AN EVALUATION OF THE STORM PREDICTION CENTER DAY ONE PROBABILISTIC CONVECTIVE OUTLOOK USING DIAGNOSTIC PARAMETERS

A THESIS

SUBMITTED TO THE GRADUATE SCHOOL

IN PARTIAL FULFILLMENT OF THE REQUIREMENTS

FOR THE DEGREE

MASTER OF SCIENCE

BY

ALEX FLYNT

DR. NATHAN HITCHENS

BALL STATE UNIVERSITY

MUNCIE, INDIANA

JULY 2021

Table of Contents

Acknowledgments

I want to thank several people for their support and help throughout this process during unusual times. I would first like to thank my fiancé, Christina Comer, for all her help and hours editing parts of my thesis. Next, I would like to thank my mom and dad for their love and support the past six years I have been in college. I can also not forget to give praise to my grandparents for their encouragement. I would also like to commemorate Teresa Wilson and Mike Hradesky in the Geography Department for the fun conversations, that were not about weather. My thesis advisor, Dr. Hitchens, was fantastic throughout this journey. He was always able to help me progress the project each time we met and guide me to collect my thoughts. I was also like to recognize my two committee members Dr. Call and Dr. Berland for their time and commitment. Dr. Call was an enormous help revising and editing my thesis. Dr. Berland was an immense help with the thesis and aiding me mentally. A special thanks to Andy Dean at the Storm Prediction Center for his time and patience with my countless questions. Andy supplied the convection outlooks and the archived mesoanalysis data; without his help the thesis would have never come to life. I would like to thank Dr. Godfrey at UNCA. I would not be the in the position I am in if it were not for his guidance in undergraduate and his teaching about writing code. Lastly, I dedicate my thesis to my dog, Justice, who passed away in April 2021. He was loving, caring, and I miss him every day.

Abstract

The Storm Prediction Center issues daily probabilistic convective outlooks forecasting the intensity and probability of severe weather hazards (tornadoes, hail, and wind). There is limited literature on wind and hailstorm environments, and the probabilistic convective outlooks environmental characteristics. To address these gaps, this research paper verifies the Storm Prediction Centers Day One 1200 UTC probabilistic outlook from March 26, 2008, through the year 2019. Bias and the Heidke skill score were used to evaluate the forecasts that had at least 10 observed reports to create three categories: overforecast, good skill, and poor skill. The archived surface object analysis from the Storm Prediction Center was used to collect parameter data of the max value within an 11.1 by 11.1 km box centered on the first report of day for each hazard's category to nearest hour to create a climatological dataset.

Analyzing verification statistics in the nearly 12-year period found there was a change in forecast ideology when marginal and enhanced risk were introduced in late 2014. The number of skillful forecasts for both wind and hail nearly doubled from 2014–2019, and in the same period there was a drastic drop in forecasts with an overforecast bias for both hail and wind. The mixed-layer lifting condensation level was the only tornado parameter was skillful in differentiating the mean between tornado overforecast and good skill. Skillful tornado forecasts were associated with higher heights accompanied by low mixed-layer convective inhibition and high effective significant tornado parameter values. Skillful hail forecasts were characterized by drier environments (low precipitable water amounts), moderate convective available potential energy, strong deep-layer shear, and moderate significant severe parameter values. Skillful wind forecasts environments were portrayed by high derecho composite parameter values, high probability of a mesoscale complex system parameter values, and strong wind shear between 0-1 km

1. Introduction

Severe weather hazards (SWX; hail, wind, and tornadoes) are some of the leading causes of injuries, property damage, and fatalities in the United States. Consequently, the Storm Prediction Center (SPC) has routinely issued categorical convective outlooks (CO) since 1955 and probabilistic convective outlooks (PCOs) since 1999 to protect life and property (Corfidi 1999; Kay and Brooks 2000). The PCO forecasts the probability of a tornado, hail, or wind event for the continental United States. The ability to forecast the location of a severe weather event can decrease the likelihood of injuries or fatalities. These forecasts are meaningful news resources for the public detailing the threat level, location, and timing of a hazard. Therefore, it is essential to verify the PCO to observe forecast skill and bias for hail, tornadoes, and wind to improve forecasts. Environmental data can be used to describe convective conditions on days a PCO verifies as an overforecast, good skill forecast, or poor skill forecast with finding similarities or differences with a severe weather type. Evaluating forecasts based on environmental characteristics can be beneficial for forecasters by understanding what conditions forecasters typically struggle with and mitigate the impacts of severe weather hazards.

Previous studies have focused primarily on tornadic storm environments using archived proximity sounding or mesoanalysis data (Rasmussen and Blanchard 1998; Rasmussen 2003; Thompson et al. 2003; Thompson et al. 2007, Guyer and Dean 2010; Thompson et al. 2012; Dean and Schneider 2012; Sherburn and parker 2014; Hampshire et al. 2018; Coffer et al. 2019). These studies have also identified what storm environments give forecasters the most trouble predicting tornadoes, and what parameters are useful in distinguishing between tornadic and hail/wind events. Forecast verification has concentrated on evaluating the SPC's CO, PCO, tornado watches produced by the SPC, and what environmental parameters can be used to distinguish between the three hazards (Doswell et al. 1990; Rasmussen 1998; Rasmussen 2003; Thompson 2003; Dean and Schneider 2012; Hitchens and Brooks 2012; Hitchens and Brooks 2013; Hitchens and Brooks 2014; Sherburn et al. 2014; Anderson-Fey et al. 2016; Sherburn et al. 2016; Hitchens and Brooks 2017; Herman et al. 2018; Coffer et al. 2019).

However, there are two gaps in severe weather research. First, there is limited to no research specifically on storm characteristics of damaging hail or wind events. Second, there has been little research on the evaluation of the SPC's probabilistic convective outlook and the examination of their convective environmental characteristics. Therefore, this paper verifies the SPC's PCO and studies the environmental characteristics. The purpose of this study is to use environmental parameters to find similarities and differences of verified PCOs and mitigate forecast biases. Furthermore, this study strives to extend the knowledge on tornadic storm environments, and significantly add to the hail and wind literature by composing a climatological dataset that characterizes the PCO environment. Identifying similarities or differences of the storm environments on days in which the forecast is either skilled, unskilled, or biased can improve the location of PCO for all three hazards.

This study raises questions to fill in the gaps discussed. What environmental characteristics are present for forecasts when predicting hail, tornado, or wind events? Are there similarities or differences of the forecasts for each hazard? What environments do forecasters struggle with or perform well in?

To answer these questions and achieve the paper's goal and purpose, the SPC's Day One 1200 UTC PCO from March 26, 2008, through the year 2019 were verified. To evaluate the forecasts, local storm reports (LSR) were grouped into 24-hr blocks and plotted on an 80.5 km grid. Since PCO and reports are considered dichotomous on a grid-cell by grid-cell basis, they can be evaluated using a 2 $\times$ 2 contingency table (Doswell et al. 1990; Hitchens and Brooks 2012). The contingency table can be used to calculate a plethora of scalar attributes and skill scores to evaluate the skill of a forecast (Doswell et al. 1990; Wilks 2011). This study uses bias and the Heidke skill score (HSS) to create three categories: poor skill forecasts, overforecast, and skilled forecasts (null dataset). To create a climatology dataset, the SPC's archive surface object analysis (SFCOA) was used to collect parameter values for reports for each hazard under the three categories (Bothwell et al. 2002). To analyze the data, the one-hour max value for the first report of the day was plotted on box plots and 2D density plots to help differentiate between the environments of the three categories and hazards.

