University of North Carolina Asheville
Journal of Undergraduate Research
Asheville, North Carolina
May 2019

Day One Storm Prediction Center Convective Outlook Verification Between 2011–2016

Guy Flynt
Atmospheric Sciences
The University of North Carolina at Asheville
One University Heights
Asheville, North Carolina 28804 USA

Faculty Advisor: Dr. Christopher Godfrey

Abstract

The NOAA/Storm Prediction Center (SPC) issues probabilistic convective outlooks for the contiguous 48 U.S. states. The Day 1 convective outlooks, issued five times per day and valid through the following morning, provide the probability of a tornado, severe hail (greater than or equal to 1 inch in diameter), or severe wind (greater than or equal to 50 knots) within 25 miles of a point. Comparisons of these convective outlooks with an observed tornado, severe hail, and severe wind reports during the period 2011–16 reveal the forecast accuracy of SPC forecasters. This effort produces accuracy measures for each hazard type both in aggregate and for each of the five daily forecast times. Scalar attributes of forecast quality are plotted on three different plots: performance diagrams, reliability diagrams, and line plots showing seasonal variability. Results indicate that the SPC may inflate the spatial extent of convective outlooks at the risk of creating false alarms in order to advance their mission to protect property and save lives.

1. Introduction

The Storm Prediction Center (SPC) has an interesting starting point in history, covered in Corfidi[1]. Corfidi[1] discussed the birth of the SPC, and the birth of the convective outlooks (CO), which has been one of the most useful products of the SPC over the past sixty years. The CO changed names before being officially established. It was originally called the "Severe Weather Bulletin" in 1952 by the U.S. Weather Bureau[1]. The bulletin was similar in form to the modern-day convective outlook. In January of 1953, the Severe Weather Unit experimented with publishing what were called "Severe Weather Discussions"[1]. After a month, the experimental discussions became an official product and were renamed "Convective Outlooks"[1].

The Storm Prediction Center has issued convective outlooks since 1955[1]. However, in March 1999, the SPC looked at other ways of producing the categorical CO. Categorical CO were just outlooks that contained a risk. The idea was to produce a forecast outlook that included probabilities of the three severe weather types; hail, wind, and tornadoes. The goal was to move away from the categorical convective outlooks. The definition of a probabilistic forecast is a direct communication of uncertainty of the forecast to the public. The categorical forecast was a word barrier, stopping straightforward information to the user. The probabilistic forecast would allow the user to look at the probabilities and make a decision[6].

The construction of probabilistic forecasts requires a climatological database of severe weather events. The database was from the National Weather Service that contained past severe weather storm reports. From there, the CO can be expressed with probabilities if one or more events happen within a radius of 25 miles or roughly 45 km[6]. This definition is used to help verify gridded convective outlooks. Hits were reported if an event was observed within the 25-mile radius of the outlook. The SPC experimented with probabilities by first grouping storm reports on an 80 by 80 grid, a single grid cell the size of the contiguous United States, and the same grid that this paper will use to verify the

convective outlooks. That is, they shaped the grid boxes to match the area of a circle with a 25-statute mile radius. The grid displacement was then smoothed in both time and space using non-parametric density estimation techniques[9].

For ease, the probabilities for the outlooks range from 0 to 100%. Due to the rarity of the events, and to translate the idea of probability forecasts instead of a categorical forecast to the forecasters, they were developed for the benefit for all forecasters. The 1200 UTC outlook was gridded for five years[6]. Storm reports from the 24-hour forecast period were placed on top of the outlooks for all risks. Observed relative frequencies (probabilities) were then computed for each report type and for each risk category, allowing a comparison between the probability of severe weather types and risk category.

The SPC has issued CO since 1955[1]. However, few researchers have taken efforts to verify SPC CO. Hitchens and Brooks[4] made these efforts and looked at the 1200 UTC CO produced by the SPC from 1973-2010. Their purpose was to examine the quality of the forecasted CO over time. The word "quality" is used to describe a good forecast[7]. Murphy[7] broke down what a good forecast is: how consistent a forecast is, the quality of the forecast, and the value of a forecast[7]. There were only three risks during the period of interest; slight, moderate, and high. The SPC later introduced the marginal and enhanced risk categories in October 2014. To verify the 1200 UTC CO, Hitchens and Brooks[4] used a contingency table and calculated scalar attributes to evaluate the SPC forecasts. Hitchens and Brooks[4] used the probability of detection (POD), frequency of hits (FOH), critical success index (CSI), and bias to assess the accuracy of the forecasts.

Hitchens and Brooks[4] used different spatial scales to evaluate different risks. These scales included 40, 80, 160, 320, 640, 1280, 2560, and 6100 km. Increasing the grid spacing led to the greatest improvement in the FOH values. Changing the grid spacing led to little change in the POD values. However, increasing the grid spacing caused the false alarm frequency to decrease. Using the 80 km spatial grid scale for both risks, POD values from 1973-1993 increased more than any other skill score. However, after 1993 the skill score did not change during the rest of the time period. From 1973 through 2010, FOH increases slightly. The increase in POD can be interpreted as an improvement in the placement of the forecast risk areas. However, the FOH was consistent during this time period, likely resulting from the increase in the size of the risk areas. After 1993, FOH values increase steadily until the end of the study period. Better precision of forecast areas reduced the rate of false alarms, making the FOH increase[4].

Hitchens and Brooks[4] also looked at moderate risk events with the same technique as that used for slight risk events. Their findings show that POD and FOH values were considerably lower than the slight risk POD and FOH values. Since moderate risks were sub-regions within slight risk areas and focused on high-severity events, the values were expected to be lower. This is not to be viewed as a comparison of the two risks, but rather as in indicator that there is a purpose for each risk category. Both the POD and the FOH improved over the study period for the slight risk category. POD increased until 1973, then FOH values improved as a direct result of making the forecast areas smaller and better placed.

Hitchens and Brooks[5] also evaluated SPC CO for day one though day three but used a different method. All time periods were included in the study, day one having five, day two having two, and day three having one forecast period. Hitchens and Brooks[5] used what is called the practically perfect forecast and used a contingency table to calculate scalar attributes. The practically perfect forecast is developed for each target period by using nonparametric density estimation with a two-dimensional Gaussian kernel to smooth the reported events. Hitchens and Brooks[5] discovered that forecast skill improved over the decades.

This paper will only consider and verify day one CO forecasts produced by the NOAA/Storm Prediction Center (SPC). As in Hitchens and Brooks[4], this paper will use 2×2 contingency tables to assess the accuracy of each CO for all forecast periods and types. Section two will discuss the verification process and scalar attributes used to assess the CO. Section three will display the results of the scalar attributes, and section four will discuss the results and describe any outcomes. The purpose of this paper is to verify whether the forecast skill changes between types, forecast updates, throughout the year, and review the skill of the SPC

2. Methodology

The SPC produces a daily CO with five different time periods. The first is issued at 0600 UTC, and valid until 1200 UTC the next day. The 0600 UTC forecast is then updated at 1200 UTC. Following the 1200 UTC update are 1300, 1630, 2000, and the 0100 UTC forecasts. These updates correspond with the daily routine of the public. The 1200 UTC update is at 8:00 AM EDT, followed 9:00 AM, 12:30 PM, 4:00 PM, and 9:00 PM EDT. Because the purpose of the SPC is to protect lives and property, these times correspond with the most active parts of the day.

This paper will only focus on the five forecast time periods associated with the day one forecast. A probabilistic convective outlook (PCO) can be broken down by category, probability, and weather type. For example, a moderate risk CO may include a PCO with a 15% chance of hail, 45% chance of wind, and a 15% change of tornadoes with 25 miles of a point. Different probabilities for each weather type will correspond with different categorical risks.

Table 1. Contingency Table.

