# Determining Criteria for Missed Events to Evaluate Significant Severe Convective Outlooks

NATHAN M. HITCHENS

*Department of Geography, Ball State University, Muncie, Indiana*

HAROLD E. BROOKS

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

(Manuscript received 22 September 2016, in final form 21 April 2017)

ABSTRACT

Among the Storm Prediction Center's (SPC) probabilistic convective outlook products are forecasts specifically targeted at significant severe weather: tornadoes that produce EF2 or greater damage, wind gusts of at least 75 mi h$^{-1}$, and hail with diameters of 2 in. or greater. During the period of 2005–15, for outlooks issued beginning on day 3 and through the final update to the day 1 forecast, the accuracy and skill of these significant severe outlooks are evaluated. To achieve this, criteria for the identification of significant severe weather events were developed, with a focus on determining days for which outlooks were not issued, but should have been based on the goals of the product. Results show that significant tornadoes and hail are generally well identified by outlooks, but significant wind events are underforecast. There exist differences between verification measures when calculating them based on 1) only those days for which outlooks were issued and 2) days with outlooks or missed events; specifically, there were improvements in the frequency of daily skillful forecasts when disregarding missed events. With the greatest number of missed events associated with significant wind events, forecasts for this hazard are identified as an area of future focus for the SPC.

## 1. Introduction

Beginning in the early 2000s, the National Weather Service's Storm Prediction Center (SPC) began issuing probabilistic convective outlooks for individual severe weather hazards (tornado, wind, and hail) alongside their day 1 categorical outlook products, while probabilities of all severe weather are forecast for products issued on days 2 and 3. Included as part of these probabilistic outlooks are forecasts specifically targeted at significant severe weather: tornadoes that produce EF2 or greater damage, wind gusts of at least 75 mi h$^{-1}$, and hail with diameters of 2 in. or greater (Hales 1988). Significant severe areas define regions where a forecaster believes a 10% or greater probability exists for these high-impact events. In this study, the accuracy and skill of significant severe outlooks are evaluated over 2005–15 and compared with other outlook products. Both accuracy and skill were defined by Murphy (1993) as the "average correspondence between individual pairs of forecasts and observations" and "accuracy of forecasts of interest relative to accuracy of forecasts produced by standard of reference," respectively. Because of the rarity of significant severe weather hazards, special attention is placed on the approach to defining and identifying missed events, since the SPC does not explicitly define minimum criteria for outlooks to be issued.

Prior research focusing on the SPC's outlooks has identified important trends in these products' performance. For instance, Hitchens and Brooks (2012) identified a change in forecast accuracy for the SPC's categorical day 1 outlooks in the mid-1990s, suggesting forecasters were reducing the size of outlook areas, while also placing them better, and improving the false alarm ratio, with little effect on the probability of detection. Further, in a study detailing a method by which synthetic forecasts, based on observed events, could assess forecast skill, Hitchens et al. (2013) showed that SPC forecasters became more skillful at issuing these outlooks, with the same mid-1990s identified as an

*Corresponding author*: Dr. Nathan M. Hitchens, nmhitchens@bsu.edu

DOI: 10.1175/WAF-D-16-0170.1

TABLE 1. A 2 × 2 contingency table for forecasts and observations. Quantities of interest: POD = $a/(a + c)$, false alarm ratio (FAR) = $b/(a + b)$, and FOH = $a/(a + b)$ = 1 − FAR. CSI = $a/(a + b + c)$ and bias = $(a + b)/(a + c)$.

| | Observed yes | Observed no | Sum |
|---|---|---|---|
| Forecast yes | $a$ | $b$ | $a$ + b |
| Forecast no | $c$ | $d$ | $c$ + d |
| Sum | $a + c$ | $b + d$ | $n$ |

important period in the longer-term trend of their forecasts' skill. This technique was later applied to outlooks with up to 48-h lead time (day 3 and day 2 outlooks) and updates to the initial day 1 outlook (Hitchens and Brooks 2014); improvement in accuracy and skill was noted with decreasing lead time, and between updates throughout day 1. Due in part to their focus on categorical outlooks, which include all of the severe thunderstorm hazards, these studies had little focus on identifying missed events and their effect on the assessment of forecast accuracy and skill; Hitchens et al. (2013) showed differences between including and not including missed events in calculating the frequency at which daily forecasts were skillful, but did not investigate this in great detail.

