforehand. The Storm Prediction Center's convective outlook slight risk areas are evaluated over the period from 1973 to 2011 using practically perfect forecasts to define the maximum values of the critical success index that a forecaster could reasonably achieve given the constraints of the forecast, as well as the minimum values of the critical success index that are considered the baseline for skillful forecasts. Based on these upper and lower bounds, the relative skill of convective outlook areas shows little to no skill until the mid-1990s, after which this value increases steadily. The annual frequency of skillful daily forecasts continues to increase from the beginning of the period of study, and the annual cycle shows maxima of the frequency of skillful daily forecasts occurring in May and June.

1. Introduction

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

In his essay on the nature of goodness in weather forecasting, Murphy (1993) defines ''quality'' (type 2 goodness) as the degree of correspondence between forecasts and events. Ten aspects of quality are defined, with ''accuracy'' and ''skill'' of particular interest when considering the forecasting of rare events. Accuracy is described as the average correspondence between individual pairs of forecasts and observations, while skill is described as the accuracy of forecasts relative to the accuracy of a forecast produced by a standard of reference.[1] So, while knowing the accuracy of a forecast is helpful, it may be more significant to know whether or not a forecast was skillful (and the degree of skill) in order to better understand the value added by a forecaster. For instance, a forecast that is accurate within 1.0° is more skillful when compared to a guidance forecast that is accurate within 5.0° than a guidance forecast that is accurate within 1.5°.

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

* Current affiliation: Department of Geography, Ball State University, Muncie, Indiana.

+ Current affiliation: AvMet Applications, Inc., Reston, Virginia.

Corresponding author address: Dr. Nathan M. Hitchens, Department of Geography, Ball State University, Muncie, IN 47306. E-mail: nmhitchens@bsu.edu

[1] Note that there may be a large number of metrics used to describe accuracy and skill depending upon the particular forecasting situation.

DOI: 10.1175/WAF-D-12-00113.1

climatological (either sample or long term) data. Such efforts attempt to limit the credit given to forecasters for making easy, correct forecasts, guessing, or forecasting the same thing all the time, particularly when the observations are dominated by nonevents 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.'' In effect, we would like to know the quality of a forecast system compared to some other standard, or baseline, forecast system.

The question of whether forecasters improve the quality of forecast guidance was examined by Ebert et al. (2004) in their verification of forecasts from the World Weather Research Programme's Sydney 2000 Forecast Demonstration Project. Forecasts for a variety of events, including the location of convection, were made using a number of radar-based nowcast algorithms, with persistence forecasts used as the no-skill baseline for the evaluation of the forecasts. No skill in this sense does not mean that the baseline forecast system is of low quality; merely that it performs at a level against which other forecast systems are to be compared. The authors observed that, in some instances, forecasters were able to make significant improvements to the persistence forecasts, but at other times forecaster intervention resulted in less desirable outcomes. It is also pertinent to note that although hail events were forecast during this project, only a single thunderstorm had hail reports associated with it, and of the 20 hail reports recorded, only 11 included usable size and location data. This illustrates the difficulties in verifying events in which reports are primarily made by the general public.

A challenge similar to Ebert et al.'s (2004) lack of hail reports was discussed in Brown et al.'s (1997) verification of in-flight icing algorithms. Pilot reports are the only actual measure of icing events, but two issues arise when considering the use of these reports for verification purposes. First, pilots are not required to report ''no icing'' conditions and have no incentive to do so. As a result the authors noted that only 25% of all pilot reports were for no-icing events. Second, regions where icing conditions are forecast to occur may be avoided by pilots due to the inherent safety issues accompanying such conditions. Thus, the authors differentiated between ''pilot reports'' and ''icing events'' and described the implications this had on the verification measures used.

For the verification of severe thunderstorm forecasts, particularly in the form of guidance products such as the convective outlooks issued by the Storm Prediction Center (SPC), the problem of verifying rare events takes on additional complexity. Much like the challenges faced by Brown et al. (1997) and Ebert et al. (2004), almost all of the observations come in from volunteer spotters, so that there is no regular temporal and spatial order to the observations. The existence of a forecast that an event is likely to occur could increase the likelihood that observers will be present to collect observations compared to a forecast of no event. Further, outlook 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 (Brooks et al. 1998; Davis and Carr 2000) and apply it to the SPC's categorical convective outlook products, specifically their ''slight risk'' areas, which we will refer to henceforth using the broad term ''outlooks.'' The dataset of outlooks used herein is described in Hitchens and Brooks (2012), with the addition of data from 2011.

By practically perfect,[2] we mean a forecast that is consistent with that which a forecaster would make given perfect knowledge of the reported events beforehand and the operational constraints associated with the forecasting system. If, as in the case of outlooks, there are explicit or implicit limits on the size of the product (e.g., outlooks are rarely smaller than 50 000 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 maximum score, and as will be shown, also the minimum score 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 dichotomous (yes–no) forecasts. It can be applied to any forecast measure and to probabilistic forecasts easily.

[2] The term ''practically perfect'' draws on the usage ''practical zero,'' in which a person offering a judgment on the probability of a very unlikely event may describe it as zero, even though they do not think the probability is exactly zero. The probability is sufficiently low to be regarded as zero in typical applications. Similarly, the ''practically'' in practically perfect does not mean that the forecast is *almost* perfect, but that the forecast is as good as could be expected in typical practice.