	Observed Yes	**Observed No**	**Total**
Forecast Yes	Hit (a)	False Alarm (b)	a + b
Forecast No	Miss (c)	Null Event (d)	c + d
Total	a + c	b + d	a + b + c +d

To verify PCO, the CO was placed onto an 80 by 80 km grid spacing. The 80 by 80 km grid is the same grid spacing as that used by Hitchens and Brooks[4] and covers a domain that is roughly the size of the United States. Severe weather storm reports from the SPC storm report database were then overlaid on top of the grid to calculate hits or misses. All forecast grid cells within 25 miles of a storm report for each forecast type and corresponding report type (i.e., tornado, wind, or hail) are considered hits (i.e., a successful forecast), while all remaining grid cells within a CO probability contour are false alarms. Grid cells within 25 miles of any storm report located outside a CO probability contour are considered misses. All remaining convective outlook grid cells farther than 25 miles from their respective storm reports and outside the probability contours are considered null events. Once the reports were compared to the grid boxes of the CO, the elements were collected and added up and stored to calculate scalar attributes. A hit can be considered as any observed report within 40.2336 km of a grid point. The CO and data would be considered dichotomous since it is layered over a grid-cell by grid cell basis. Thus, a 2×2 contingency table can be used to evaluate the PCO[4]. Table 1 shows how the data would be organized to calculate skill score. Here, if a forecaster were to forecast a tornado and the tornado occurred, it was a hit (a). If the forecast for a tornado proved incorrect, meaning what was forecasted did not occur, then it was a false alarm (b). If the forecaster saw conditions unfavorable for tornadoes but a tornado did occur, then it was a miss (c). If the forecaster did not forecast a tornado and one was not observed, then it was a null event (d). The variables from this table then can be used to calculate scalar attributes. Scalar attributes are used to assess the accuracy of the forecast, not the skill.

For this paper, the probability of detection (POD), frequency of hits (FOH), false alarm ratio (FAR), and the critical success index (CSI) will be used to highlight the accuracy of the PCO. Equations 1-4 are the mathematical equations for the scalar attributes.

$$POD = \frac{a}{a+c} \tag{1}$$

$$FOH = \frac{a}{a+b} \tag{2}$$

$$CSI = \frac{a}{a+b+c} \tag{3}$$

$$FAR = \frac{b}{a+b} \tag{4}$$

A perfect POD, FOH, and CSI score would be one, with zero being the worst. A perfect FAR score would be zero, with one being the worst. Having a FAR skill score of zero indicates that the forecast landed all hits and no false alarms. A large-scale PCO could lead to a great number of false alarms, with few hits, yet forecasters prefer to eliminate false alarms to increase other scalar attributes. To achieve good scores for POD, CSI, and FOH, the number of hits would need to be greater than missed forecasts and false alarms. For a complete description and for more scalar attributes, see Doswell et al.[2,4].

To find seasonal variation, the data for each month of 2011–2016 in the contingency table were averaged and scalar attributes were calculated. The scalar attributes were then plotted on a line graph over a monthly period. A performance diagram allows a user to display all three scalar attributes, including bias, on one graph. To understand more about performance diagrams, refer to Roebber[8]. The performance diagrams in this paper looked a FOH versus POD, with FOH on the y-axis and POD on the x-axis. This allows the viewer to see the FOH, POD, and CSI scores for any period.

All forecast types and issued CO time periods were plotted on performance diagrams to observe any differences in skill throughout the study period. A reliability diagram helps to visualize the goodness of the forecast. This diagram shows forecast probability bins versus observed frequency. The diagrams show whether the forecasts under- or overforecasted for any probability bin. Each forecast type has a different probability bin range. For tornadoes, the forecast probabilities include 0, 2, 5, 10, 15, 30, 45, and 60%. Hail and wind probabilistic forecasts include 0, 5, 15, 30, 45, and 60%. Underforecasting results when the forecast did not forecast as many observations that were observed. On the diagram, this would be above the one to one line. Overforecasting refers to when a forecaster forecasts more events that were observed and is shown on a reliability diagram as a point under the one to one line. Good reliability is apparent when the forecast and observed frequencies are close to the one to one line. This paper looks at the five CO forecast time periods for each forecast type and year through the use of reliability diagrams.

3. Results

3.1 Performance Diagrams

3.1.1 hail performance diagrams

Perfect skill on a performance diagram tends toward the top right part of the diagram. The 0100 UTC forecast period shows the worst skill, with low POD and FOH, compared with the daytime forecasts. The other time periods had similar POD, FOH, and CSI scores. Between 2012–2014 the POD was low compared to POD scores in 2011, 2015, and 2016. In the years where the POD was low, the FOH and CSI increased slightly. Low POD scores derive from a larger number of misses compared to hits and result from either missed forecasts or smaller COs. The smaller CO would Between 2012–2014, the POD scores were low, but the FOH and CSI scores were higher compared to the FOH score in 2011, 2015, and 2016. The lower number of false alarms increased the FOH and CSI scores.

The yearly hail performance diagram shows a low POD resulting from a large number of misses. In the same years where the POD was low, FOH and CSI scores were slightly higher than in years that had high POD scores. Forecast periods after the 0100 UTC update showed an improvement in FOH, as in Figure 1.

The hail performance diagram shows that skill decreased in the 0100 UTC forecast period. The higher FOH and CSI scores correspond with smaller CO. The SPC could have issued large PCO for hail in 2011, 2015, and 2016, increasing POD, but decreasing FOH and CSI scores because of the number of false alarms. The larger CO would limit the misses but inflate false alarms. This is the opposite of the POD scores from 2012–2014. The SPC PCO could be smaller than the other years, decreasing false alarms but increasing the odds of missing an event. A larger PCO results in a higher POD but low FOH and CSI. A smaller PCO increases the FOH and CSI score but lowers POD. However, SPC shows great forecast skill for hits (POD), but this success comes at the cost of false alarms. The forecast has high quality, even with the inflated false alarm ratios.

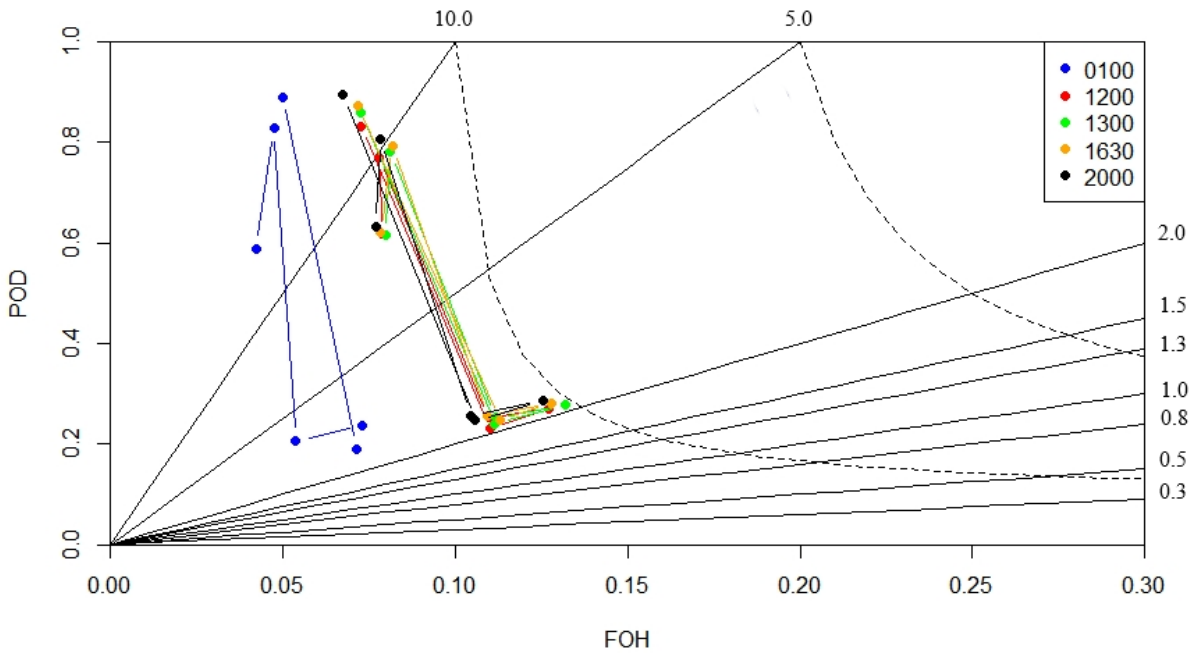

Figure 1. Hail performance diagram for all forecast periods. CSI values (dashed curves) start at 0.1 at lower left and move to 0.2 at upper right. Bias (straight lines) is given by the labels at the top and right of the plot.