The present study builds upon previous analyses of the SPC's outlook products, while also serving as a precursor to future studies that will focus on the evaluation of their probabilistic outlooks. With respect to the latter, focusing on significant severe outlooks provides the opportunity to investigate approaches to best evaluate forecasts with a specified minimum coverage, which can later be applied to probabilistic forecasts of severe weather hazards.

## 2. Data

Coordinates representing vertices of probabilistic outlooks were retrieved from the SPC's website for 2005–15, and plotted on a latitude–longitude grid with nominal grid spacing of 80 km, approximating the SPC products' spatial definition.[1] The resulting dataset includes outlook areas for all three hazards for day 1 forecasts, which are issued at 0600 (06D1), 1300 (13D1), 1630 (16D1), and 2000 (20D1) UTC, and outlook areas for all three hazards combined ("any severe") for day 2 at 0600 (06D2) and 1730 (17D2) UTC, and day 3 at 1200 (12D3) UTC; significant severe outlook areas are issued for individual hazards on day 1 forecasts and all hazards on days 2 and 3.

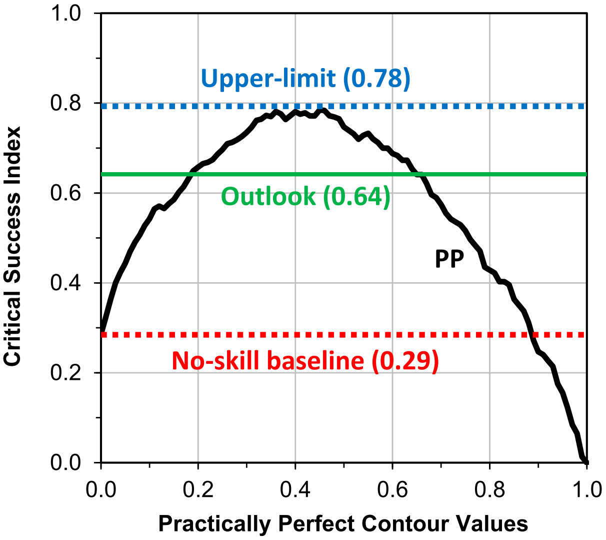

FIG. 1. Example of the CSI values (black curve) resulting from the 2 × 2 contingency tables constructed for each PP contour on a particular day. In calculating the relative skill for this example (0.71), the position of the outlook's CSI (OTLK; green line) is determined relative to the no-skill baseline CSI value (NS; red dashed line) and the practical upper-limit CSI value (UL; blue dashed line). The formula used to calculate the relative skill is (OTLK − NS)/(UL − NS).

Observed severe weather reports (OSRs) from 2005–15 were obtained from the SPC's Warning Coordination Meteorologist's web page (http://www.spc.noaa.gov/wcm#data) and aggregated over the valid period for days 2 and 3, and the 06D1 outlooks (24 h beginning at 1200 UTC of day 1); OSRs were also aggregated for the 13D1, 16D1, and 20D1 outlooks, beginning from their issuance through 1200 UTC the following day. Aggregated OSRs that met significant severe criteria were plotted onto grids with similar characteristics as the outlooks, with a grid box assigned a value of "1" if it contained at least one OSR.

## 3. Methods

To assess the performance of the SPC's significant severe outlooks, forecast and OSR grids were compared for each forecast–report set over the study period, resulting in 2 × 2 contingency tables representing each (Table 1). From these tables, standard measures such as probability of detection (POD), frequency of hits (FOH), and critical success index (CSI) were calculated [see Doswell et al. (1990) for a description of these measures]. For the assessment of forecast skill, the

[1] According to the SPC (http://www.spc.noaa.gov/misc/SPC_probotlk_info.html), "For all outlooks, the probability values represent the chance of severe weather occurring within 25 miles of any point, which is about the size of a major metropolitan area."