Chapter 2 reviews the literature of storm environments of severe weather hazards and the SPC's forecast skill. Following the literature review in Chapter 2, a discussion of the data and methods are found in Chapter 3. The results of the study are presented in Chapter 4. Following Chapter 4, the discussion and implication of the results are discussed in Chapter 5. Lastly, Chapter 5 discusses future recommendations and ideas of the study.

2. Literature Review

The following paragraphs review the forecast skill of the SPC and discuss what storm environments are likely to produce a tornado, wind, or hail event. Furthermore, this section reviews what storm environments forecasters struggle with, and which hazard the SPC has the most trouble forecasting for. It also evaluates the storm environment of various storm modes and regional differences.

2.1 Storm environments

Past studies have analyzed storm environments that produce significant tornadoes, hail, and wind events. The real challenge is identifying a likely environment, a range of parameters, or the parameter with the highest forecast skill that could produce a severe or significant weather hazard. Overall, a majority of research has focused on forecast challenges, environmental characteristics that distinguish between significant tornadoes and nontornadic supercells, or between quasi linear convective storms (QCLS) and discrete right moving supercells. However, the dependence on a single diagnostic, composite, or environmental parameters to forecast severe weather is unwise. Parameters used as forecast variables, either gained from model data or proximity soundings, decay with time (Doswell and Schultz 2006). Moreover, there is no parameter that becomes a “silver bullet” to completely forecast a severe weather event in the future. Nonetheless, the use of these variables helps to distinguish between storm modes and environments and aid forecasters.

2.1.1 Tornado

Rasmussen and Blanchard (1998) used soundings from 1992 to evaluate storm characteristics. The authors divided the storm into three categories: TOR is a tornadic supercell with reports of F2+, SUP is a supercell without a significant tornado, and ORD is a storm with no severe weather reports. The authors found LCL (HSS score of 0.348), VGP (HSS score of 0.295), and EHI (HSS score of 0.267) were the best parameters to distinguish between TOR and SUP environments and had the highest forecast skill (see Table 2.1 for description of parameters). The authors also found EHI (0.386), VGP (0.315), SRH3 (0.268) were the most effective to distinguish TOR+SUP and ORD environments. Furthermore, Rasmussen (2003) updated the parameter dataset and found EHI1, using SRH1, had the highest forecast skill overall (0.355), as SRH1 (0.349) and CAPE3 (0.336) also had high forecast skill and usefulness distinguishing between TOR and SUP environments. Moreover, SRH1 (0.356) and EHI1 (0.355) had high forecast skill and were useful to distinguish between TOR+SUP and ORD environments.

Table 2.1. List and description of environmental parameters used in the literature review.

Abbreviations	Description	Units
CAPE	Convective Available Potential Energy	J/kg
CAPE3	0–3 km CAPE	J/kg
DCAPE	Downdraft CAPE	J/kg
EHI1	0–1 km Energy helicity index	numeric
ESHR	Effective shear	m/s
ESRH	Effective storm relative helicity	m2/s2
LCL	Lifting condensation level	m
LR3	0–3 km lapse rate	C/km
LR75	700–500 hPa lapse rate	C/km
MLCAPE3	0–3 km 100-hPa mean mixed layer CAPE	J/kg
MLEHI1	0–1 km 100-hPa mean mixed layer EHI	numeric
MLLCL	100 hPa mean mixed layer LCL	m
MUCAPE	Most unstable parcel CAPE	J/kg
SCP	Supercell composite parameter	numeric
SHERB	Severe hazards in environments with reduced buoyancy parameter	numeric
SHERBE	Effective shear magnitude SHERB	numeric
SHERBS3	0–3 km shear magnitude SHERB	numeric
SHERBS6	0–6 km shear magnitude SHERB	numeric
SHERBW3	0–3 km wind magnitude SHERB	numeric
SHERBW6	0–6 km wind magnitude SHERB	numeric
SHR1	0–1 km wind shear magnitude	m/s
SHR2	0–2 km wind shear magnitude	m/s
SHR3	0–3 km wind shear magnitude	m/s
SHR6	0–6 km wind shear magnitude	m/s
SHR8	0–8 km wind shear magnitude	m/s
SHR10	0–10 km wind shear magnitude	m/s
SRH500	0–500 m storm relative helicity	m2/s2
SRH1	0–1 km storm relative helicity	m2/s2
SRH3	0–3 km storm relative helicity	m2/s2
SRW2	0–2 km storm relative winds	m/s
SRW3	0–3 km storm Relative winds	m/s
SRW8	0–8 km storm relative winds	m/s
SSP	Significant severe parameter	m3/s3
STP	Strong tornado parameter	numeric
STPeff	Significant tornado parameter effective-layer	numeric
STPfix	Significant tornado parameter fixed-layer	numeric
STP500	Significant tornado parameter using 0–500-km SRH	numeric
VGP	Vorticity generation parameter	numeric

VTP	Violent tornado parameter	numeric
WIND1	0–1 km wind speed magnitude	m/s
WIND2	0–2 km wind speed magnitude	m/s
WIND3	0–3 km wind speed magnitude	m/s
WIND4	0–4 km wind speed magnitude	m/s
WIND5	0–5 km wind speed magnitude	m/s
WIND6	0–6 km wind speed magnitude	m/s
WIND7	0–7 km wind speed magnitude	m/s

In a similar study, Thompson et al. (2003) assessed 413 soundings in close proximity to significant tornadic (F2+), weakly tornadic (F0/1), and nontornadic supercells. The MLLCL heights were statistically significant when discriminating between significant tornadic and nontornadic supercells, but not between significant and weakly tornadic supercells. Additionally, the authors found MLCAPE, SHR1, SRH1, SRH3, BRN shear, MLEHI1, SCP, and STPfix were useful in distinguishing between significant tornadic and nontornadic environments. However, SCP was the only parameter that could distinguish between significant and weak supercells. Also, there were no parameters with statistical significance that could distinguish between weakly tornadic and nontornadic supercells. Cravens and Brooks (2004) found STP, significant severe parameter (SSP), SHR1, SHR3, MLLCL, and LR3 were useful to distinguish between significant tornadoes and significant hail/wind events. In addition, shear values were higher for significant events, and MLLCL heights and LR3 were lower compared to significant wind/hail or severe events. The authors also found STP, SSP, SHR1, SHR6, MLLCL, and LR3 were able to distinguish between significant and severe weather events, similar to findings in Thompson et al. (2003). Additionally, Thompson et al. (2007) found ESRH and SRH1 were useful to distinguish between significant tornadic and weakly tornadic environments and between significant tornadic and nontornadic environments.

Hampshire et al. (2018) used proximity soundings to compare the environmental characteristics of tornadoes based on the EF-Scale. Reports from 1973–2015 was used to analyze significant tornadoes (EF2/3) and violent tornadoes (EF4+) and reports from 2013–2015 for weak tornadoes (EF0/1). The parameter, STPeff, was statistically significant in differentiating between violent and significant/weak tornadoes. Moreover, MLCAPE was the only component variable of STPeff to be statistically significant in distinguishing between significant and violent tornado environments. Furthermore, the authors analyzed low-level instability parameters and found MLCAPE3 and LR3 were both statistically significant in distinguishing between violent and significant tornado environments. The violent tornado parameter (VTP) was found not only to be useful in the distinction between violent and significant tornadoes but also for weak tornado events.