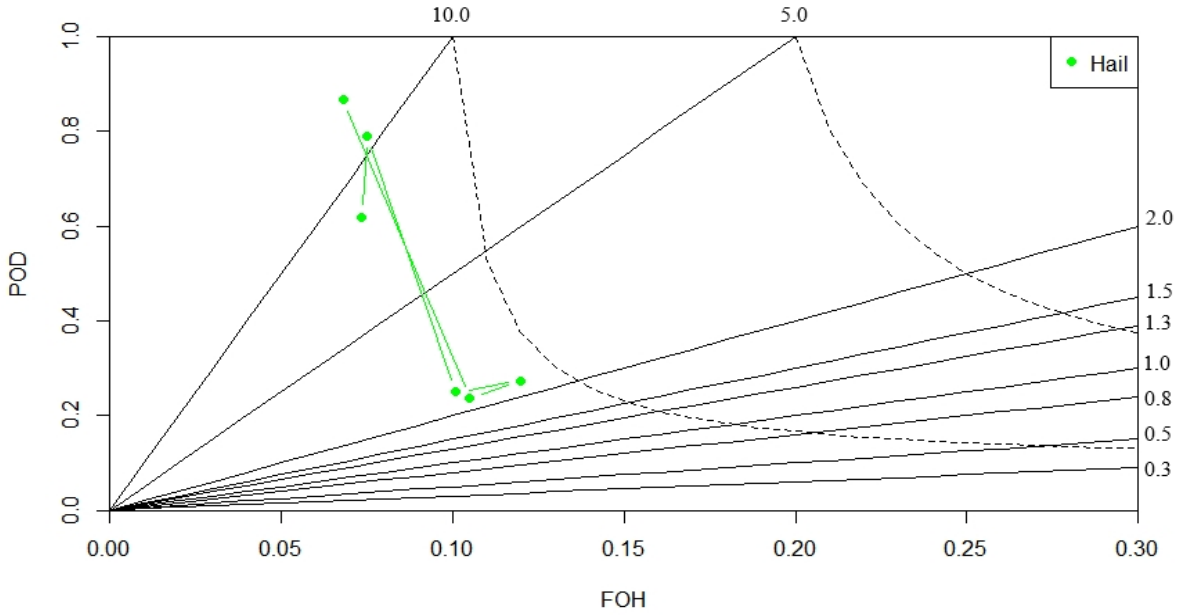

Figure 2. As in Figure 1, but for hail for years 2011–2016.

3.1.2 *tornado performance diagrams*

The forecast periods did not vary in skill as the data is tightly bunched in the upper left portion of the diagram. The 0100 forecast time period had the highest POD compared to other forecast periods, but a lower FOH and CSI score. Like Figure 2, the concentration of forecast periods in Figure 3 can be seen in Figure 4. In Figure 4 the POD is good, but with a low FOH and CSI score for the six-year period. It is not clear whether the skill increases during the period

from 2011–2016, as the data do not show any conclusive evidence for this trend. As in Figures 1 and 2, Figures 3 and 4 show that when there is a lower POD score, the FOH and CSI scores go up.

The tornado performance diagrams in Figures 3 and 4 show that the SPC issued PCO with a large area. The large PCO would capture all the hits for the day but includes many false alarms. This can be seen in Figure 3 and 4 with a low CSI and FOH score. The SPC was able to forecast well for tornadoes for all forecast periods. The SPC forecast well for tornadoes, even with low FOH and CSI values. The high hit rate for tornadoes makes the forecasts good. It means that the SPC were successful when forecasting for tornadoes. They were able to forecast for a large number of tornadoes that were within the CO.

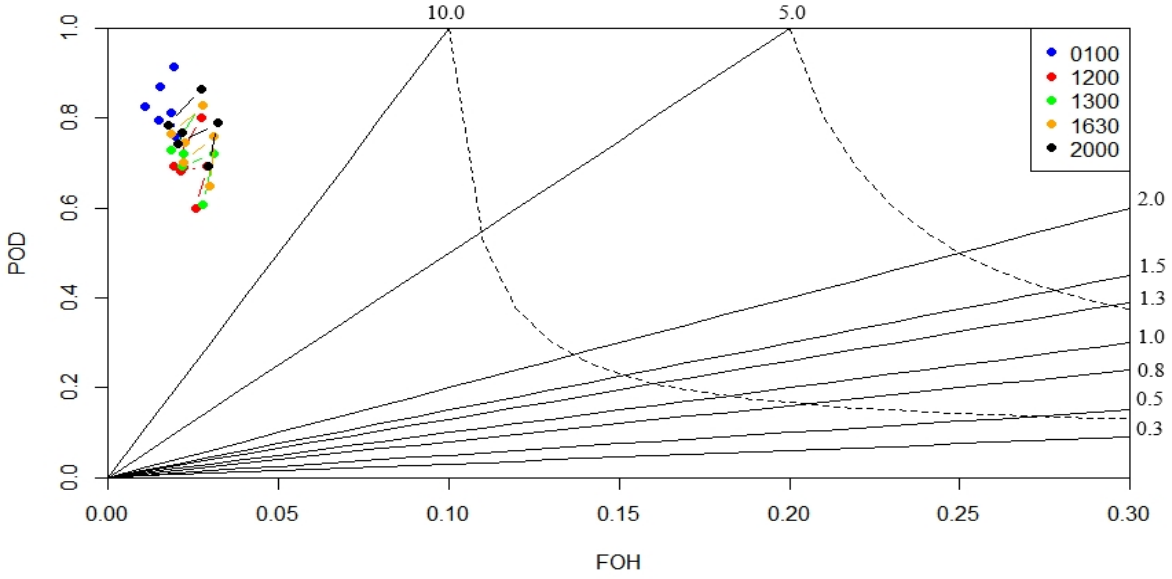

Figure 3. As in Figure 1, but for tornadoes.

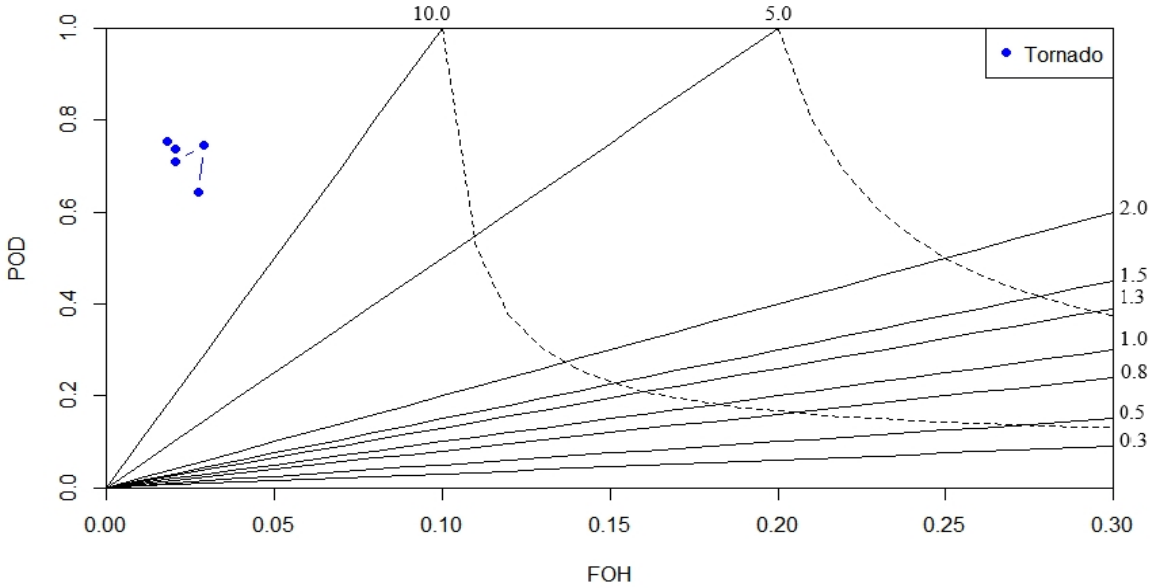

Figure 4. As in Figure 2, but for tornadoes.