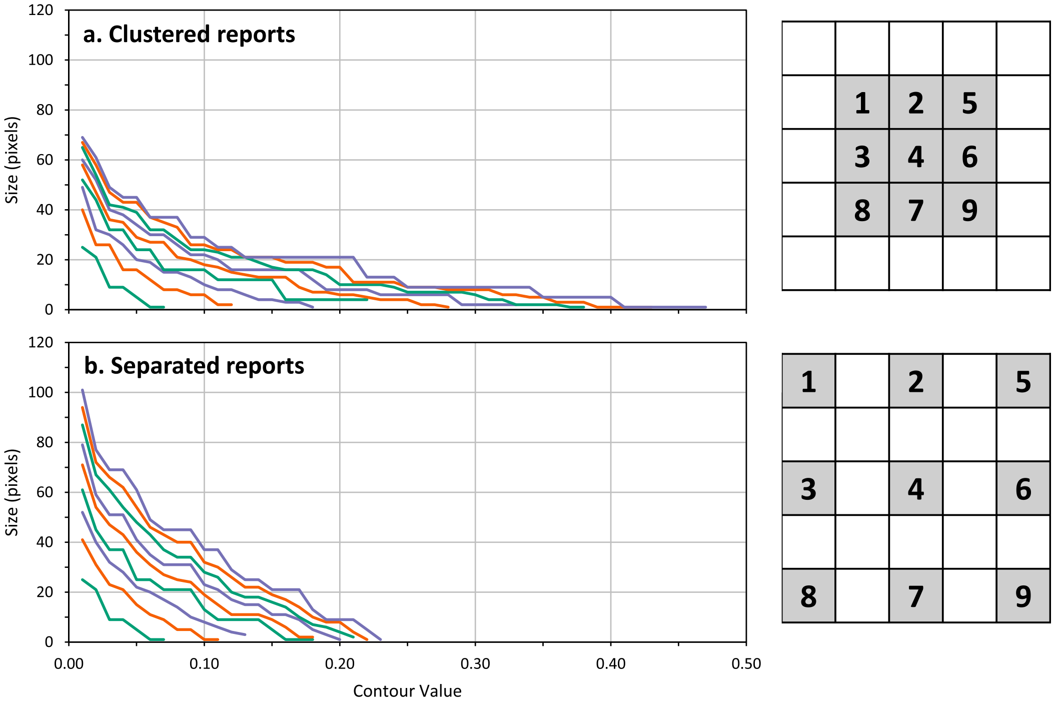

FIG. 2. PP contour size and value for increasing numbers of artificial report grid boxes, either (a) clustered together or (b) separated by one grid box. Reports are added to the central area of a 20 × 20 grid, with the placement of the first through ninth reports indicated to the right of each panel, where 1 represents the placement of the first report, 2 represents the placement of the second report, up to the ninth report. The line at the bottom left shows the results from using the PP technique on a single report, the next line to the right shows the results from two reports, and the rightmost line shows the results from nine reports; colors are used to distinguish between lines and have no significance.

practically perfect (PP) smoothing technique described in Hitchens et al. (2013) was used, whereby a 2D nonparametric Gaussian smoother was applied to each OSR grid, resulting in areas representing forecasts that would be made with perfect knowledge of locations of OSRs in advance, while still adhering to size and shape constraints characteristic of typical outlooks. PP forecasts consist of contours beginning with the lowest value (0.01), while not exceeding 1.00, although the maximum contour value depends on the number of grid boxes containing OSRs and their proximity to one another. By comparing the area created by a particular contour of a PP forecast with its associated OSR grid, contingency tables were constructed, and verification measures were calculated. Outlook skill was determined by comparing outlook CSI values to a no-skill baseline determined from corresponding PP forecasts, which in this case is the CSI achieved by linearly extrapolating the 0.00 PP contour from the 0.01 and 0.02 contours; a skillful forecast's CSI value exceeds the no-skill baseline value, meaning the forecaster demonstrated some skill beyond that of a person with no severe weather forecasting knowledge. Additionally, outlooks are bounded by an upper limit by identifying the PP contour that results in the maximum CSI value from all contours; the relative location of the outlook CSI value between the no-skill baseline and the upper-limit is referred to herein as the relative skill (Fig. 1). Conceptually, this upper limit represents an outlook that would be issued by a forecaster given perfect knowledge of the location of the reports, in a manner consistent with the guidelines for producing outlooks.

### *Missed events*

When evaluating forecasts of rare events, the definition of what constitutes a missed event, or a day that a forecast should have been issued but was not, is critical in analyzing these forecasts over time. When assessing forecast skill, a missed event is treated as a nonskillful forecast, especially when calculating the frequency that forecasts are skillful over time. The SPC does not explicitly define minimum criteria for the issuance of outlooks, stating that outlooks are intended to forecast organized convection. Therefore, special emphasis in this study is placed on determining a reasonable definition of minimum criteria for missed events.