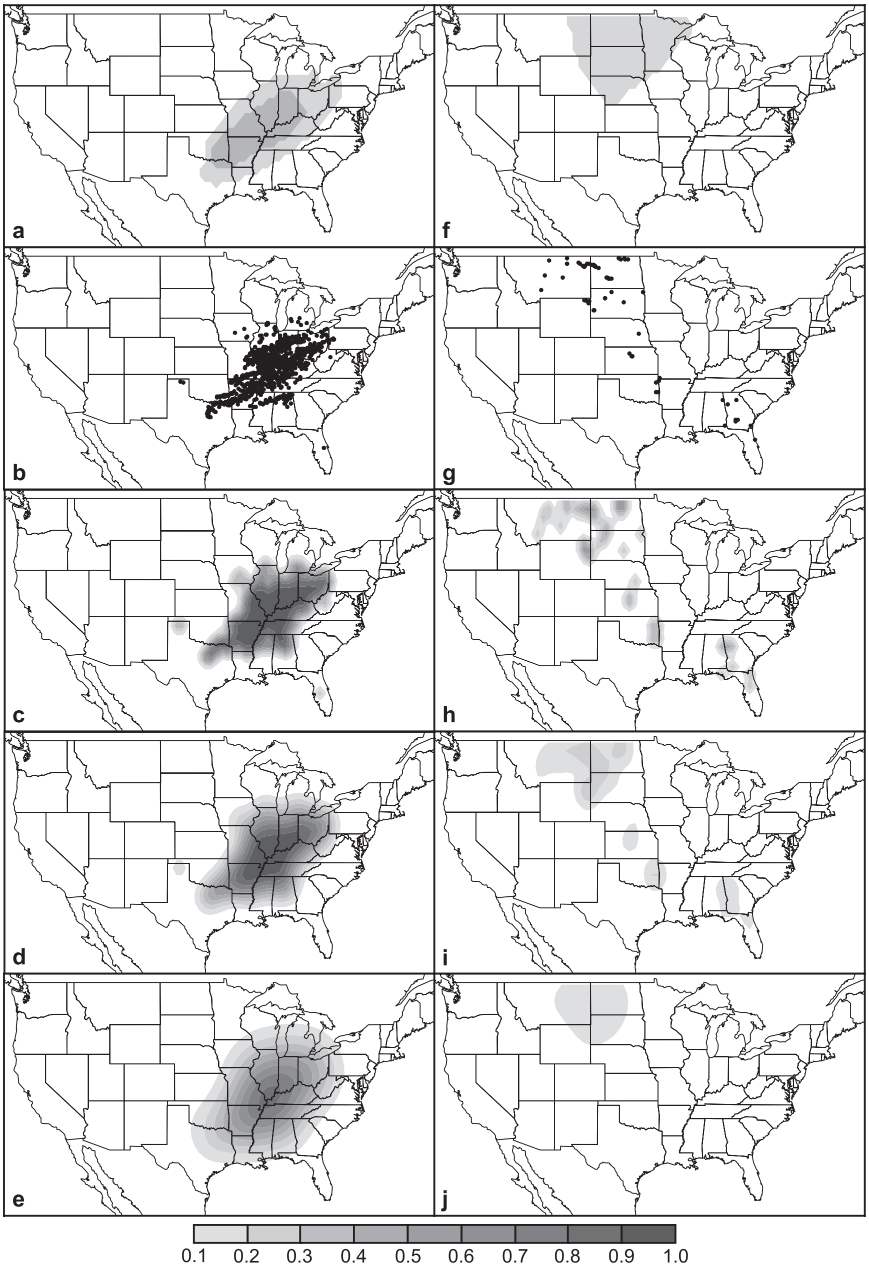

FIG. 1. (a),(f) SPC convective outlooks, (b),(g) locations of storm reports, and PP forecasts for σ = (c),(h) 0.75, (d),(i) 1.5, and (e),(j) 3.0 for the 24-h periods beginning at 1200 UTC on (a)–(e) 19 Apr 2011 and (f)–(j) 1 Aug 2010. Light shading in (a),(b) denotes slight risk areas and darker shading indicated moderate risk areas.

2. Practically perfect forecasts

To develop the PP forecast, we will begin with the reports of events, as recorded at the SPC. Reports of severe weather are put on a grid with each grid box representing an area approximately 80 km × 80 km, roughly equivalent to the area associated with SPC forecast definitions of the probability of an event occurring within 25 miles (mi) of a point. 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 the intensity and number of reports, but we will limit the procedure to the simplest case.) 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

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

where d_n is the distance from the forecast grid point to the nth 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 equivalent to increasing the spatial uncertainty associated with the forecast as one would do with increasing lead time of the forecast. That is, within 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 field of f gives an artificial forecast that is as good 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 impact of the choice of the σ value (0.75, 1.5, or 3.0), we examine events representing the two extremes on the spectrum of severe weather forecast scenarios: a large ''outbreak'' (case 1; Figs. 1a–e) and a relatively small number of reports with little spatial concentration (case 2; Figs. 1f–j). There are 1045 severe weather reports covering 209 grid boxes in case 1, and 73 reports covering 18 grid boxes in case 2. For both cases the PP forecast size decreases with increasing probability threshold values, with the rate of decrease being larger at higher σ values (Fig. 2). Concentrating on each σ value, for 0.75 the size of the 0.01 or greater probability area for case 1 (case 2) is 1.79 (1.27) million km^2, and reduces to no area at a probability of 94% (57%). Likewise, for a σ value of 1.5 (3.0), the size of the 0.01 probability area is 3.04 (4.76) million km^2 for case 1, and 2.60 (4.00) million km^2 for case 2. The probability threshold at which it reduces to no area for case 1 is 100% (84%) and is 28% (17%) for case 2.