3.1.3 *wind performance diagram*

In figure 5, there is a clear outlier. In 2011 the POD was low, but the FOH and CSI was the highest for all forecast time periods and in this study period. This year had the highest FOH and CSI score in the forecast period yet had the lowest POD. Figure 5 shows no difference in skill between 0100, 1200, 1300, 1630, and 2000 UTC, except for the 0100 UTC forecast period in 2011 and 2016. The lowest POD in the study was in 2016. Between 2012–2015, the skill for each forecast period was similar. Figure 6 shows similar results to Figure 5, but for all forecast periods. The highest FOH and CSI score happen in 2011, and 2016 had the lowest POD, FOH, and CSI score. The rest of the time periods had similar values.

For Figures 5 and 6, FOH and CSI were highest in 2011. For 2011, the SPC forecasted PCO that resulted in high FOH and CSI scores, but a low POD score. The low POD score could have resulted in bad forecast or busts. The other years, PCO observed hits, resulting in a higher POD but also increased the number of false alarms. The SPC forecast quality for wind is good, even with the two outlier years. Case studies of those two years would need to be conducted to understand why these two years were so different in score compared to the other years.

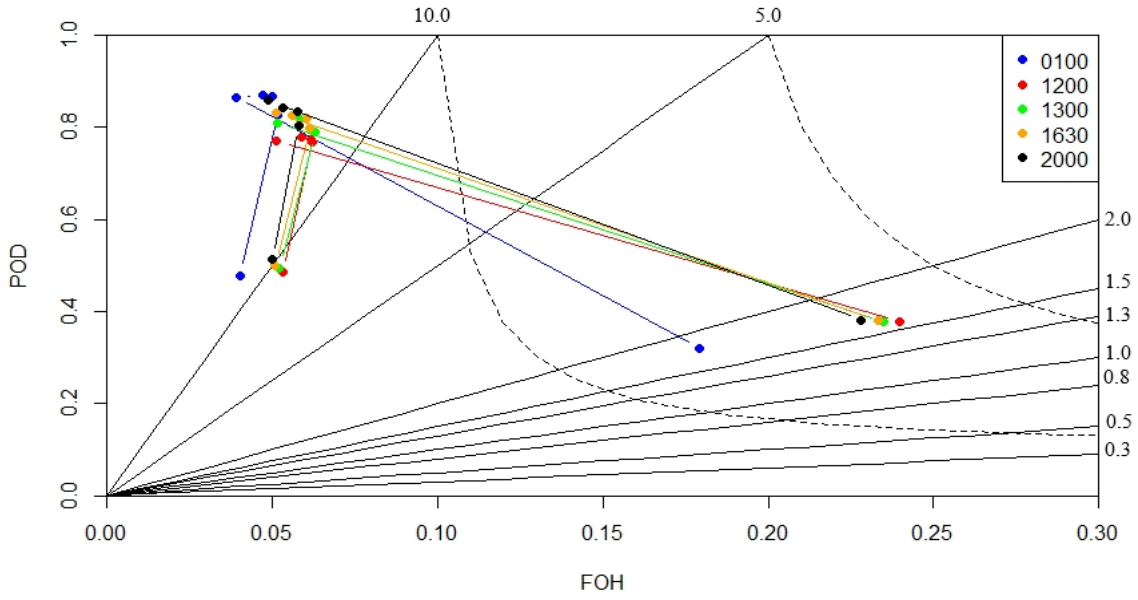

Figure 5. As in Figure 1, but for wind.

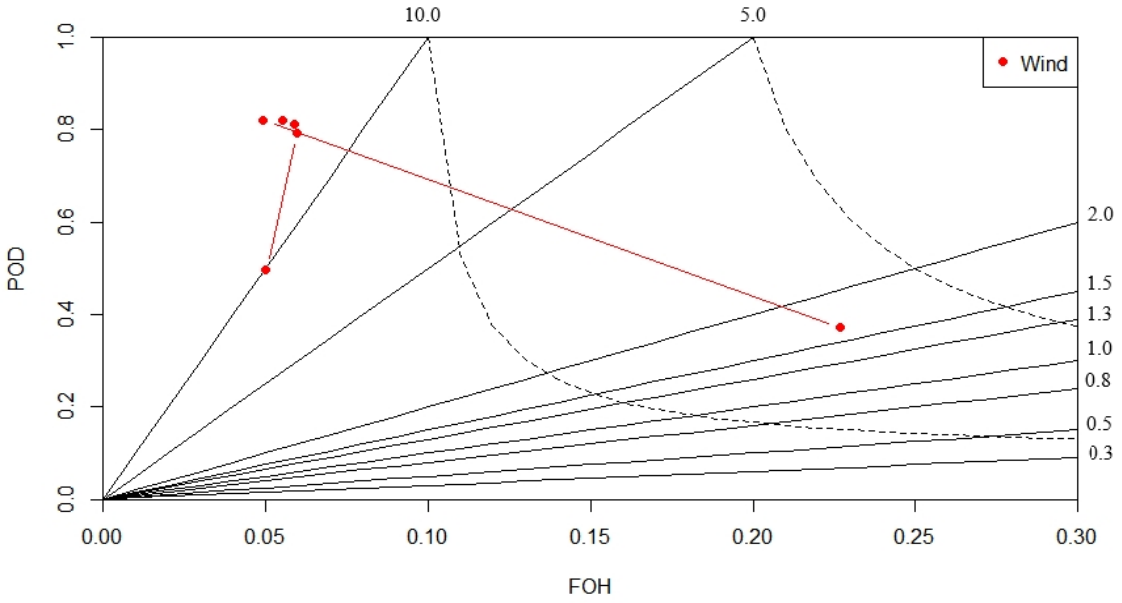

Figure 6. As in Figure 2, but for wind.

3.2 Reliability Diagrams

3.2.1 *hail reliability diagram*

Good reliability is when the forecast probability matches the observed frequency. Each time period was used to determine whether reliability changed throughout the study period and forecast period. The top left panel of Figure 7 shows that all years overforecasted for the 0100 UTC forecast period as probability increased. The 1200 and 1300 UTC forecast periods show good reliability, though the years 2013–2016 tended to overforecast as the probability increased. The reliability diagram for the 1630 UTC forecast period shows that 2011 possessed great reliability as the probability increased. From 2013–2016 the SPC overforecasted as probability increased. Hail for the 2000 UTC forecast period was overforecasted by the SPC in 2013 as probability increased, specifically for hail forecasts of 60%. A possible explanation for this is that the forecaster expected a huge outbreak, but the event did not happen. The lower probability bins, except for 0100 UTC, tend to show good reliability, possibly because these events are more common than a 60% hail probability event. The bust in the 2000 UTC period could be because 60% hail probability days are rare compared to 5% hail probability, or for that day the forecasted probability should have been a lower probability. Forecast periods of 1200, 1300, 1630, and 2000 UTC had better reliability than the 0100 UTC forecast period. Concluding the daytime forecasts were better and had good reliability compared to the nighttime forecast, 0100 UTC.