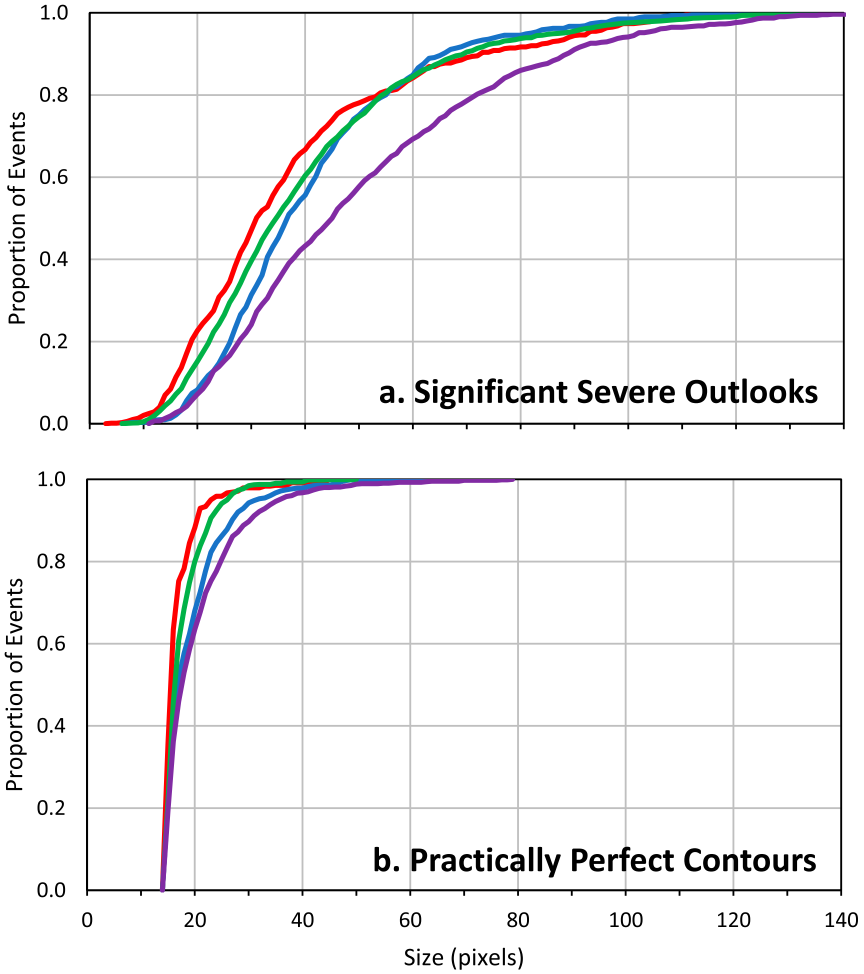

FIG. 3. Cumulative distribution function of the size of (a) significant severe outlooks and (b) PP contours of significant severe reports representing the maximum CSI value. Representing day 1 outlooks, tornadoes are shown in red, wind gusts in blue, and hail in green, while all significant severe cases are shown in purple, representing day 2 and 3 outlooks.

Since the contours from the PP smoothing technique are sensitive to the number and spacing of the grid boxes with OSRs, a series of tests was conducted using a 20 × 20 grid, with grid boxes in the center of the domain ‘‘activated’’ one at a time to simulate different configurations of OSRs. When increasing numbers of reports are clustered together (Fig. 2a), the maximum probability contour increases and the area enclosed by the contours increases, with the highest contour value for a single report (0.07) being much lower than the highest contour value for the nine reports clustered together in a 3 × 3 pattern (0.47). The increase in the maximum contour value is much greater than for the increase in the size of the contour; the 0.01 contour (representing values of at least 0.01) for one report covers 25 grid boxes, while for nine reports it is 69. A slightly different configuration of reports (Fig. 2b), separating each by one grid box, results in a relatively smaller increase in the highest contour value (0.23 for nine), and a relatively larger increase in size (101 for the 0.01 contour using nine reports). The first case, where reports are most clustered, represents the maximum contour value that could be attained for each number of reports, and the minimum size of each contour value for each number of reports.