The choice of the value of σ should at least in part be made with the goal of emulating the range of sizes made by forecasters. A comparison of quantiles of outlook size and PP forecast size from 1982–91 and 2002–11 (Fig. 3) reveals a shift from outlooks being more closely associated in size to PP forecasts at the 5% threshold to PP forecasts at the 10% threshold. Forecasts from the mid-1990s were purposely excluded due to apparent changes in forecasting philosophies that significantly impacted the size of outlook areas (Hitchens and Brooks 2012). Quantiles of the first two σ values (0.75 and 1.5) correspond well to the line $y = x$, with noticeable curvature toward the y axis at lower quantiles, suggesting that there is greater variability among outlook sizes at these quantiles. The quantiles for $\sigma = 3.0$ are sharply curved toward the x axis at lower quantiles, indicating more dispersion among PP forecasts at these quantiles and, in general, show much less correspondence to the line $y = x$ compared to the other σ values.

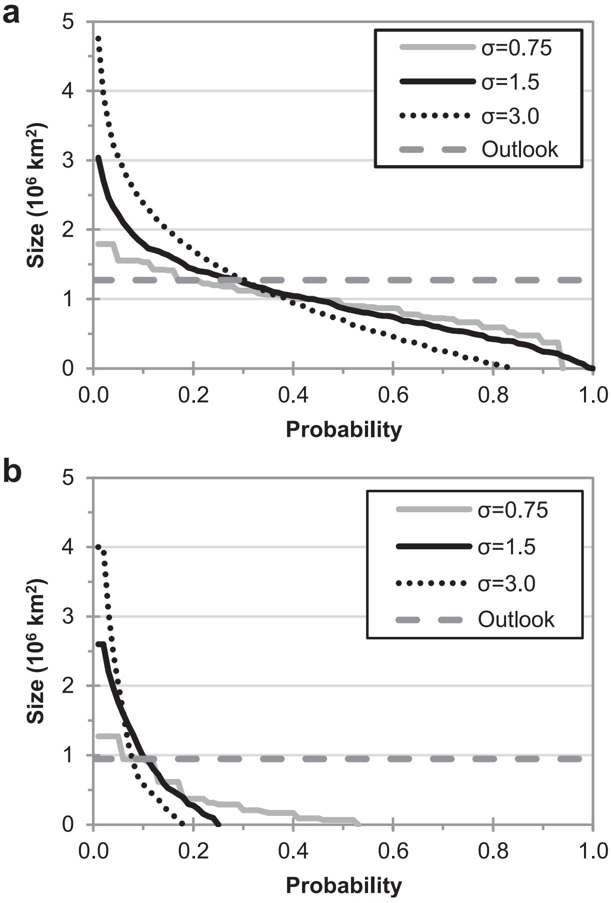

FIG. 2. The PP forecast size by probability threshold for the 24-h period beginning at 1200 UTC on (a) 19 Apr 2011 and (b) 1 Aug 2010. The size of the SPC's outlook area for the same period is displayed as a dashed horizontal line.

By considering each increasing PP forecast probability value as an individual threshold, we can convert the range of probabilities into a set of dichotomous (yes or no) forecasts of severe weather. This allows for the development of a 2 × 2 contingency table for each probability threshold (Table 1) and the calculation of standard measures of performance [see Doswell et al. (1990) for a complete description of the measures used herein]. For each probability threshold the critical success index (CSI), a performance measure that compares the number of correct forecasts to the union of all forecasts and observations, is calculated. The value of the CSI at a