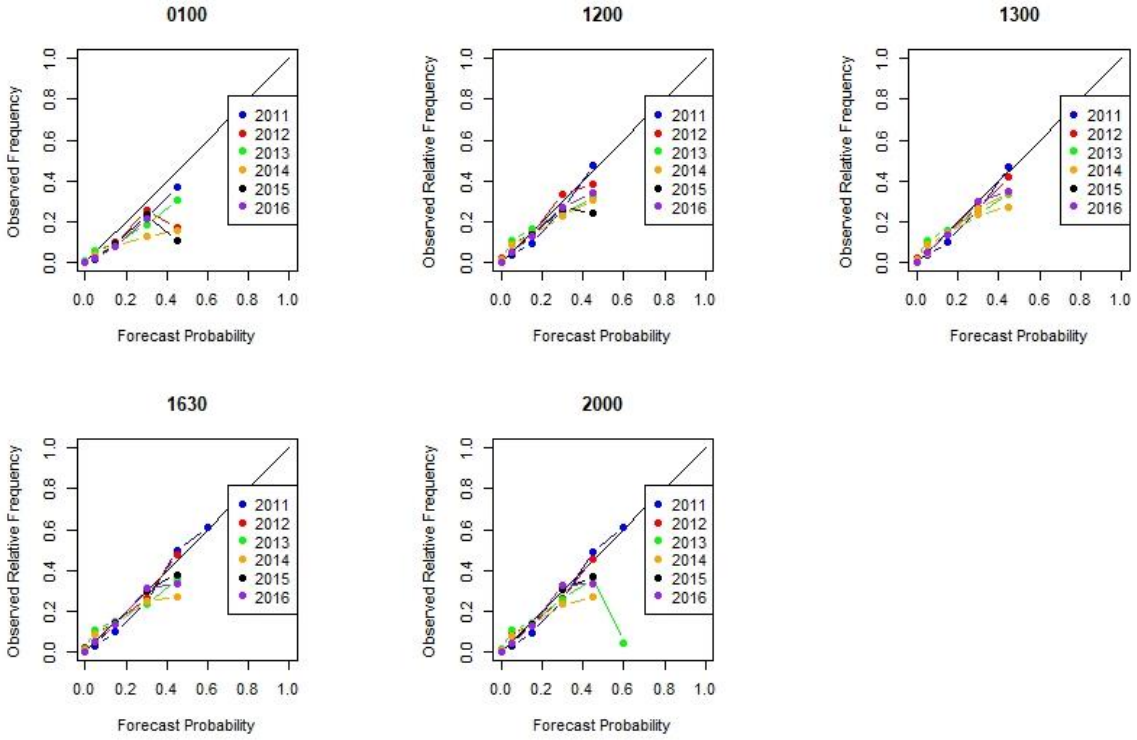

Figure 7. Hail reliability diagrams for all forecast periods. The forecast probability bins are 0, 0.05, 0.15, 0.45, and 0.6. Dots above the line indicate underforecasting and dots below the line indicate overforecasting.

3.2.2 *tornado reliability diagram*

Figure 8 shows that for all time periods and years, the overall reliability was good. For the 1200 UTC forecast time period, 2013 and 2016 tended to overforecast as the probability increased but was inconsistent in reliability as probability increased. The 2000 UTC time periods show an underforecasted bust in 2012 for 45% probability. Like hail, this could be due to the rarity of the event. The overall best reliability for all forecast periods was in 2011. The SPC was superb in forecasting for tornadoes in 2011, specifically with a massive tornado outbreaks in this year. To reference, the April 27-28 massive tornado outbreak occurred in 2011.

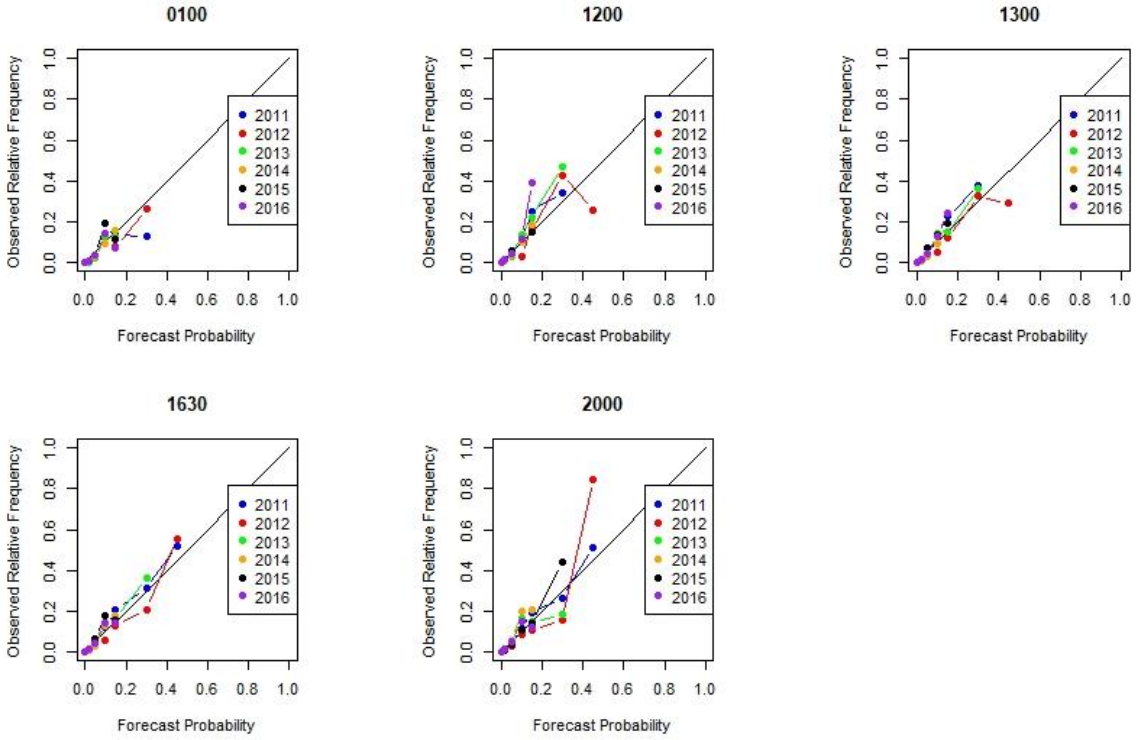

Figure 8. As in Figure 7, but for tornadoes.

3.2.3 *wind reliability diagrams*

Figure 9 shows that the 2011 wind forecasts exhibited bad reliability in the 0100 UTC forecast period, and forecasters underforecasted for the rest of the forecast periods. Forecasters generally overforecasted for the upper probabilities in 2012 and 2016. The years 2013–2016 had good reliability for all five forecast periods. The only bust was in 2016 in multiple forecast time periods and the 2000 UTC forecast period in 2013 for 60% wind probability days. Since 60% of wind probability days are infrequent, the reliability can go either way. The best reliability was in 2014 out of all forecast periods.

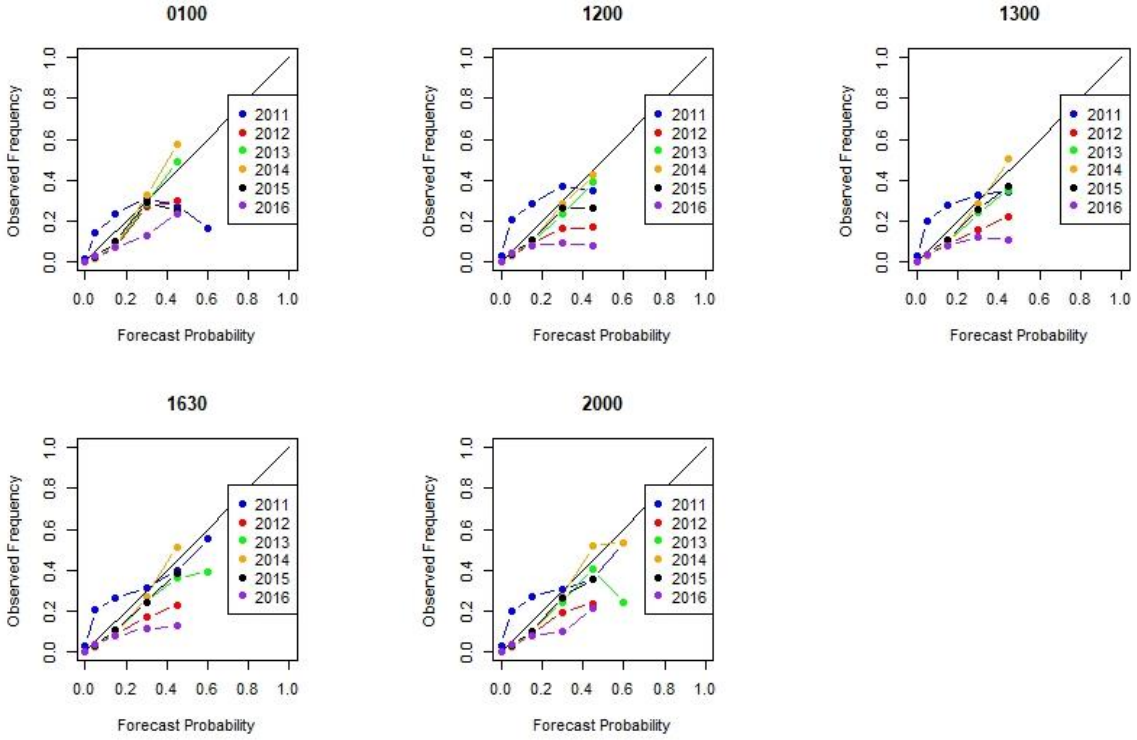

Figure 9. As in Figure 7, but for wind.