Based on these tests, it is reasonable to conclude that criteria for a significant severe weather event should include both a minimum PP contour value and the minimum size of that contour. Examination of significant severe outlook sizes (Fig. 3a) shows that, generally, significant severe tornado outlooks tend to be relatively small. Day 2 and 3 areas for any significant severe outlook are larger, with 99% of hail and wind areas being at least 10 pixels in size, and 98% of tornado areas and all

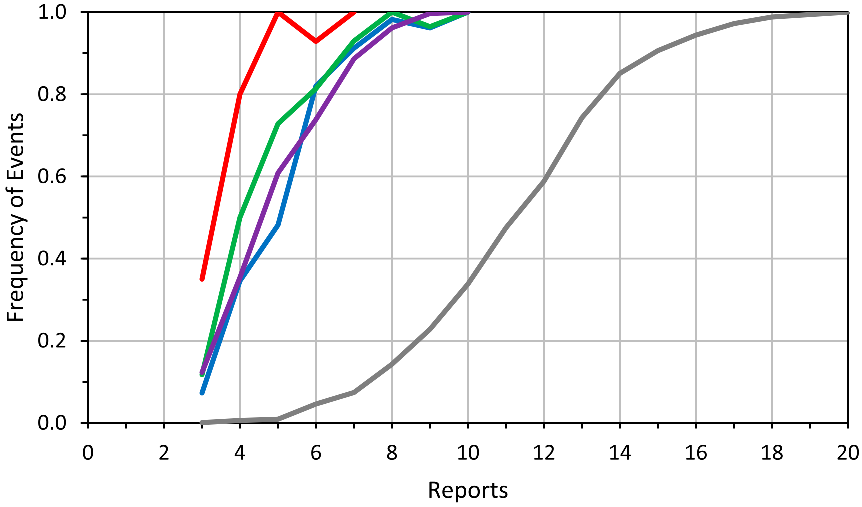

FIG. 4. The frequency of events, which are defined as a day in which the 0.10 PP contour is at least 10 pixels in size. The gray curve represents the results of 1000 random placements of reports on a 20 × 20 grid, beginning with 3 reports, and increasing to 20. The observed frequencies of significant severe tornadoes (red), wind gusts (blue), hail (green), and all significant severe cases (purple), using data from 2005–15, are also shown.

day 2 and day 3 areas having at least 10 pixels. Considering the distributions in Fig. 2, a corresponding reasonable minimum threshold for missed events is the existence of the 0.10 PP contour at least 10 pixels large. This effectively eliminates all instances where two or fewer grid boxes contain OSRs and requires tight clustering for three or four grid boxes. On days when these criteria were met, the PP contour used for the maximum CSI must meet additional criteria: it must be at least 15 pixels large to account for differences between the smallest outlooks for individual hazards and all severe cases, and it must have OSR coverage of at least 10% of its total area, following the definition set by the SPC. If no contour meets these criteria, the maximum CSI value used to calculate the relative skill remains the maximum value possible (1.00). As expected, most PP contours that qualify as the maximum CSI are not much larger than the minimum (Fig. 3b), with 35% of PP tornado areas and about 20% of all other significant severe PP areas at exactly the minimum size. It is recognized that these criteria are somewhat arbitrary, as there are no well-defined guidelines for patterns of OSRs that warrant forecasts, but these thresholds seem to correspond with what might be expected for SPC forecasters with perfect foreknowledge to use.

To further investigate the effects of the minimum criteria for a missed event, trials were conducted with $N$ grid boxes randomly assigned as reports using the same 20 × 20 grid. Here, $N$ was increased from 3 to 20 with 1000 trials at each value of $N$. Less than 10% of the time, $N < 8$ resulted in an event (Fig. 4), but by $N = 12$, event criteria were met over half the time. Not surprisingly, the rate at which OSRs qualify as events is much higher than the rates from these trials. This implies that clustering of significant severe reports tends to occur relatively often, with four grid boxes for tornadoes and hail necessary to exceed a frequency of 50% event identification, while all significant severe cases require five grid boxes, and wind six.