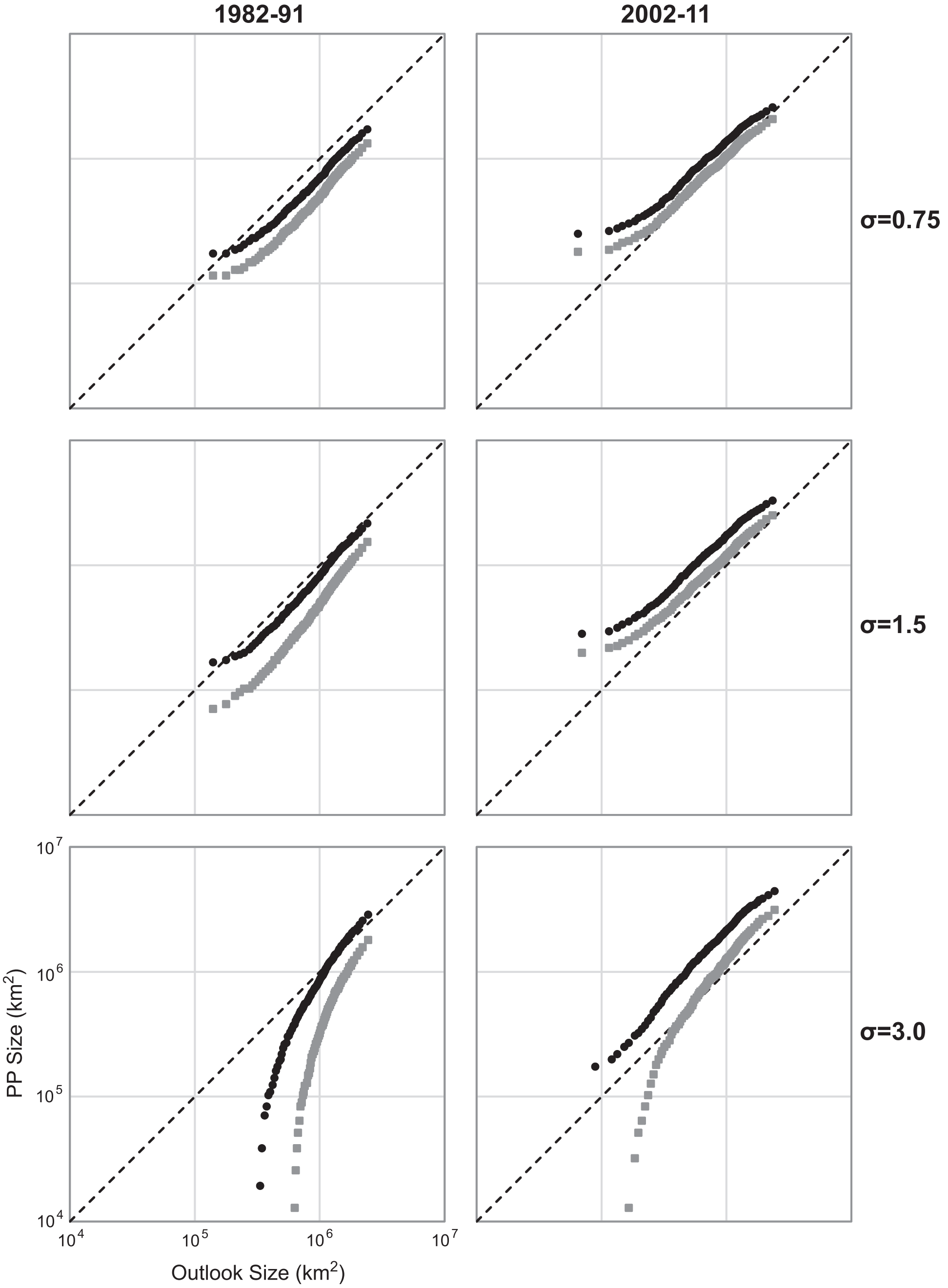

FIG. 3. The *Q–Q* plots comparing quantiles of the size of the SPC's outlook areas and the PP forecast 5% (black circles) and 10% (gray squares) probability areas during (left) 1982–91 and (right) 2002–11 for σ = (top) 0.75, (middle) 1.5, and (bottom) 3.0. The line $y = x$ is displayed as a black, dashed line.

probability threshold 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 attain 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 probability threshold of 1% to a threshold of 0%, and considering the value that CSI approaches as the probability threshold approaches zero. The 2% and 1% CSI values are used to linearly extrapolate this adjusted lower bound. In the two cases described above the CSI values at the 0% threshold are 0.06 and 0.02 for cases 1 and 2, respectively, while the values for the adjusted lower bound are 0.29 and 0.08 (for $\sigma = 1.5$).

TABLE 1. A 2 × 2 contingency table for forecasts and observations.

	Observed yes	Observed no	Sum
Forecast yes	*a*	*b*	*a* + *b*
Forecast no	*c*	*d*	*c* + *d*
Sum	*a* + *c*	*b* + *d*	*n*

Quantities of interest: probability of detection (POD) = *a*/(*a* + *c*), frequency of hits (FOH) = *a*/(*a* + *b*), critical success index (CSI) = *a*/(*a* + *b* + *c*), and bias = (*a* + *b*)/(*a* + *c*).

As σ increases, the maximum value of CSI generally decreases (Fig. 4). Assuming that a value of σ can be found that 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. Based on the size comparison above, $\sigma = 3.0$ would not be a suitable choice, and of the remaining two values, $\sigma = 1.5$ better represents the outlooks issued by the SPC. For example, the maximum CSI value attained using $\sigma = 0.75$ in case 1 (case 2) is 0.90 (0.84), while for $\sigma = 1.5$ the maximum is 0.78 (0.31). Not only does $\sigma = 1.5$ provide a more reasonable upper bound, especially in the case of an event that is less concentrated (e.g., Fig. 4b), but it also compares better visually to the SPC's outlooks (Fig. 1). Thus, $\sigma = 1.5$ will be used hereafter when PP values are calculated. Using this approach to define practical lower and upper bounds for severe weather forecasts, we can determine the skill of a forecast by calculating its relative position between the two bounds. In our two cases the relative skill of the first case is 0.71 and 0.13 for the second, providing increases from their CSI values of 0.64 and 0.11.