Coffer et al. (2019) looked at various environmental parameters to further the improvement in forecasting tornadoes. Tornadic and nontornadic events from 2005–2017 were analyzed as they pose the greatest forecast challenge. Tornadoes are produced by right-moving supercells and therefore were solely used in the study (Bunker et al. 2006; Smith et al. 2012). Using the true skill statistic (TSS), the authors find higher forecast skill when using lower layers of the SRH parameter. Additionally, ESRH and the STPeff, and STPfix also showed high skill when discriminating between significant tornadoes and significant hail/wind events. Overall, the SRH500 (0.529) and SRH1 (0.508) have the highest forecast skill. However, there are regional differences in forecast skill. Parameters such as SRH500 and ESHR had the highest forecast skill in the Northeast, South Atlantic, Lower Mississippi Valley, and Upper Mississippi Valley regions. For the Plains (southern and northern) it was STPfix and STP500, and in the Western U.S., MLCIN and MLLCL had the highest forecast skill. From the author's results, shear parameters have the highest skill in the Southeast and Northeast, while for the Plains it is STP500 and STPfix. Furthermore, versions of the STP parameter do show skill for all regions, specifically STPfix and STP500.

Bunkers et al. (2006) assessed environments of LLS, moderate long-lived (MLS), and short-lived supercells (SLS) events with the use of proximity soundings. Parameters for LLS such as SHR8, SRW8, MLBRN, SRW3, SRH3, MLLCL, SRH1, SHR1, SHR3, and SHR6 do well when discriminating between MLS and SLS. Therefore, deep-layer shear variables tend to do well discriminating between the longevity of supercells. A sample of tornado reports from 2003–2011 was analyzed and assigned a storm mode (Thompson et al. 2012). Tornadoes produced by discrete and cluster right moving (RM) supercells had higher MUCAPE values than tornadic linear storms. Furthermore, cluster and discrete storms also have higher SCP values compared to quasi linear convective storms (QLCS). When broken down by E(F) rating, parameter values for signs produced by RM discrete cells had higher MLCAPE, ESRH, and STPeff values compared to significant tornadoes produced by QLCS. Moreover, in trying to distinguish tornadic storms, shear values for nontornadic storms were significantly lower than tornadic storms (both discrete RM and QCLS). Also, 45 percent of QLCS tornado events were associated with MLCAPE values of less than 500 J/kg. The authors found the biggest difference in storm environments was in the winter between discrete RM and QCLS storms, where 75 percent of QCLS storms occurred with MLCAPE < 450 J/kg and in high shear environments. Finally, the values of the STPeff and their components tended to be higher for discrete RM cells throughout all seasons in comparison to QCLS storms.

Childs et al. (2018) looked at the environmental characteristics of different storm modes and (E)F1+ tornadoes during the cool months (November–February) between 2003–2015 in the Mississippi Valley and the Southeast. Discrete and cluster supercells were associated with higher STPfix, SBCAPE, EF-scale rating, and deaths compared to the QLCS storm mode. Furthermore, the QLCS storm mode is affiliated with lower SBCAPE and higher SRH1. Tornadoes rated (E)F1+ during the cool months can be characterized by lower MLCAPE, SBCAPE, and STPfix values but higher wind shear values when compared to annual values. Anderson-Frey et al. (2019) assessed tornado warnings and events in the Southeast between 2003–2017. Tornado events during winter are characterized by high shear and low CAPE. Most of the tornado events occurred in the spring, fall, and winter. Significant tornadoes were associated with MLCAPE values greater than 750 J/kg, SHR6 greater than 25 m/s, SRH1 greater than 250 m^2/s^2, and STPeff values greater than 1. STPeff values were particularly higher in the spring, where each of its components at the time are also high. When broken down by storm mode (RM and QLCS), MLCAPE values for RM discrete storms were doubled for tornado ratings (E)F1+ compared to QLCS tornado events. Moreover, QLCS tornado events typically occur in low MLCAPE values and high SRH1 and SHR6 values during spring and winter. Additionally, MLCAPE, SHR6, SRH1 values increase as tornado rating increases for both storm modes. In terms of skill, the authors found the lowest POD and highest FAR scores occurred in storm environments of low shear and low CAPE, and high shear low CAPE environments.

Schneider et al. (2006) found tornado environments in the Southeast and Mid-Atlantic are characterized by moderate to strong shear and low CAPE (MLCAPE <1000 J/kg). Whereas tornadic environments in the Plains are associated with high MLCAPE (MLCAPE ≥ 2000 J/kg) and moderate shear. Additionally, they found 97 percent of significant tornadoes occur when SHR6 is greater than 15 m/s; a high volume of tornado occurs when MLLCL values are below 1000 m and SHR1 values are greater than 10 m/s. Furthermore, Schneider and Dean (2008) found equivalent results as they focused on tornadic environments from 2003–2007. The authors found 84 percent of tornado reports occurred when SHR6 was greater than 15 m/s, and 65 percent occurred with shear greater than 20 m/s. Moreover, moderate deep-layer shear (SHR6 > 15 m/s) and low CAPE (MLCAPE < 1000 J/kg) environments are more common and likely to produce a potential tornado than high CAPE environments. Significant tornadoes are characterized by stronger deep layer shear, with more than 98 percent of significant tornadoes occurred with SHR6 greater than 15 m/s, and 86.7 percent occurred with shear greater than 20 m/s in addition, Dean and Schneider (2008) found forecast skill was at its lowest in environments of high shear and low CAPE.

Guyer and Dean (2010) investigated weak or low CAPE (MLCAPE ≤ 500 J/kg) tornadoes from 2003–2009. A bulk of weak CAPE tornadoes were lower-rated tornadoes, (E)F0-1. Diurnally, many weak CAPE tornadoes occurred between 1600–0300 UTC. Seasonally, most weak CAPE tornadoes occurred during the winter and early spring and fall months. With the higher count being in April, May, September, and March. Furthermore, weak CAPE and significant tornadoes occurred predominantly in the Southeast and Mid-Atlantic during the cool season months. Although, the Midwest also sees a moderate frequency of weak CAPE and significant tornadoes. In comparison to weak CAPE tornadoes versus tornadoes with MLCAPE > 500 J/kg, parameters such as SHR6, SRH1, SRH3, LR75, and LR3 were all able to distinguish between the two environments. Likewise, weak CAPE tornadoes tended to have higher MLLFC heights. For composite parameters, SCP and STPeff values were significantly lower for weak CAPE tornadoes. Thus, shear, lapse rates, and composite parameters were useful when discriminating between weak CAPE tornadoes and tornadoes with CAPE greater than 500 J/kg. Dean and Schneider (2012) analyzed tornadic environments from January 2003 through June 2012 using the SPC's SFCOA. The authors found the probability of a tornado increases with increasing MLCAPE and SHR6. Furthermore, tornadoes occurred between values MLCAPE from 0–2500 J/kg and SHR6 10–36 m/s. A majority of significant tornadoes occur frequently in environments of moderate-to-strong deep-layer shear, usually with values greater than 15 ms^{-1}. Although, there were regional differences in tornado environments. In the Southeast and Midwest are represented by low CAPE (MLCAPE < 1000 J/kg) and moderate deep-layer shear

(SHR6 $\geq$ 20 m/s) compared to the Plains where the bulk of tornado events occurred in moderate CAPE (MLCAPE $\geq$ 1000 J/kg) and moderate deep-layer shear (SHR6 $\geq$ 20 m/s). Furthermore, forecast skill was lower for HSLC environments in the Southeast. Therefore, the Southeast poses as the greatest forecast challenge for tornadoes in HSLC environments. Additionally, Anderson-Fey et al. (2016) used a 13-year climatology to evaluate forecast skill and found forecast skill to be lower in situations of low shear and low CAPE, and high shear and low CAPE.