## 4. Results

The quality of significant severe outlooks was analyzed using a performance diagram (Roebber 2009), allowing for the simultaneous comparison of POD, FOH, CSI, and bias; these measures are calculated both for 1) only those days when an outlook was issued and 2) days when an outlook was issued or qualified as a significant severe event without an outlook. Figure 5a shows tornado outlooks have the highest POD values, ranging from 0.49 to 0.65 for all days, and 0.74–0.79 for days only with outlooks, while outlooks for all significant severe cases have slightly better FOH values than the individual hazards, 0.12–0.13, contributing to better CSI values, 0.05–0.09 and 0.11. Most evident is the large difference between the POD values for significant severe wind outlooks, 0.09–0.17 (filled) and 0.48–0.51 (hollow), which is caused by differences between the number of outlooks issued each year (annual average of 10 at 12D1) and the number of days each year when an outlook should have been issued (51). By including missed events, the

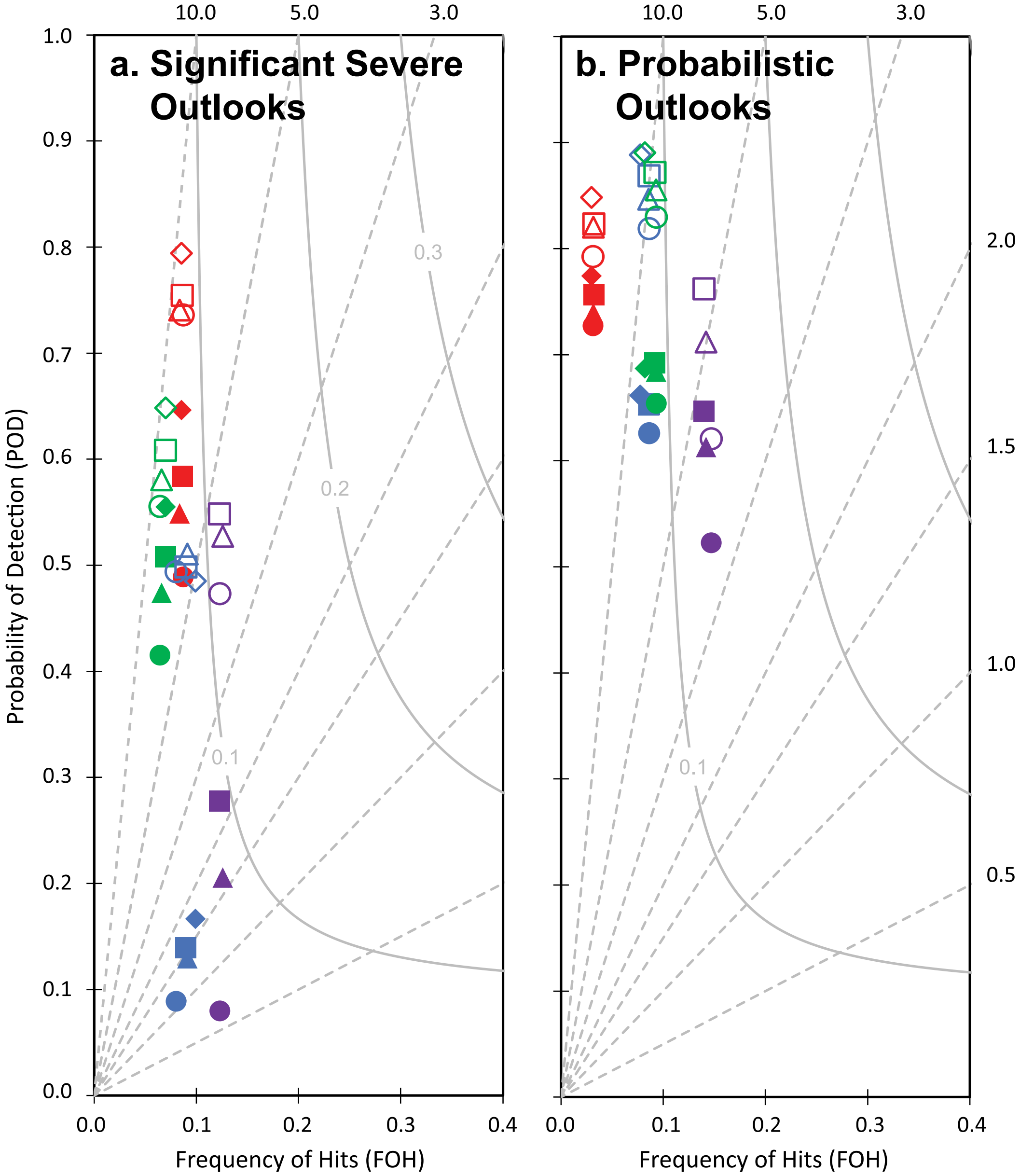

FIG. 5. Performance diagrams (Roebber 2009) showing (a) significant severe outlooks and (b) probabilistic outlooks. Tornadoes are shown in red, wind gusts in blue, hail in green, and all severe/significant severe cases in purple. Circles denote outlooks issued at 06D1/12D3, triangles for 13D1/06D2, squares for 16D1/17D2, and diamonds for 20D1. Filled shapes represent the performance on days for which outlooks were issued or a missed event was indicated by the PP values, while hollow shapes represent performance only on days for which outlooks were issued, excluding possible missed events.