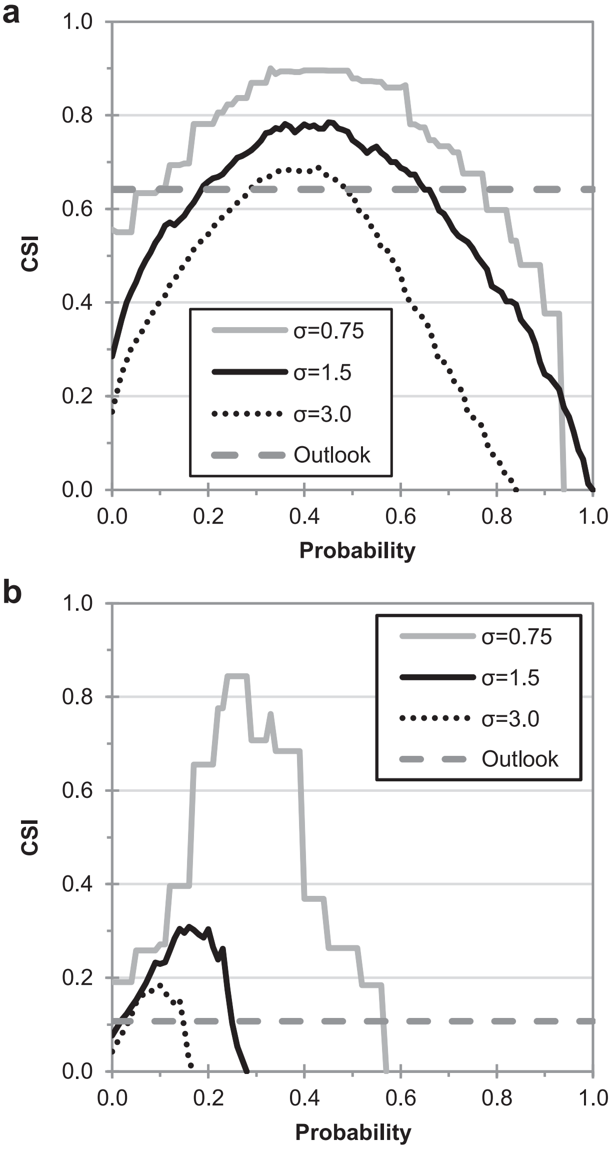

FIG. 4. CSI value by PP forecast probability threshold for the 24-h period beginning at 1200 UTC on (a) 19 Apr 2011 and (b) 1 Aug 2010. The CSI value calculated from the SPC's outlook area for the same period is displayed as a dashed horizontal line.

3. Application to categorical forecasts

Expanding this analysis to the entire digital dataset of SPC outlooks (1973–2011), it is seen from a 365-day running mean[3] that these forecasts show little to no skill until 1987, after which the relative skill remains above zero (Fig. 5). Further, until the mid-1990s the relative skill shows little variation, ranging between −0.03 and 0.06, but beginning in 1995 there is a steady increase until the end of the period of record, with a maximum relative skill of 0.25 in 2010. This corresponds with the improvements in POD and frequency of hits (FOH) observed in Hitchens and Brooks (2012) for the same forecasts. A number of factors could have contributed to this increase in relative skill, including advancements in numerical weather prediction, changes in forecasting philosophies within the SPC, improved physical understanding of severe thunderstorms, and the completion of upgrades to the national radar network. It is impossible to determine the relative importance of these, or other, factors, particularly considering that we cannot put strong limits on the interannual variability of the difficulty of the forecasts. Through the first 25 yr of data the CSI value for

[3] In this study 365-day running means are computed by constructing a 2 × 2 table that sums all 365 forecasts centered on each day. In the case of maximum CSI from PP forecasts, the 2 × 2 table associated with each day's maximum CSI value is used in the construction of the table for the 365-day period.

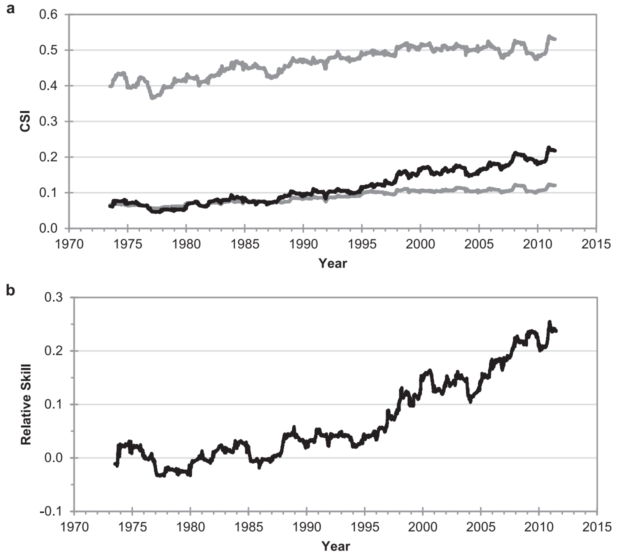

FIG. 5. (a) CSI values calculated using a 365-day running mean. The maximum and minimum CSI values from PP forecasts are plotted as gray lines and the CSI value from SPC outlook areas is plotted as a black line. (b) The skill of the SPC's outlook areas is calculated as the relative position of the outlook CSI value between the corresponding maximum and minimum CSI values from the PP forecast.

both the upper and lower bounds increased steadily, by ~0.10 for the upper bound and ~0.04 for the lower, but from 1998 onward neither displayed a noticeable trend. Likewise, during the first two decades the CSI of the outlooks increased at nearly the same rate as the lower bound, but after 1995 it increased at a higher rate than the lower bound, accounting for the increase in relative skill over that time.