3.3 Seasonal Variability of Averaged Scalar attributes for Each Forecast Type

3.3.1 *seasonal variability of averaged scalar attributes for hail*

Seasonal variation of scalar attributes for hail tells an interesting story (Figure 10 and Table 2). The hits, false alarms, misses, and null events were averaged for each month, then used to calculate FAR, CSI, FOH, and POD. The maximum FAR skill score occurred in May. The FAR reached and stayed at is maximum throughout the months. In the same month, FOH and CSI hit their maxima. POD reached its maximum in June and November, with the minimum in September. When FAR reaches its maximum so does FOH and CSI. This could be due to how the scalar attributes are calculated. When the FAR is at its lowest, a good score, the total false alarms in that month is low. With a lower number of false alarms, the direct result would be that the FOH and CSI scalar attributes would increase.

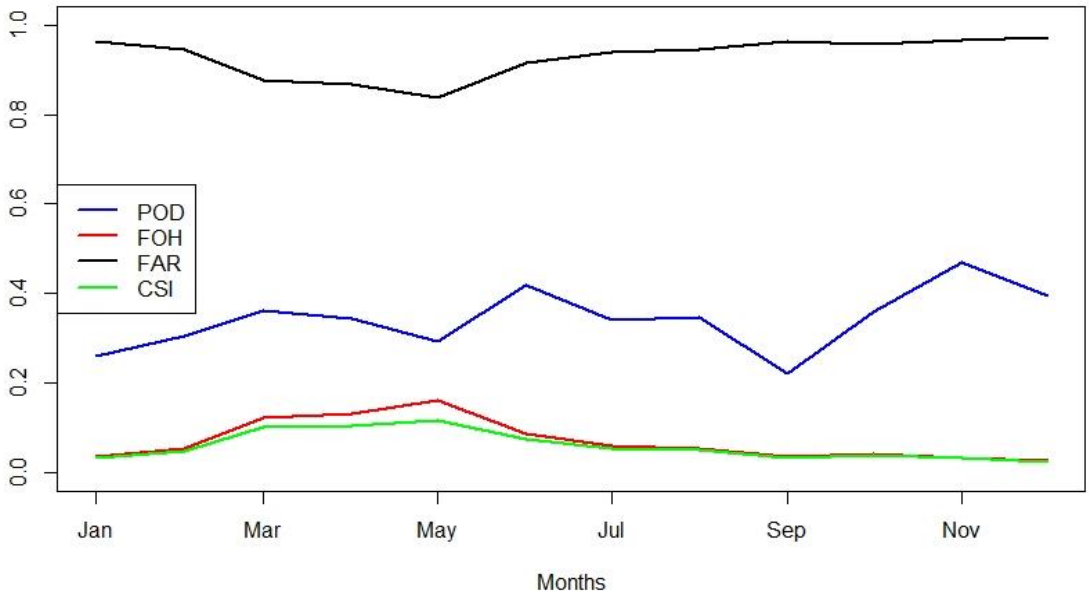

Figure 9. Monthly mean scalar attributes for hail.

Table 2. Contingency table for all hits, false alarms, misses, and null events for hail forecasts.

	Forecast Yes	**Forecast No**	**Total**
Forecast Yes	19558505	195718740	215277245
Forecast No	38294823	7848383755	7886678578
Total	57853328	8044102495	8101955823

3.3.2 *seasonal variability of averaged scalar attributes for tornadoes*

In Figure 11, the average monthly CSI line overlies the FOH line because the scores are identical. The hits, false alarms, misses, and null events were averaged for each month, then used to calculate FAR, CSI, FOH, and POD. The FAR line is constant and near one for all the months. The only skill score that varies seasonally is POD. The maximum for POD occurred in January and November, both cool months. The POD minimum occurred in August, an offseason month for severe weather. From Figure 11, the high FAR score and low FOH and CSI score indicates a higher amount of false alarms compared to hits and misses. The months with high POD values shows that the hits outweighed the misses. Refer to Table 3 for the sample size of tornadoes.

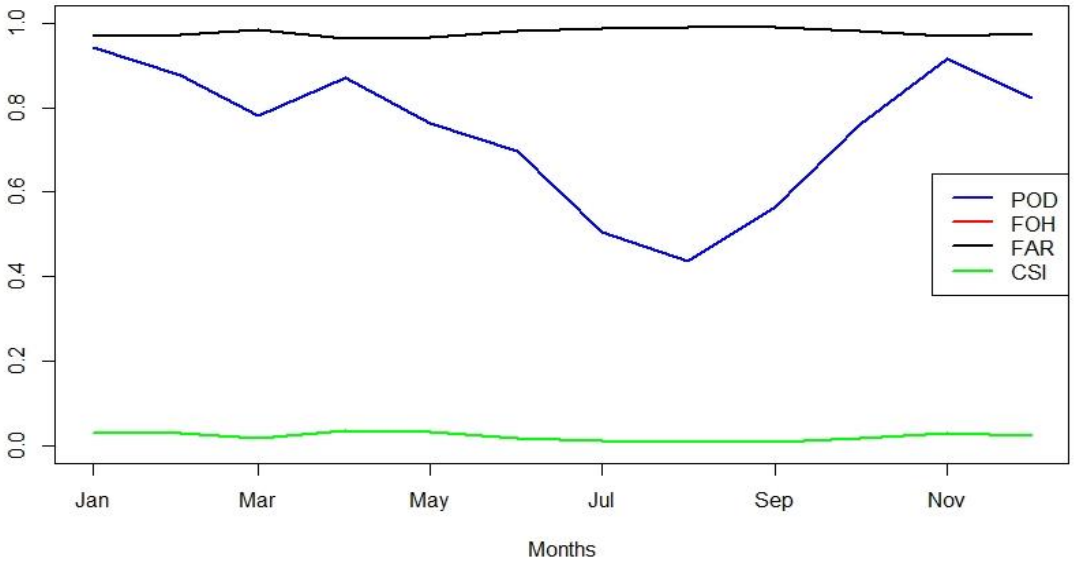

Figure 11. As in Figure 10, but for tornadoes.

Table 3. As in Table 2, but for tornado forecasts.

	Observed Yes	**Observed No**	**Total**
Forecast Yes	1845003	75726684	77571687
Forecast No	628299	8023755837	8024384136
Total	2473302	8099482521	8101955823

3.3.3 *seasonal variability of averaged scalar attributes for wind*

The average seasonal accuracy measures for wind forecasts show unique findings (Figure 12 and Table 4). The hits, false alarms, misses, and null events were averaged for each month, then used to calculate FAR, CSI, FOH, and POD. The POD maximum occurred in January and December, which are cool months with relatively little severe weather. The POD minimum occurred between March and September. The FAR, FOH, and CSI maxima occurred in June. When FAR hits its maximum, so does FOH and CSI. Therefore, the three scalar attributes correspond with each other. When the false alarm total is low, FOH and CSI scores increase.

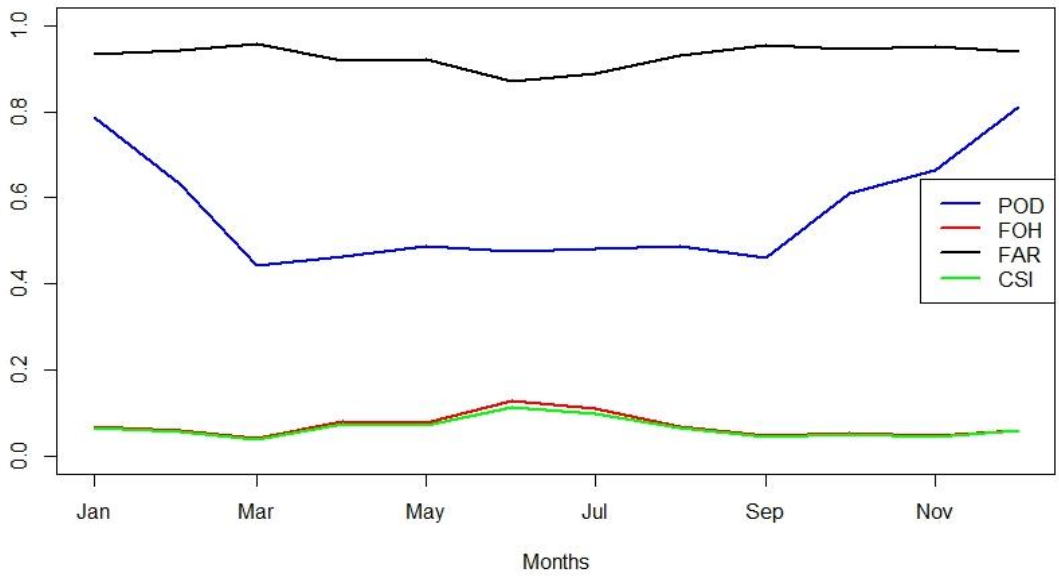

Figure 10. As in Figure 10, but for wind.