In terms of challenges, high-shear, and low-CAPE (HSLC) environments prove to be difficult for forecasters. Sherburn and Parker (2014), looked at significant HSLC events from fall 2006 through spring 2011, with the requirement of SBCAPE $\leq$ 500 J/kg, 0–6-km shear $\geq$ 18 m/s, and MUCAPE $\leq$ 1000 J/kg. Additionally, the authors used the TSS to evaluate the skill of various environmental parameters in HSLC events. The most skillful parameters were LR3 and LR75 that could distinguish between HSLC significant events and null events. Wind and shear parameters in combination with lapse rates increased the forecast skill. Thus, the authors developed a series of parameters labeled as severe hazards in environments with reduced buoyancy parameter (SHERB) combined with a shear parameter. Overall, SHERBE (0.588), SHERBS3 (0.539), SHERBS6 (0.531), SHERBW6 (0.531), STPfix (0.501), EHI (0.501), SHERBW3 (0.481) had the highest skill when discriminating between HSLC significant tornadoes and null reports. Regionally, both SHERBS3 and SHERBE were the most skillful parameters when discriminating between HSLC significant severe reports and nulls. Sherburn et al. (2016) looked at EF1+ tornadoes and significant wind reports from 2006–2011 to analyze HSLC events. The authors followed the same criteria as in Sherburn and Parker (2014). The LR3 was the most skillful individual thermodynamic parameter and the ESHR had the highest skill overall for an individual parameter. The authors found the SHERBE and SHERBS3 (using the

3–5 km lapse rate) had the highest forecast skill. Furthermore, the authors developed two variations of the SHERB parameter called the MOSH and MOSHE. They proved to have even higher skill than SHERBE and SHERBS3.

2.1.2 Hail

Schneider and Dean (2008) found there is a higher dependence on strong deep-layer shear for severe hail reports, as a majority of severe and significant hail reports occurred with SHR6 greater than 15 m/s like results in Rasmussen (1998). Sherburn and Parker (2014) found the maximum forecast skill for HSLC significant hail events was SHERBE (0.376), SHERBS6 (0.273), and SHERBW6 (0.224). Johnson and Sugden (2014) evaluated the correspondence of hailstone diameter ranges and various parameters from 2003–2011. Large hailstone size (> 1.25 in) was associated with moderate-to-strong deep-layer shear values (ESHR and SHR6). Furthermore, SHREL and SHR10 were capable of distinguishing between hailstone diameter ranges and more specifically hailstone size > 2 in. The authors, using the HSS, found the LHP and SSP had the highest forecast skill in distinguishing between significant hailstone size (≥ 2 in) and hailstone greater than 3.5 in. Additionally, LR75 and deep-layer shear, such as the SHR6 and SHR10 parameters, had high forecast skill when distinguishing between hailstone sizes.

2.1.3 Wind

Evens and Doswell (2001) analyzed storm environments that produced derechos using proximity soundings. The authors found strong forced (SF) events had higher SHR2 and SHR6 values when compared to weak forced (WF) and hybrid events. Furthermore, SF events had higher SHR6 and BRN shear values compared to WF and hybrid events. Interestingly, WF and hybrid events had higher MLCAPE and MUCAPE on average compared to SF events. The authors found SHR6, SHR2, and SRW2 were useful in distinguishing between derecho and nonderecho events. Kuchera and Parker (2006) examined wind reports from 2003 to understand environments that support wind damage from convective storms. They found WIND1 (0.224), WIND2 (0.245), WIND3 (0.239), WIND4 (0.242), WIND5 (0.236), WIND6 (0.23), WIND7 (0.23) were all statistically significant and had high Optical Pierce skill score (OPSS) scores when distinguishing between damaging and nondamaging wind events. Wind magnitudes were more useful than shear values in distinguishing between wind and non-wind events. In addition, thermodynamic parameter such as DCAPE and MUCAPE also had use in characterizing between wind and non-wind events. Moreover, Sherburn and Parker (2014) found SHERBW3 (0.247), SHERBS3 (0.231), and SHERBW6 (0.224) had maximum forecast skill for HSLC significant wind events. Thus, low-level instability was useful in analyzing HSLC significant wind events.

2.2 Forecast verification

Forecast verification for tornado warnings from the NWS, tornado watches or convective outlooks from the SPC, and excessive rainfall outlook from the Weather Prediction Center (WPC) are necessary to evaluate the skill of the forecast. Precisely, if these forecasts are not verified, then neither the forecasters nor the public are taking them seriously (Doswell et al. 1990). Forecasts can be verified by various methods. For example, a 2 × 2 contingency table is used for a dichotomous forecast. Furthermore, a plethora of skill scores and scalar attributes can be formulated from this table. However, using a single skill score or scalar attribute to evaluate the skill of a forecast or environmental parameter can be unjust. Doswell et al. (1990) goes into detail about the use of several skill scores and scalar attributes but focuses on the HSS and the true skill statistic. Both skill scores are calculated using the contingency table. The authors find the true skill statistic has disadvantages in situations involving forecasts of rare events such as severe weather, but the HSS avoids these disadvantages. Thus, Doswell et al. (1990) implies the true skill statistic is an improper skill score for forecasting rare events. This is due to rare events being dominated by correct null forecasts, which then influences the skill score. However, the HSS is a reliable judgment of skill as it avoids the issues involved with the TSS, and a robust way to evaluate forecasts.

It can be difficult to evaluate, or even verify, what is a good or bad forecast. The term good or bad forecast is coined throughout the meteorological literature and discipline, but there can be a lack of clarity or understanding between the forecast and user. In Murphy (1993) journal, the author describes how to evaluate the goodness of a forecast. Specifically, about the differences between a measure of accuracy (bias) vs a measure of skill (HSS) under second type of goodness quality. Murphy (1993) describes accuracy as an average correspondence between individual pairs of forecasts and observations. The difference between the forecast and the observation is error. Higher accuracy is a result with fewer errors. The author describes skill as the accuracy of forecasts of interest relative to accuracy of forecast produced by standard of reference. Forecasts could be more accurate and have more skill due to various forecast types being easier like forecasting for rain compared to damaging wind events. Both terms have been frequently used to describe the quality of a forecast.

Hitchens and Brooks (2012) verify the Day One 1200 UTC COs produced by the SPC from 1973–2010. The authors choose the 1200 UTC outlook because they are valid for a 24-hour period until 1200 UTC the next day. The slight and moderate outlooks had the greatest longevity and therefore were chosen. The outlooks were verified by grouping storm reports from the SPC into 24-hour blocks, 1200–1200 UTC, and plotted on the same grid size (80-km spacing) as the SPC. Since both COs and reports are considered dichotomous on a grid-cell by grid-cell basis, they can be evaluated using a 2 × 2 contingency table. A variety of verification measures can be calculated using the contingency table, but the authors choose the probability of detection (POD), frequency of hits (FOH), critical success index (CSI), and bias. For slight risks, the POD between 1973–1993 increased more than any scalar attribute. However, POD remains steady for the rest of the period. Over the same period, the FOH improved slowly over time. It should be noted that it is unlikely that these changes could be achieved or resulted through increases in severe weather reports. If more reports are found, then false alarms become hits and null events become misses, which would have trivial effect on POD but would increase FOH. The increase in the FOH for the remainder of the period is indicative of well-placed risk areas that decreased in size, reducing the number of false alarms. For moderate risks, POD values are remarkably lower than slight risks, but FOH values are considerably higher. The maxima for the POD occur most frequently during the spring (March–May), while minima occur most frequently in autumn (September–November). Similarly, the FOH maxima occur most frequently in the spring, and minima occur most frequently during autumn and into December.