POD and bias values both decrease for individual hazards and all significant severe cases, since the inclusion of missed events results in more grid boxes with OSRs, but the same number was detected (lowering the POD). Additionally, with no more grid boxes with outlooks added, the bias decreases. The magnitude of the difference in POD values between including and excluding missed events is directly related to the number of missed events.

In comparison, the performance of SPC forecasts for the lowest probability of each hazard (2% tornadoes; 5% wind, hail, and any severe cases) is better than the significant severe forecasts in terms of POD for all three hazards (Fig. 5b), especially wind outlooks, and for FOH

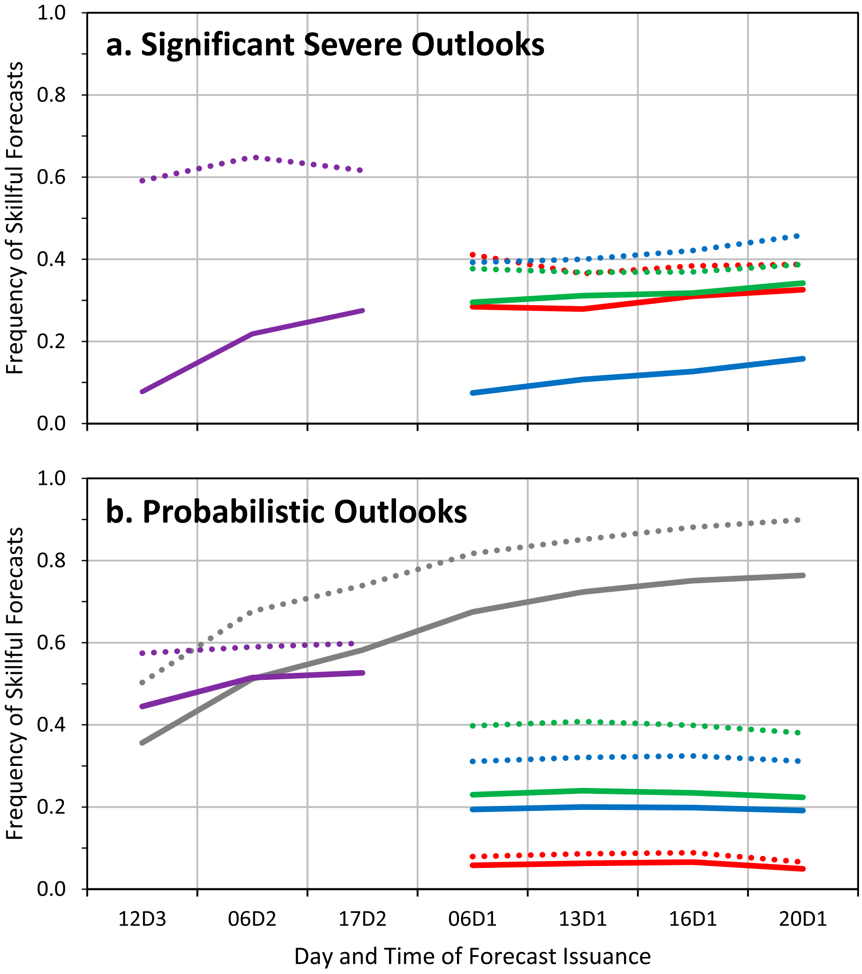

FIG. 6. The frequency at which (a) significant severe outlooks and (b) probabilistic outlooks are skillful. Tornadoes are shown in red, wind gusts in blue, hail in green, and all severe/significant severe cases in purple; for reference, categorical slight risk areas are shown in gray. Solid lines represent days on which outlooks were issued or a missed event was indicated by the PP values, while dashed lines represent only days for which outlooks were issued, excluding possible missed events.