The SPC outlooks and the practical upper and lower bounds can be further analyzed using a performance diagram of annual mean values to examine additional measures related to these forecasts (Fig. 6). As expected, the lower bound has a POD of 1.0 since these PP forecast areas include all reports, but these large areal extents limit the FOH to 0.05–0.11. Of more interest is the upper bound, which shows a moderate increase in POD through time (0.50–0.69), but exhibits the most improvement in FOH (0.38–0.65). While the POD of the outlooks increased significantly during the first 20 yr of the study period, it has not displayed a steady improvement in the remaining years, but only trails the upper bound by 0.12 for 2011. Although the FOH values of the outlooks have been increasing since the mid-1990s, the best value (0.26 in 2011) has not yet reached the lowest value of the upper bound (0.38 in 1977) and trails the FOH of 2011 (0.63) by 0.37. The recent improvement in the performance of the outlooks, as well as the potential for continued improvement, can be attributed to the placement of outlook areas (i.e., correctly forecasting as many events as possible while seeking to minimize outlook size).

Another useful application of the upper and lower bounds is for investigating how often individual outlooks are skillful (i.e., positive, nonzero values of relative skill). Throughout the period of record the annual frequency of skillful forecasts has increased steadily from 22% in 1973 to 75% in 2011 when considering all forecast days,[4] with a maximum of 78% during 2008 (Fig. 7). These frequencies are slightly higher when considering only days with both an outlook and at least one observation, with a maximum of 86% during 2009. The annual cycle of skillful forecast frequency for all days (Fig. 8) shows a peak in May (55%) during the 10-yr period covering 1982–91, and for 2002–11 the peak frequency occurs in June (86%), with a distinct secondary peak in November (69%). For both decades January is the month with the lowest frequency of skillful forecasts (23% and 54%, respectively). It should be noted that the entirety of the frequencies from the latter period are above 50%, while only two (April and May) from the prior period exceed this number. Frequencies in the first decade are impacted by a larger proportion of false alarms and missed events during the cool season (October–February) compared to the latter decade, while the proportions of these events were similar in both decades for the remaining months. The 90% confidence interval for each decade is constructed from 101 trials with random samples without replacement sized at half the number of events for a particular month. These confidence intervals suggest that the frequencies from each decade are well separated, with larger intervals during the cool season in both periods.

Over the same two 10-yr periods, the frequency of the relative skill values for all days is calculated (Fig. 9) illustrating an increase in skillful forecasts from 44% to 73%. This increase is the result of forecasts that showed small negative (0.0 to −0.2) to small positive (≤0.1) skill during 1982–91 instead showing more skill in the moderate to high positive values (0.2–0.6). From a forecasting perspective this change in the distribution of relative skill is partly caused by the reduction of false alarms by 58% (a reduction in frequency of 0.03), which are considered to have zero relative skill, as well as increases in the skill of forecasts with relative skill values of at least −0.2 by

[4] The term ''all forecast days'' includes days when an outlook was issued and no reports were recorded (''false alarm''), and days when no outlook was issued but reports were recorded (''missed events''). In the latter scenario the area of the upper bound must be at least as large as the smallest regular outlook area (~64 000 km^2).

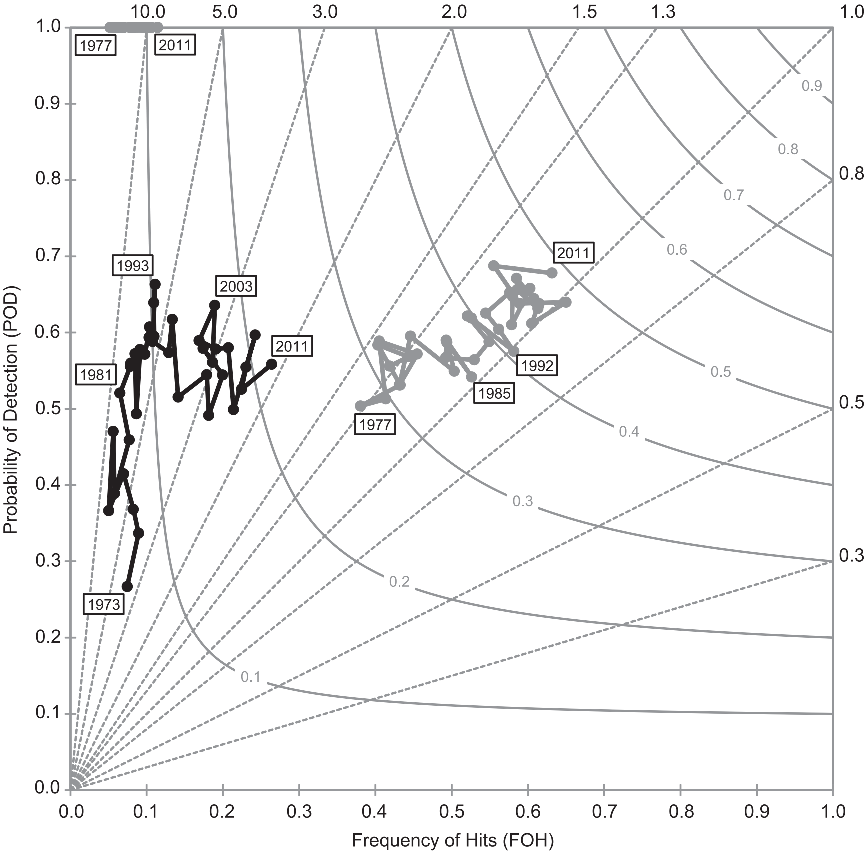

FIG. 6. Performance diagram (Roebber 2009) showing annual performance from 1973 to 2011 for outlook areas (black circles and lines) and PP forecasts (gray circles and lines) in terms of POD and FOH. The dashed lines represent bias (*B*), while the curved lines show CSI. The (center) maximum and (top left) minimum annual PP forecasts are determined using annual CSI values. Labeled years are provided for context.