Table 4. As in Table 2, but for wind forecasts.

	Observed Yes	**Observed No**	**Total**
Forecast Yes	24341983	242676233	267018216
Forecast No	25739094	7809198513	7834937607
Total	50081077	8051874746	8101955823

3.4 Seasonal Variability of Averaged Scalar attributes of Significant Forecast Types

3.4.1 *seasonal variability of averaged scalar attributes for significant hail*

Significant weather forecast is also part of the CO (Table 5). The hits, false alarms, misses, and null events were averaged for each month, then used to calculate FAR, CSI, FOH, and POD. This is indicated in Figure 13 The POD, FOH, and CSI scores for significant hail, which is hail greater than two inches in diameter, are low compared with regular hail forecasts. However, there are still seasonal variations in the significant hail forecasts. POD reaches its maximum in April and November. The FAR maximum happens from April to May. When the FAR is at its lowest, FOH and CSI scores increase. FOH and CSI scores for significant hail are higher than for hail. This could be because the forecast area is smaller than the hail contour area, which reduces the possibility of false alarms.

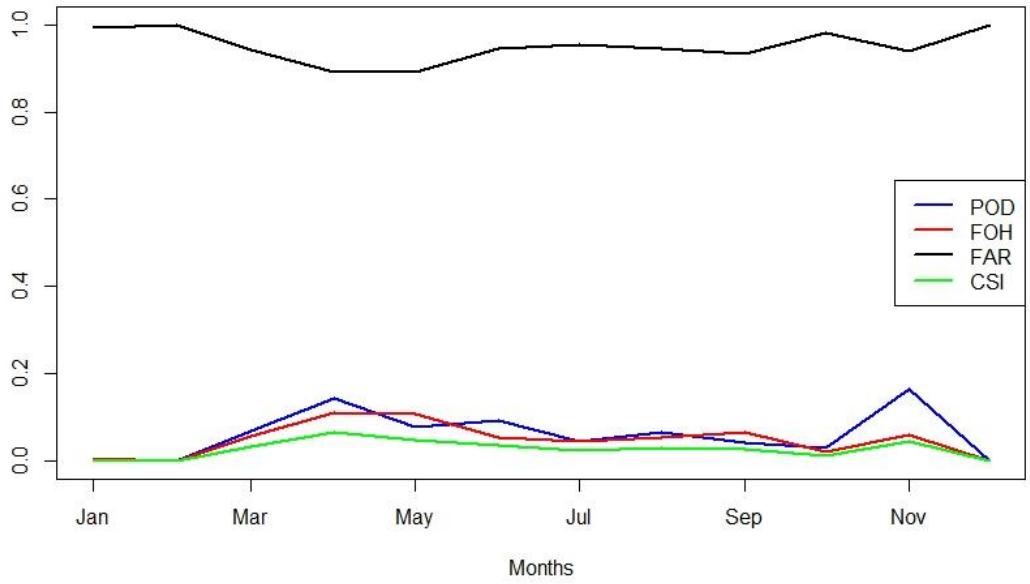

Figure 11. As in Figure 10, but for significant hail.

Table 5. As in Table 2, but for significant hail forecasts.

	Observed Yes	**Observed No**	**Total**
Forecast Yes	665036	7570566	8235602
Forecast No	7191053	8086529168	8093720221
Total	7856089	8094099734	8101955823

3.4.2 *seasonal variability of averaged scalar attributes for significant tornadoes*

The difficulty of forecasting for significant tornadoes can be seen in Figure 14 (Table 6). The hits, false alarms, misses, and null events were averaged for each month, then used to calculate FAR, CSI, FOH, and POD. In Figure 14, the maximum for POD is in April and from November to January, with a clear minimum in March. The period from August to September did not have significant tornadoes. The FAR score was high throughout the months. However, the FOH and CSI scores were better than the non-significant FOH and CSI scores overall. A maximum in FAR corresponds with a maximum in FOH and CSI.

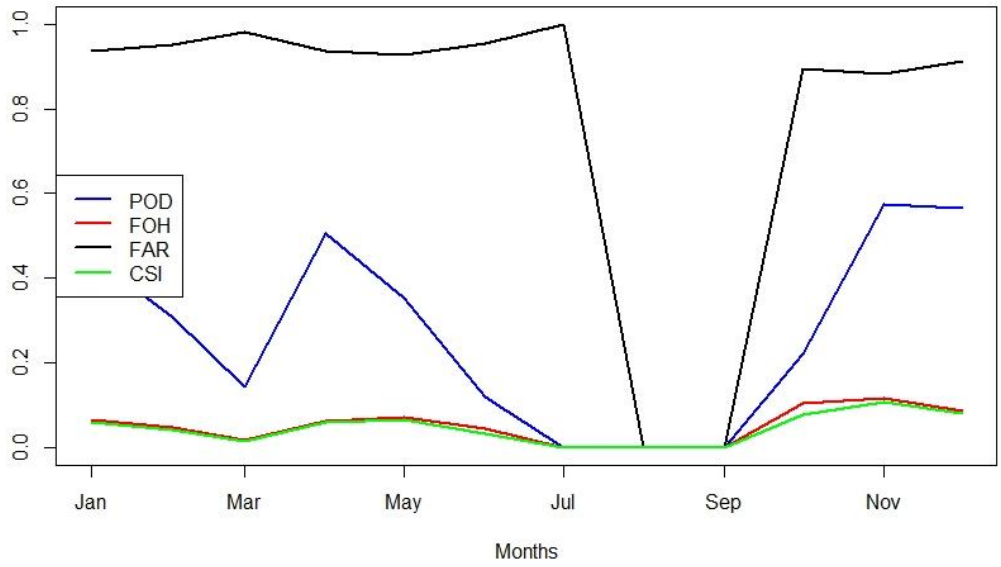

Figure 12. As in Figure 10, but for significant tornadoes.

Table 6. As in Table 2, but for significant tornadoes.

	Observed Yes	**Observed No**	**Total**
Forecast Yes	139454	2014855	2154309
Forecast No	263287	8099538227	8099801514
Total	402741	8101553082	8101955823

3.4.3 *seasonal variability of averaged scalar attributes for significant wind*

There are interesting aspects of the mean seasonal variation of scalar attributes for significant wind (Table 7). The hits, false alarms, misses, and null events were averaged for each month, then used to calculate FAR, CSI, FOH, and POD. A significant wind event is when the wind speed is higher than 65 knots. The POD maximum occurred in January. The FAR maximum occurred in August and January but varied throughout the months. When the FAR score was low, the FOH reached its maximum in August and January. The CSI reached its maximum in January. The FOH for most of the months was higher than the CSI score. In the months where the FOH score was high, the hits outweighed the false alarms. For CSI, the misses outweighed the hits, lowering the score.

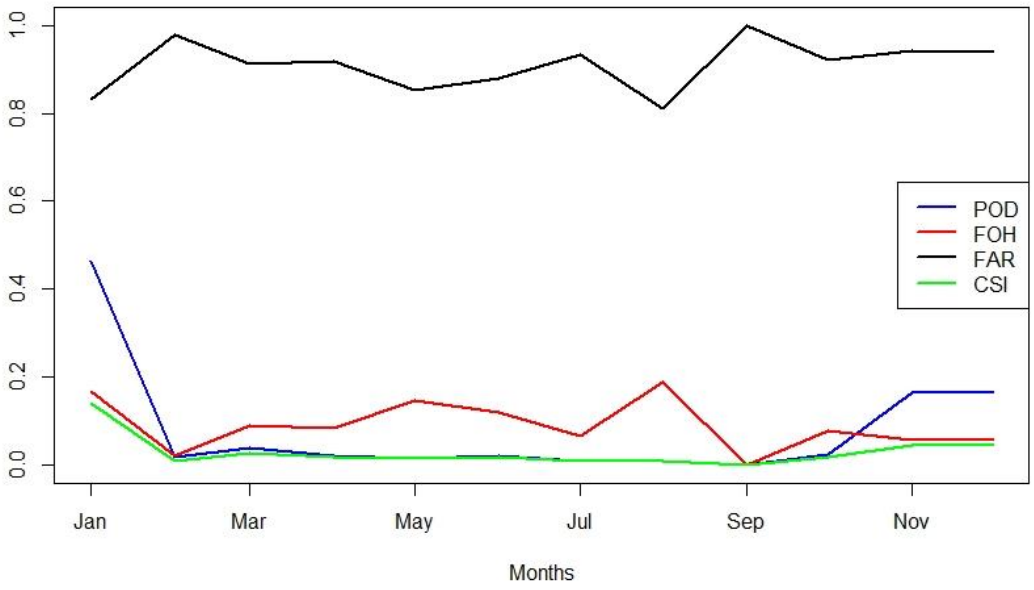

Figure 13. As in Figure 10, but for significant wind.