values for all hazards except tornado forecasts, resulting in higher CSI values for forecasts of hail, wind, and any severe weather. Higher values of POD for low-probability forecasts are a reflection of more outlooks being issued and fewer missed events, as is evident by the relatively small magnitude of the change in POD when excluding missed events, as compared to significant severe forecasts; the improvements in FOH by low-probability forecasts indicate a lower false alarm ratio, likely because of the magnitude/size threshold for significant severe events. There is far less difference between the probabilistic performances based on all days compared to only days with outlooks, which further suggests the SPC is not forecasting significant severe outlooks as frequently as they should, with the possible exception of hail.

Differences in the rates of significant severe outlook issuance and events that require outlooks also affect how frequently SPC forecasts are skillful (Fig. 6), defined as the number of days with relative skill values greater than zero compared with the number of days with forecasts or missed events; the greatest difference, 0.51, is seen with forecasts for all significant severe cases on days 2 and 3. This is due partially to underforecasting the wind events, resulting in frequency differences of 0.29–0.32 (Fig. 6a). The SPC is more frequently skillful with significant severe outlooks for tornadoes compared with probabilistic tornado forecasts and, similarly, for wind and any severe weather when only considering outlook days.

## 5. Concluding remarks

Significant severe outlooks issued during 2005–15 were analyzed from day 3 through the 2000 UTC update during day 1 for days when outlooks were issued, and days when outlooks were issued or criteria were met for a missed event. Accuracy measures were better for each hazard, and all three combined, when only considering days with outlooks, especially for significant severe wind. A similar pattern emerged comparing the frequency of skillful forecasts, suggesting the SPC should consider adding a focus on identifying situations conducive to significant severe weather, especially wind events, to forecast these situations at rates similar to their occurrence.

To assess the rate at which significant events occurred, specifically to determine on which days events were missed, experiments were conducted to examine the PP's sensitivity to the number and location of OSRs. While there is no explicit minimum criterion from the SPC for issuing outlooks, a combination of PP contour value and size was chosen to identify events, with three or more reports necessary, and for OSRs to be tightly clustered with lower numbers. Using these criteria, days with low numbers of significant severe weather OSRs frequently qualified as events, occurring more frequently than would be expected at random.

*Acknowledgments.* The scientific results and conclusions, as well as any views or opinions expressed herein, are those of the authors and do not necessarily reflect the views of NOAA or the Department of Commerce. The constructive comments and suggestions made by Matthew Bunkers and an anonymous reviewer helped to improve the manuscript. Funding for this work for the first author was provided by Ball State University, and for the second author by the NOAA National Severe Storms Laboratory.

## REFERENCES

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, doi:10.1175/1520-0434(1990)005<0576:OSMOSI>2.0.CO;2.

Hales, J. E., Jr., 1988: Improving the watch/warning program through use of significant event data. Preprints, *15th Conf. on Severe Local Storms*, Baltimore, MD, Amer. Meteor. Soc., 165–168.

Hitchens, N. M., and H. E. Brooks, 2012: Evaluation of the Storm Prediction Center's day 1 convective outlooks. *Wea. Forecasting*, **27**, 1580–1585, doi:10.1175/WAF-D-12-00061.1.

——, and ——, 2014: Evaluation of the Storm Prediction Center's convective outlooks from day 3 through day 1. *Wea. Forecasting*, **29**, 1134–1142, doi:10.1175/WAF-D-13-00132.1.

——, ——, and M. P. Kay, 2013: Objective limits on forecasting skill of rare events. *Wea. Forecasting*, **28**, 525–534, doi:10.1175/WAF-D-12-00113.1.

Murphy, A. H., 1993: What is a good forecast? An essay on the nature of goodness in weather forecasting. *Wea. Forecasting*, **8**, 281–293, doi:10.1175/1520-0434(1993)008<0281:WIAGFA>2.0.CO;2.

Roebber, P. J., 2009: Visualizing multiple measures of forecast quality. *Wea. Forecasting*, **24**, 601–608, doi:10.1175/2008WAF2222159.1.