A different method of verifying SPC's COs was conducted by Hitchens et al. (2013). They used the practically perfect (PP) forecast probability values as a threshold and thus allowed them to create a 2×2 contingency table to analyze the SPC's day 1 slight COs from 1973–2011. Two ten-year periods were additionally picked to study: the first period 1982–1991 and the second period 2002–2011. The purpose of the PP forecast was to evaluate the skill of the forecaster if they had prior knowledge of the severe weather events that day. On a 365-day average, relative skill was steady until 1995; from 1995 relative skill increased throughout the rest of the study period. The increase in skill is similar to findings in Hitchens and Brooks (2012) as in the same year and afterward, POD and FOH values increased. For both periods, a maximum in the frequency of skillful forecast occurred in the spring months (March–June).

Hitchens and Brooks (2014) used the PP forecast and the 2×2 contingency table to evaluate the slight risks outlooks (Day 3 to Day 1) to assess skill with lead time. Data are available from 2002 to 2011 for the 1200 UTC day 3 CO, 2000 to 2011 for 1200 UTC and 1700 UTC day 2 CO, and 1973 to 2011 for the 1200 UTC day 1 CO. When placed on a performance diagram using the 2000–2011 data for all outlooks, the authors found with decreasing lead time, the day 1200 UTC CO possessed the higher POD and CSI values. However, with other outlooks there appeared to be no pattern of improvement. They inferred with decreasing lead time, the location of outlooks was either better, or the size of the outlooks increased, improving the values. Although, each outlook had improved in relative skill since the start of 2000. Like findings in Hitchens et al. (2013), the frequency of skill in daily forecasts reached a maximum in April–June for each outlook. Looking into each update for Day One, relative skill increased for the 1200 Day One CO and the proceeding updates. Lastly, the 1630 and 2000 UTC update showed to have the highest relative skill and frequency of skillful forecasts.

A recent study conducted by Hitchens and Brooks (2017) looked at verifying significant probabilistic convective outlooks (PCOs) from day 3 through the 2000 UTC Day One CO update from 2005–2015. The authors used the PP forecasts and contingency table, same as in Hitchens et al. (2013), to evaluate the PCOs. The authors find there are differences between the significant convective outlook and probabilistic outlooks for all three hazards. For tornadoes, both the significant and probabilistic outlook had high POD values. But the tornado significant outlook had higher FOH and CSI values in comparison to the tornado probabilistic outlooks. The Hail probabilistic outlooks had higher POD, FOH, and CSI values in comparison to the significant outlook. There was a stark difference between the wind probabilistic and significant outlook in the POD values. Wind probabilistic outlooks POD values were more than 0.4 higher than the significant outlooks. Overall, wind significant and probabilistic outlooks had lower forecast skill than tornadoes and hail, specifically significant wind events. Interestingly, tornadoes and hail had higher bias scores than wind outlooks. So, the higher POD values for tornadoes and hail could be due to the larger size of the convective outlook.

Rather than using a contingency table to analyze COs produced by the SPC, Herman et al. (2018) used the Brier skill score, reliability diagrams, and ArcGIS Pro to evaluate SPC's PCOs. Results from the study show the SPC has the highest skill forecasting for severe wind and the lowest skill for severe tornadoes. However, the opposite is true for significant outlooks. The authors found significant tornadoes had the highest skill, and significant winds had the lowest forecast skill. The poor forecast skill for significant wind PCOs is comparable to the results in Hitchens and Brooks (2017).

Lastly, the change in forecast skill with decreasing lead time of the SPC outlooks can be explained by Hitchens and Brooks (2014). First, a forecaster does not start with a blank slate when forecasting severe weather. They begin the process aware of the previous update outlooks. Second, consistency is key in not drastically changing the location of the outlook and or the message intended for the users. Later updates give opportunities for forecasters to change locations of the CO in certain situations. Other factors leading to a higher forecast skill is the better location of an outlook, change in the size of the outlooks, lead time, the change in forecast philosophies, and the improvement of numerical models could aid forecast skill (Hitchens and brooks 2012 and 2014).

2.3 Gaps and Summary

The summary of findings of environmental parameters that are useful to distinguish between storm modes and severe weather parameters follow. For tornadoes, VTP, SCP, STPfix, STPeff, STP, CAPE3, LR75, LR3, EHI, SRH1, SHR1, MLCAPE, ESRH, and MLCAPE3 were useful parameters to distinguish between violent/significant and weak tornadic storms (Rasmussen and Blanchard 1998; Rasmussen 2003; Thompson et al. 2003; Hampshire et al. 2018; Coffer et al. 2019). Low-level shear and thermodynamic parameters were specifically useful in distinguishing between significant and weakly tornadic storms. Authors also found VGP, SRH3, SRH1, EHI1, MLLCL, MLCAPE, SCP, STP, VTP, BRN shear, LR3, and ESRH are useful to distinguish between significant tornadic and nontornadic environments (Rasmussen and Blanchard 1998; Rasmussen 2003; Thompson et al. 2003; Thompson et al. 2007; Hampshire et al. 2018).

Few studies have been conducted to differentiate between hailstone size and hailstorm environments. Although, authors have found deep-layer shear, SHERBE, SHERBS6, SHR6, SHREL, SHR10, LHP, SSP, and LR75 were useful to distinguish between hailstone size (Rasmussen 1998; Craven and Brooks 2004; Schneider and Dean 2008; Sherburn and Parker 2014; Johnson and Sugden 2014). There has been a large concentration that has left a gap in the research on what environments produce damaging wind or severe hail events. There is a better understanding of the environments in which convection produces large hailstone sizes, but there is not enough on the topic on what environments aid in a success forecast for hail. This is especially true for research on damaging wind events. Wind magnitudes, wind shear, and thermodynamic parameters could be useful in forecasting damaging wind events; however, forecasting wind events is more complex than the other two hazards (Evens and Doswell 2001; Kuchera and Parker 2006; Sherburn and Parker 2014). Organized wind events can come from several types of organized convection like discrete cells, QCLS, derechos, and MCS systems (Dean). As a result, studies have found the SPC has poorer skill in forecasting for damaging wind and significant wind events (Hitchens and Brooks 2017). Lastly, forecast verification has focused on the forecast (SPC CO or tornado watches), but little research has pursued why the forecast was successful or did poorly. This research hopes to find storm environments and parameters that can differentiate between forecasts that are skill, biased, or poor.

3. Data and Methods

The following two sections detail the data used in this study and the methodology to fill in the gaps described in the introduction and literature review. The following questions should be kept in mind throughout the results, discussion, and conclusion. What environmental characteristics are present for forecasts when predicting hail, tornado, or wind events? Are there similarities or differences in the forecasts for each hazard? What environments do forecasters struggle with or perform well in?

3.1 Data

The SPC produces four daily COs and one severe weather outlook that covers various time periods from when the forecast was issued and are valid for 24-hours from 1200–1200 UTC the next day (Hitchens and Brooks 2014). The products are the day one, two, and three convective outlooks, and the day 4–8 outlook. The Day One CO is issued at 1200 UTC and updated four times throughout the day (The Day One CO is issued at 6Z but is recognized as the 1200Z outlook). The first update is the 1300 UTC CO, the second is 1630 UTC CO, the third is 2000 UTC CO, and fourth is the 0100 UTC CO all valid until 1200 UTC the next day. The Day Two CO is issued at 0600 UTC (30 hours prior to the valid forecast time) and updated at 1730 UTC (18.5 hours prior to the forecast's valid time). The Day One and Day Two describe the threat of severe weather using probabilities. The Day Three forecast severe weather describes the risk of severe weather using categorical risks. The Day Three CO is issued at 1200 UTC, 48 hours prior to forecasts valid time. The day 4–8 Severe Weather Outlook depicts the probability of severe weather 72 to 192 hours prior to the 1200 UTC forecast's valid time using 15% and 30% probabilities.