making better forecasts (namely increasing FOH). Another interesting change observed in Fig. 9 is the increase in the frequency of forecasts with relative skill values between −0.2 and −0.3. This increase (0.0228) is entirely the result of a rise in missed events (up 49%), contributing 0.0235 to the frequency of the relative skill values in that range. The discrepancy is likely due to slight improvements in forecasts with accompanying observations (i.e., forecasts that were not ''false alarms'') between the two decades. Some of these ''missed events'' are rather substantial, with 46% having at least 20 grid boxes containing reports, and 5% with at least 50 grid boxes. However, the majority of these apparent misses are forecast by the SPC as ''see text''; locations where a threat of severe weather exists, but that threat is not sufficient to issue a slight risk. These forecasts were first issued publicly beginning in 1999, and their spatial extents are not explicitly defined.

As demonstrated in the performance diagram, improvements in CSI (and relative skill) are influenced more by improvements in FOH. This relationship is supported in a plot of FOH values as a function of relative skill (Fig. 10), with a 0.77 coefficient of determination (R^2). For relative skill values from just below 0.0 to

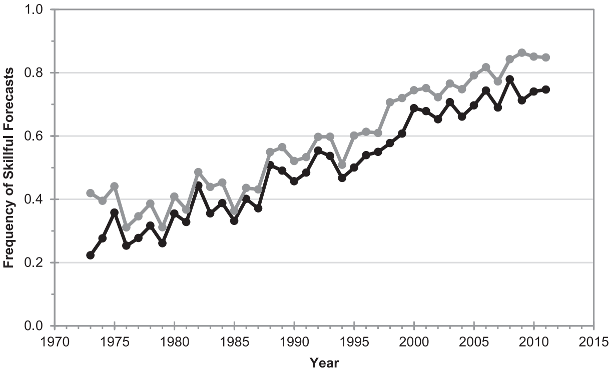

FIG. 7. Frequency of skillful daily forecasts by year for all days (black circles and line) and only those days in which both an outlook was issued and severe weather was reported (gray circles and line).

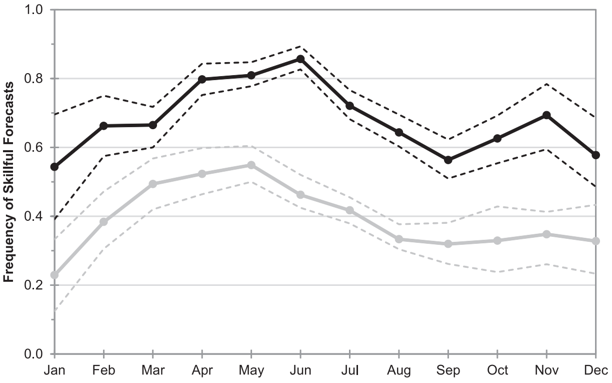

FIG. 8. Frequency of skillful daily forecasts by month (lines with circles) with 90% confidence intervals (dashed lines) for 1982–91 (gray) and 2002–11 (black).

approximately 0.2 there appears to exist a well-defined minimum threshold of FOH values. In contrast, there is no significant relationship between the size of the outlook areas or the observation areas and relative skill (Fig. 11). Although no relationship exists with the size of outlook areas ($R^2 = 0.001$), there is a tendency for outlook sizes related to more extreme relative skill values to be smaller relative to some outlooks with relative skill values near zero. Outlooks with smaller areas are more likely to have extreme relative skill values since the FOH plays such a large role in the calculation; a small-sized outlook that misses a very localized cluster of reported events will have an FOH and POD of zero, while in the same scenario, if the small-sized outlook is well placed, it will have a large FOH and POD. Larger outlooks are less likely to have greater values of FOH due to a higher likelihood of false alarms. On the other hand, the size of the observation areas ($R^2 = 0.22$) is relatively uniform across the range of relative skill values, suggesting that forecasters do not seem to perform better (or worse) based on the areal extent of an event.

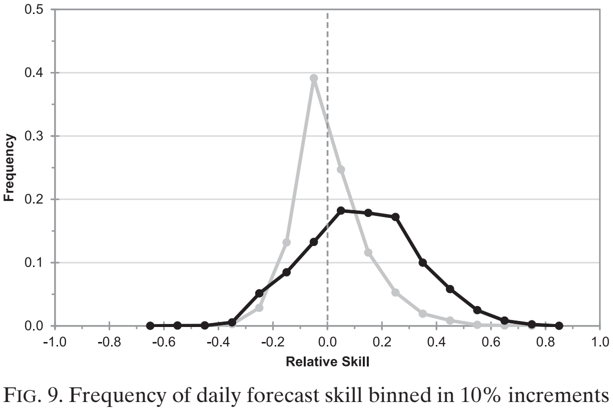

FIG. 9. Frequency of daily forecast skill binned in 10% increments for 1982–91 (gray) and 2002–11 (black).