Table 7. As in Table 2, but for significant wind.

	Observed Yes	**Observed No**	**Total**
Forecast Yes	201140	1826879	2028019
Forecast No	10007861	8089919943	8099927804
Total	10209001	8091746822	8101955823

4. Conclusion and Discussion

The purpose of this paper is to show how CO forecast skill changes with respect to the types of probabilistic forecasts made by the SPC, as well as the timing of the forecasts and their seasonality. The data from the PCO are dichotomous, which allowed the use of a 2×2 contingency table to evaluate PCO. The table was used to calculate POD, FAR, CSI, and FAR for all forecast types and forecast periods. Performance, reliability, and line diagrams allowed the data to be displayed and show differences between forecast periods and type.

The performance diagrams showed interesting characteristics of the PCO size. For hail, the 0100 UTC time period showed the worst skill. The years 2011, 2015, and 2016 exhibited high FOH and CSI scores, but low POD. POD from 2012–2015 was high for all time periods but was low for FOH and CSI. The yearly scalar attributes show similar patterns for the forecast time periods. The tornado performance diagrams show tightly-packed POD and FOH scalar attributes. Both in aggregate and for individual forecast periods, POD was high but FOH and CSI were low. The wind performance diagrams differ from the hail and tornado diagrams. The year 2011 exhibited the highest FOH and CSI for the forecast time periods, but low POD. In contrast, 2016 had low POD and FOH.

The hail reliability diagrams (Figure 7) showed that as the forecast probability increased, except for in 2011 and 2012, where the SPC tended to overforecast. The reliability increased after each forecast period update. The biggest forecast bust was the 2000 UTC forecast period for the 60% forecast probability bin. This bin dramatically overforecasted severe weather events, but the higher probability events can vary in reliability due to the rarity of the event. The tornado reliability diagrams (Figure 8) showed good reliability for all forecast periods. The reliability increased after each forecast update, particularly after the 1200 UTC update. The biggest bust was in 2012 for the 2000 UTC time period due to the rarity of a 45% tornado probability day. The wind reliability diagram (Figure 9) shows the worst reliability in terms of different years. The trend for 2011 for all probability bins and forecasts periods was to underforecast for the lower forecast probabilities, then come close to the one to one ratio line once the

probability increased. Years 2012 and 2016 overforecasted for all time periods. The remaining years showed good reliability for all time periods. Overall, the tornado reliability diagrams for forecast periods show the best reliability out of the forecast types.

Assessments of seasonal variability through an examination of monthly mean scalar attributes for all forecast types have yet to appear in the verification literature. For hail (Figure 10) the POD was at its maximum in June and July, with the minimum for POD in September. The maxima for FAR, CSI, and FOH were in May. POD seemed to change from season to season, with the maximum being in the cool season, and the minimum being in the offseason of severe weather. The maximum of POD in November may be due to good forecasting or little severe weather. For tornadoes (Figure 11), FAR, CSI, and FOH scores do not change over the months. The seasonal variability in POD, however, the maximum is in November and January, with the minimum in August. Season variability for wind in Figure 12 shows that POD has its maximum in the cool season, and minimum in spring and summer. When the false alarm reached its peak maximum value, so did the FOH and CSI scores. Each forecast type had a higher score that was better than other forecast types. Seasonal variability for tornadoes (Figure 11) had the highest POD out of the forecast types but had low FOH and CSI. Hail (Figure 10) had the highest FOH and CSI of the forecast types.

Average seasonal variability was also calculated for all significant forecast types. For hail (Figure 13), CSI, POD, and FOH were low, with the maximum for POD in November. For significant tornadoes (Figure 14) the POD varies throughout the months. Significant wind (Figure 15) show unusual results for all scalar attributes compared to the other significant forecast types. The FOH was the highest for wind than for the other significant forecast types, with the maximum in August. POD was at its maximum in January and decreased to nearly zero for the rest of the months until October. Significant wind (Figure 15) had the highest FOH out of the forecast groups, but significant tornadoes had the highest POD.

The increase in FAR corresponds to an increase in CSI and FOH scores. This idea applies to Figures 1–6 and 10–15. A lower FAR score would mean fewer false alarms and result in increasing the FOH and CSI score. However, the complexity of this idea is that the SPC has to decide whether to issue a PCO that warns all of the population and records every hit, or issue a smaller and more precise PCO which tells the public where exactly the severe weather would be, but risk the possibility of missing severe weather outside of the PCO. This can be seen in the figures for forecast types.

In conclusion, the SPC forecasts exhibits good reliability for hail and tornadoes, but overforecasts for wind (Figure 9). SPC nailed the forecast for tornadoes and hail in 2011 with great reliability (Figure 8). The nighttime forecast period, 0100 UTC, had the weakest reliability and low skill. The daytime forecast had good skill and reliability. The SPC PCO seems to be large for all forecast types in order to issue a severe weather forecast for every event, just to cover their bases. But this leads to a large number of false alarms, shown by the high FAR in Figures 10–15. The result is a low FOH and CSI score. Throughout the seasons, tornadoes have the highest POD, but low FOH and CSI scores. Forecasting for tornadoes is a difficult task to achieve, but it is interesting that the POD of hail and wind is lower than tornadoes.

Sample sizes of each month for the forecast types would affect the scalar attributes, but any grid points that did not possess a hit, miss, or false alarm would go into the null event category. Future work would include why forecast for different years of forecast type and period were either good or bad, and whether the missed forecast were due from dominant mesoscale forcing, or whether the good forecast were from synoptic scale dominants that allowed an easier forecast. The knowledge of synoptic, mesoscale , and planetary conditions during the good and bad forecasted years would help understand more about SPC forecasts. Model forecasting or bad data assimilation is a theory of why SPC forecast either performed well or not in certain years. Break out years, or out of the norm from climatology, of the forecast types could explain why some years the SPC did well in forecasting or under performed when forecasting for severe weather.

The SPC forecasts large CO to maintain their goal of protecting life and property while sacrificing making sharp forecasts. With the difficulty of predicting severe weather, the SPC has done a great job forecasting severe weather for the contiguous 48 states.

5. Acknowledgments

The author would like to thank Dr. Christopher Godfrey for his time and help with completing this project. I would also like to thank Ty Higginbotham for his contribution to the project, and to the reviewer, for their revisions to make the paper better. Finally, I would like to thank the SPC for providing their data to allow this project to start.

6. References

1. Corfidi, S. F., 1999: The birth and early years of the Storm Prediction Center. *Wea. Forecasting*, **14**, 507–525, doi:*10.1175/1520-0434(1999)014<0507:TBAEYO>2.0.CO;2*.

2. 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.

3. Herman, G. R., E. R. Nielsen, R. S. Schumacher, 2018: Probabilistic verification of Storm Prediction Center convective outlooks. *Wea. Forecasting.* **33,** 161-184, doi: *10.1175/WAF-D-17-0104.1*.

4. 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.

5. Hitchens, N. M., and H. E. Brooks, 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.

6. Kay, K. P., H. E. Brooks, cited 1999: Verification of probabilistic severe storms forecast at the SPC. [Available online at https://www.spc.noaa.gov/publications/mkay/probver/]

7. 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.

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

9. Simonoff, J. S., 1996: *Smoothing Methods in Statistics*. Springer, New York, 338 pp