The Day One 1200 UTC PCOs were collected and stored from March 26, 2008, through the year of 2019 as second edition (or Edition 2) GRIdded Binary (GRIB2) data files. The GRIB2 files are only reliably available starting on March 26, 2008 (A. R. Dean 2019, personal communication). Inside the GRIB2 files are categorical CO and PCO for severe and significant weather hazards. The CO risks and PCO contours are stored as a series of coordinates that outline individual polygons (Hitchens and Brooks 2012). Inside the PCO are probability contours for each severe weather hazard that depend on the risk for a certain day; the probability of severe weather hazards corresponds to the level of risk (https://www.spc.noaa.gov/misc/about.html#Convective%20Outlooks). Only the 2% (tornado) and 5% (wind/hail) polygons within the PCOs were used in the study. The 2% contour for tornadoes and 5% for wind/hail encloses the higher probability contours. For example, the tornado PCO on May 21st, 2019, the 5 and 10% contours are enclosed by the 2% contour (see Fig. 3.1). The lowest probability for each severe weather type is intended to encompass each observed report, while the higher probability contours would see a higher density in observed reports. Thus, the lowest probability contour looks at the overall forecast for each severe weather type.

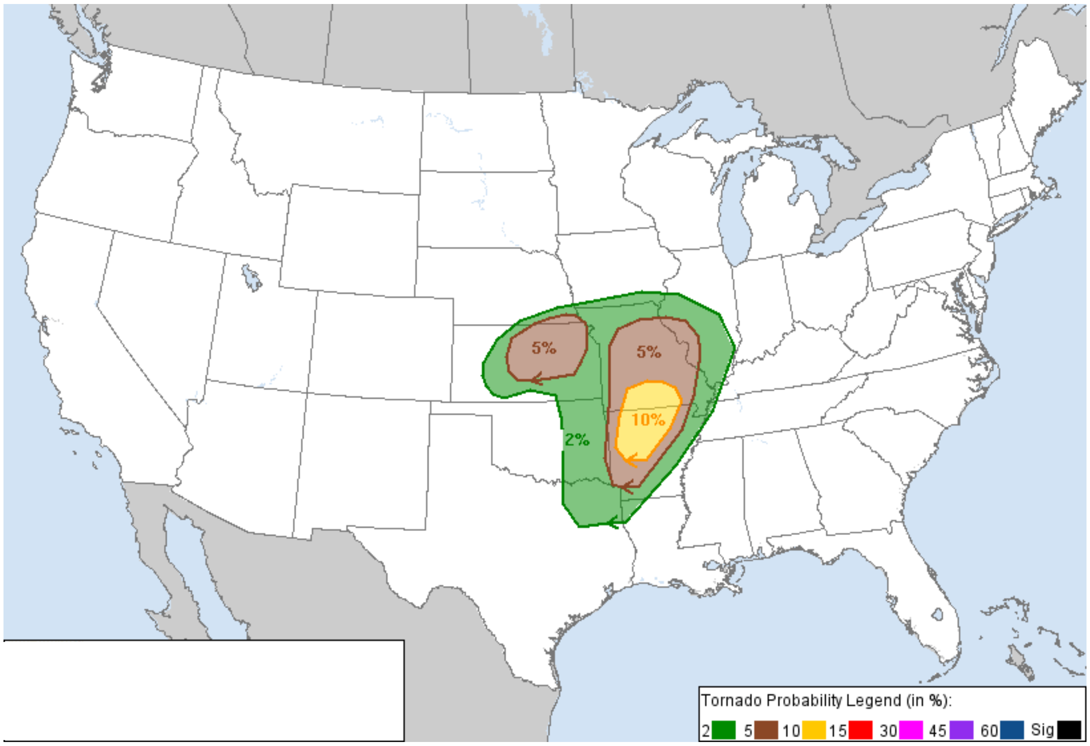

Fig. 3.1. Image of the SPC's Day One tornado PCO showing the 2, 5, 10% probability contours on May 21st, 2019. This image was gathered and available at https://www.spc.noaa.gov/exper/archive/event.php?date=20190521.

Archived severe weather reports from March 26, 2008, through the year of 2019 were collected from the SPC's warning coordination meteorologist page (https://www.spc.noaa.gov/wcm/). This particular time period was selected to match the available PCO data. The time and date, EF rating for tornadoes, wind speed for damaging wind events, hailstone size for hail events, and starting latitude and longitude were used to verify the outlooks. All reports from the SPC's warning coordination meteorologist page is filtered to remove inaccurate data and duplicate reports. The current NWS criteria for severe reports is a tornado, winds ≥ 50 kts (58 mph), and hail ≥ 1 in.

The archived surface object analysis (SFCOA) was used as the environmental data for each storm report. The SFCOA uses the 40-km grid resolution Rapid Update Cycle (RUC) model to derive sounding, composite, and diagnostic parameters (Bothwell et al. 2002). To collect the data, the General Environmental Meteorological Package (GEMPAK) was used to gather parameter data to the nearest grid point and the hour the storm report occurred to create a database (Sherburn and Parker 2014; Sherburn et al. 2016).

3.2 Verification process

The SPC 1200 UTC Day One PCO is the first probabilistic forecast issue for that day and is valid until 1200 UTC the following day. The 1200 UTC PCO was chosen because it is valid for 24-hours and is the last forecast made for the entire 24-hour period. Additionally, the other day one updates only cover a portion of the 12–12 UTC period. Hitchens and Brooks (2014) found after each day one update the forecast skill improved since forecasters have more information about how the environment would evolve. Therefore, the 1200 UTC PCO was used to understand forecast skill/bias and their environmental characteristics at the beginning of the day for all three severe weather hazards. Lastly, only days that the SPC issued a PCO and had at least 10 reports with respect to each hazard were used. For instance, if on one day there were nine hail reports, 30 tornado reports, and four wind reports, only the tornado PCO would be used. The 10 report threshold was used for multiple reasons. First, the threshold concentrates on days with a PCO that has multiple storms. Second, it eliminates days where a single storm produces only a few reports and days when there are no reports. Third, it is an efficient way to verify the PCOs to analyze storm characteristics and allows enough samples for a robust dataset.

Storm reports were grouped into 24-hour blocks from 1200 to 1200 UTC for each day. The PCO was verified using an 80.5 km grid spacing, the same grid size used in the SPC products, and Hitchens and Brooks (2012 and 2014). The reports and outlooks were verified on the same grid size. The reports were then compared to the outlooks on a grid-point by grid-point basis. Each grid cell was considered dichotomous, where one cell with multiple reports does not have greater influence over a cell with one report (Hitchens and Brooks 2012). Lastly, if the outlooks are in any part of the grid box, then the whole box is considered to be considered as part of the forecast.

Since both PCOs and storm reports are considered dichotomous based on a grid-cell by grid-cell basis, a 2 × 2 contingency table can be used to evaluate the forecasts. The 2 × 2 contingency table (Table 3.1) contains four components; hits (a), false alarms (b), misses (c), and null events (d). A hit (a) is an event that was forecasted and observed. An event that was forecast for but was not observed is a false alarm (b). A miss (c) is if no forecast was made but a report was observed. Lastly, if no forecast was made and a report was not observed then it is classified as a null event (d).

Table 3.1. A 2 × 2 contingency table with forecasts on the y-axis and observed on the x-axis.