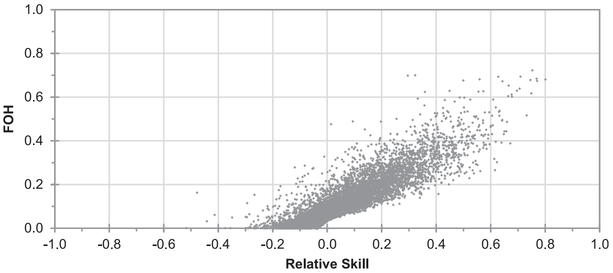

FIG. 10. Distribution of outlook FOH values by daily forecast skill for days in which both an outlook was issued and severe weather was reported.

4. Concluding remarks

The primary objective of this paper was to develop objective, practical baselines for forecasts of rare events. This is accomplished by using practically perfect forecasts to identify a ''no skill'' lower bound and a practical upper bound of CSI values for any particular forecast. The position of the CSI value of a forecast relative to the values attained from a PP forecast indicates the skill of that forecast. The choice of σ in the calculation of PP forecasts plays an important role, and can be effectively used to simulate the level of uncertainty inherent to a particular forecast (e.g., convective outlooks versus severe weather watches). For the purpose of evaluating the SPC's convective outlooks, $\sigma = 1.5$ was chosen based on a comparison of the sizes of PP forecasts and outlooks, as well as the range of CSI values for the PP forecasts.

Analysis of convective outlooks issued from 1973 to 2011 using PP forecasts for CSI baselines revealed that

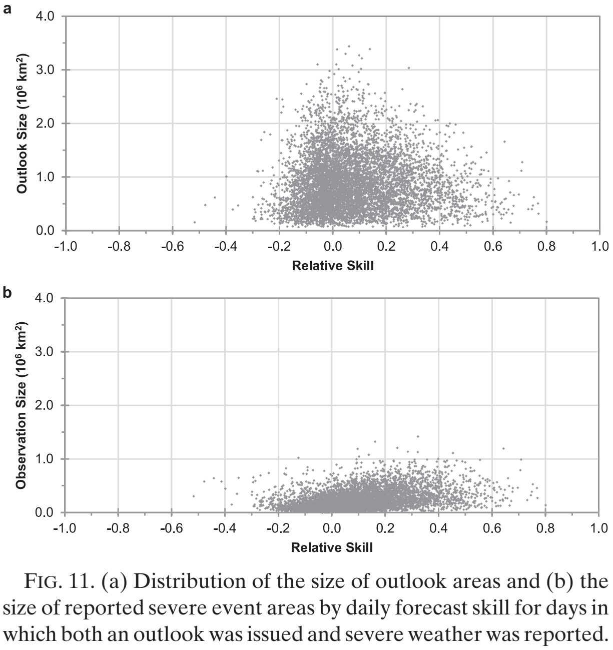

FIG. 11. (a) Distribution of the size of outlook areas and (b) the size of reported severe event areas by daily forecast skill for days in which both an outlook was issued and severe weather was reported.

these forecasts showed little to no relative skill for two decades, but after 1995 the relative skill of the outlooks has shown steady improvement. The timing of this increase in relative skill is not unexpected since Hitchens and Brooks (2012) found that FOH values for these forecasts began to increase at a similar point in time. Additionally, the annual frequency of skillful forecasts has continued to increase, with at least 50% of forecasts each year since 1995 showing some skill.

We plan to extend the analyses using PP forecasts to the full suite of convective outlook products—the ''day 2'' and ''day 3'' forecasts valid for the same 24-h period examined in this study, and the various updates to the 0600 UTC ''day 1'' outlook—and the probabilistic convective outlooks that were introduced in the early 2000s. Future work will also explore approaches to include the number and intensity of events within the PP forecast framework.

Acknowledgments. This research was performed while the first author held a National Research Council Research Associateship Award at the National Severe Storms Laboratory. The impetus for the origins of the research date back to conversations the second author had with Allan Murphy in the mid-1990s. The constructive comments and suggestions made by the three anonymous reviewers helped improve the manuscript.

REFERENCES

Brooks, H. E., M. Kay, and J. A. Hart, 1998: Objective limits on forecasting skill of rare events. Preprints, *19th Conf. on Severe Local Storms,* Minneapolis, MN, Amer. Meteor. Soc., 552–555.

Brown, B. G., G. Thompson, R. T. Bruintjes, R. Bullock, and T. Kane, 1997: Intercomparison of in-flight icing algorithms. Part II: Statistical verification results. *Wea. Forecasting,* **12,** 890–914.

Davis, C., and F. Carr, 2000: Summary of the 1998 workshop on mesoscale model verification. *Bull. Amer. Meteor. Soc.,* **81,** 809–819.

Doswell, C. A. III, R. Davies-Jones, and D. L. Keller, 1990: On summary measures of skill in rare event forecasting based on contingency tables. *Wea. Forecasting,* **5,** 576–585.

Ebert, E. E., L. J. Wilson, B. G. Brown, P. Nurmi, H. E. Brooks, J. Bally, and M. Jaeneke, 2004: Verification of nowcasts from the WWRP Sydney 2000 Forecast Demonstration Project. *Wea. Forecasting,* **19,** 73–96.

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

Hitchens, N. M., and H. E. Brooks, 2012: Evaluation of the Storm Prediction Center's day 1 convective outlooks. *Wea. Forecasting,* **27,** 1580–1585.

Murphy, A. H., 1993: What is a good forecast? An essay on the nature of goodness in weather forecasting. *Wea. Forecasting,* **8,** 281–293.

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

Roebber, P. J., 2009: Visualizing multiple measures of forecast quality. *Wea. Forecasting,* **24,** 601–608.

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