	Observed yes	Observed no
Forecast yes	a	b
Forecast no	c	d

Several verification measures can be calculated using the table, but only the scalar attribute bias and the HSS are used (Doswell et al. 1990; Wilks 2011). Bias represents the forecasters' tendencies and serves as a useful verification measure to evaluate forecasts. Furthermore, bias can be used to evaluate the size and placement of the PCO. Bias compares the size of the forecast area (PCO) to the area covered by observed reports. Mathematically, it describes the frequency of forecast yes events (hits + false alarms) compared to the number of observed yes events (hits + misses). The scalar attribute can also be used to categorize forecasts into three classes: overforecast, underforecast, and unbiased. An overforecast indicates that size of the forecast area is much larger than the area covered by observed reports with a numerical number greater than one. The opposite is underforecast, the area covered by observed reports is greater than the forecast area with a numerical number under one. Finally, unbiased forecasts are a one-to-one ratio where the size of the forecast area is the same as the area of observed reports and has a numerical number of one.

$$\text{Bias} = \frac{a+b}{a+c} \tag{1}$$

The other verification measure, HSS, evaluates the performance of a forecast relative to the proportion of correctly forecasted events that would be achieved by random forecasts that are statistically independent of the observation (Wilks 2011). Thus, if the forecast is better than what could be predicted at random, it is skillful. A perfect forecast would be if HSS equals one, and no skill is when HSS less than or equal to zero. The HSS was used over other verification measures as gives credits and acknowledges correct forecast of null events null, and the effect of false alarms is also considered (Doswell et al. 1990). Correct null forecasts dominate the contingency table, as there are far more null events compared to hits or false alarms on a typical day. The inclusion of null events better portrays the total forecast. Moreover, the HSS is useful in discerning between skillful and unskillful forecasts for severe weather hazards (Doswell et al. 1990). It is possible that an overforecast or underforecast could be skillful, and that an unbiased forecast could be unskillful. Thus, the two verification measures were used to create three categories: poor skill, overforecast, and good skill forecasts (null dataset).

$$\text{HSS} = \frac{2(ab-bc)}{(a+c)(c+d)+(a+b)(b+d)} \tag{2}$$

3.3 The three categories of forecast skill

The 85th and 15th percentile of bias and HSS were used to sort the verified PCO days with 10 or more reports into three categories: overforecast, good skill forecast, and poor skill forecast (See Table 3.2). All of the PCO days that meet the 10 report threshold are exclusive within each category so that there is no overlap. The 85th and 15th percentile for each hazard were chosen to ensure there were enough data samples, but also to analyze the severe cases of the verified hail, tornado, and wind PCO days at both ends of the spectrums (Table 3). For example, the HSS 85th percentile looks at the skilled forecasts, and bias 85th percentile analyzes the extreme days where the size or placement of a PCO was substandard. Furthermore, the percentiles are unique and with respect to each hazard. Finally, the verified PCO were divided into three categories for each hazard to study the environmental characteristics and compare the differences or similarities in each of the categories. The data distribution of each hazard and category can be seen in Fig. 3.2.

Table 3.2. Verification statistics and percentiles of bias and HSS for each hazard, and the number of days that meets the category threshold for each hazard.

Description	Hail	Tornado	Wind
Number of days with 10 reports or more	1964.000	444.000	2089.000
Number of days with an overforecast	270.000	53.000	267.000
Number of days with good skill forecasts	165.000	42.000	171.000
Number of days with poor skill	55.000	8.000	49.000
15th percentile bias	4.678	6.793	5.715
85th percentile bias	21.899	39.067	31.367
15th percentile HSS	0.034	0.013	0.013
85th percentile HSS	0.241	0.187	0.193

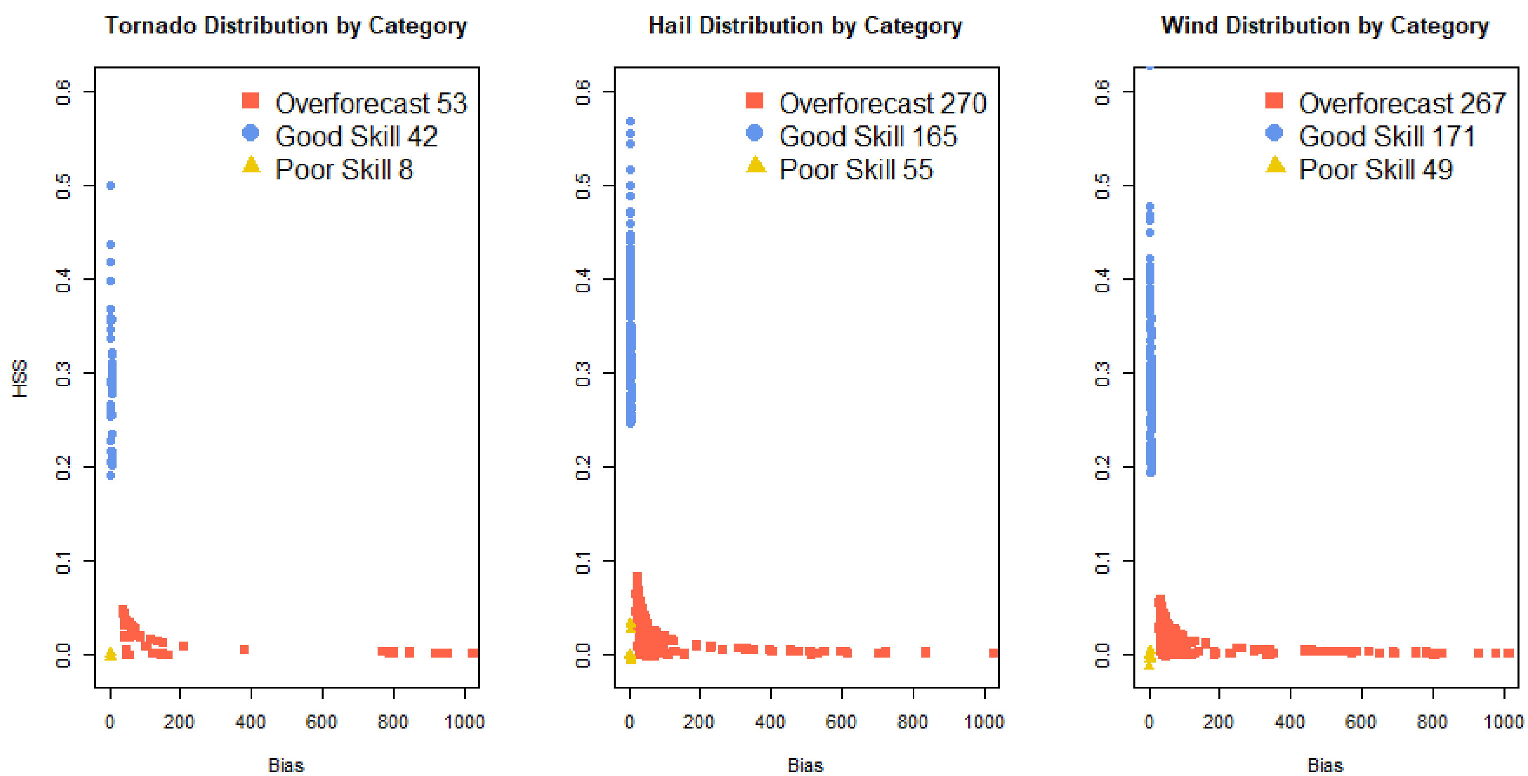

Fig. 2.2. Forecast outlook distribution data by category, starting with tornado to the left, hail in the middle, and wind to the right. The square symbol represents overforecasts, circles good skill, and triangles for poor skill.

Overforecast was defined as a PCO with 10 or more reports and with a bias greater than the 85th percentile with respect to each hazard (refer to Table 3.2). This category was used to analyze the environmental characteristics of days with severe overforecast for all three hazards. Moreover, the overforecast category studies days with a large outlook area but only a few reports. It was chosen to use each hazard's 85th percentile bias as the size of the PCOs varies for each hazard.

Analyzing the distribution of bias values of the PCOs, it was found the SPC rarely had a PCO that verified as a underforecast (bias ≤ 1). Thus, using this category would result in a low amount of sample. The number days with a bias ≤ 1 was for tornadoes three, for hail 51, and wind 32. Looking at Table 3.2, the 15th percentile of bias for all hazards represents a low overforecast value. Thus, the poor skill category, created by using bias and HSS, was used to add more samples to the dataset. Poor skill forecasts were defined as a day with 10 and more forecasts, an HSS less than the 15th percentile, and a bias less than the 15th percentile (refer to Table 3). The poor skill category was used to observe the storm characteristics on days where the PCO for all hazards was either too small or misplaced. Thus, both classes were created to compare the environmental differences or similarities of all three hazards.

The good skill category was used as a null dataset to compare to overforecast and poor skill forecasts. The good skill forecasts were defined as a day with 10 or more reports, an HSS greater than the 85th percentile with respect to each hazard, and a bias greater than one but less than the 85th percentile with respect to each hazard (refer to Table 3.2). Thus, the storm environment of the good skill forecast can be used to compare with the other two categories and three hazards.

3.4 Storm environments

There are several severe weather parameters that can be used to study each hazards category. Though there are some parameters that are more useful than others, using a single parameter as a silver bullet to forecast a severe weather hazard is unwise. Thus, analyzing various environmental aspects of the PCO can further the understanding of when forecasters struggle or do well in (Doswell and Shultz 2006). To understand the environments of each category, a handpicked list of 21 parameters from SFCOA were used (see Table 3.3).

Table 3.3. List and description of environmental parameters within the SFCOA database.

Description	Units	Abbreviation	Hazard studied
0–3 km MLCAPE	J/kg	MLCAPE3	Tornado
0–3 km VGP	numeric	VGP3	Tornado
100 mb mean mixed CAPE	J/kg	MLCAPE	Tornado and Hail
100 mb mean mixed CIN	J/kg	MLCIN	Tornado
100 mb mean mixed LCL height	m	MLLCL	Tornado
Craven-Brooks Sig Severe	m^3/s^3	SSP	Hail
Derecho composite parameter	numeric	DCP	Wind
Downdraft CAPE	J/kg	DCAPE	Wind
Effective helicity	m^2/s^2	ESRH	Tornado
Effective shear	knots	ESHR	Tornado
Effective-layer STP	numeric	STPeff	Tornado
Lapse Rate from 700 to 500 mb	C/km	LR75	Hail
Lapse rate surface to 3km	C/km	LR3	Tornado
MCS maintenance probability	%	MCSM	Wind
Most Unstable CAPE	J/kg	MUCAPE	Hail and Wind
MUCAPE from -10C to -30C	J/kg	FMUCAPE	Hail
Precipitable water	inches	PWAT	Hail
Storm relative helicity surface to 1 km	m^2/s^2	SRH1	Tornado
Surface to 1 km shear	m/s	SHR1	Wind
Surface to 6 km shear	m/s	SHR6	Hail, Wind, Tornado
Surface to 8 km shear	m/s	SHR8	Wind

In the literature of storm environments, tornadoes have been the primary focus. Past articles have discussed tornadoes frequently occur in environments of moderate SHR6 and MLCAPE and are used extensively used in forecasting (Thompson et al. 2003; Craven and Brooks 2004; Thompson et al. 2007; Thompson et al. 2012; Dean and Schneider 2012; Hampshire et al. 2018). Additionally, ESHR, ESRH, and SRH1 are useful distinguishing between tornadoes and the other two severe weather hazards and forecasting organized convection that produces either a severe or significant tornado ((Thompson et al. 2003; Craven and Brooks 2004; Thompson et al. 2007; Thompson et al. 2012; Sherburn et al. 2016; Hampshire et al. 2018; Coffer et al. 2019). Low-level instability can be a key to tornadogenesis and organized convection. Therefore LR3, MLCAPE3, and VGP3 were selected (Rasmussen and

Blanchard 1998, Rasmussen 2003; Davies 2006; Hampshire et al. 2018). Mixed-layer LCL and CIN have been documented to be useful in differentiating between significant and weakly tornadic tornadoes, and useful predicting tornadic convection (Thompson et al. 2003; Craven and Brooks 2004; Thompson et al. 2012; Sherburn et al. 2016; Hampshire et al. 2018). Lastly, the STPeff was created using MLCAPE, MLCIN, MLLCL, ESHR, and ESRH (Thompson et al. 2012). It is also used repeatedly to forecast for significant tornadoes. It has been found to be useful forecasting tornadic convection and distinguishing between tornadic and hail/wind convection (Thompson et al. 2012; Sherburn et al. 2016; Hampshire et al. 2018; Coffer et al. 2019). Thus, these parameters were picked to analyze the environments of the verified PCOs.

Deep-level shear (SHR6) is commonly used forecasting organized convection, and but also for the development for a strong updraft that is key component for hail events (Schneider and Dean 2008). In addition, CAPE is another vital part for forecasting organized convection and the establishment of a strong updraft. Therefore, MUCAPE and MLCAPE have been known to be useful for forecasting hail events (Craven and Brooks 2004; Schneider and Dean 2008; Johnson and Sugden 2014). Environments with high MUCAPE between -10°C to -30°C values, temperature zone is in the prime hail growth zone, can increase the probability of a hail event (Johnson and Sugden 2014). Steep lapse rate between 700–500 mb has been useful to discriminate between hail and tornado/wind events (Cravens and Brooks 2004; Johnson and Sugden 2014). Plus, strong lapse rates in the middle of the atmosphere increase the buoyancy and updraft in a storm. Lastly, SSP has been useful forecasting for hail events and organized convection (Craven and Brooks 2004).

The literature on forecast parameters used to predict damaging wind events is even less comprehensive than hail and tornadoes. The DCP composite parameter and MCS maintenance probability parameter have been useful predicting wind events (Evens and Doswell 2001; Coniglio and Corfidi 2006). In addition, moderate DCAPE and MUCAPE values have been found useful forecasting for wind events from various storm types (Evens and Doswell 2001; Kuchera and Parker 2006). Lastly, wind shear is a key ingredient and organized convection, and could be used to forecast wind events. Thus, several wind shear parameters were used (Kuchera and Parker 2006; Schneider and Dean 2008).

Since the purpose of the paper was to investigate the similarities and differences of the PCOs environments, SFCOA was used to collect environmental data to the nearest hour of a storm report. The max parameter value of the first report within an 11.1 by 11.1 km box was recorded. It should be noted that search radius may not occur in the location of most favorable conditions. The first report was considered if it was within the time range of 12 UTC through 1159 UTC the next day. The 11.1 by 11.1 km box was made by using the starting latitude and longitude of the report, and this would make the total search area a 22.2 by 22.2 km square centered on the first report. The max parameters value of the first report within the grid and time range was used to represent the day of each hazard's category environment. To illustrate, a box was made of the latitude and longitude of the first hail report under the overforecast category. Then, the max value for the seven parameters (see Table 3.3) within the grid was collected and would represent the environment for that day.

The advantage of distilling it down to one value compared to using all reports was it would prevent one day with multiple reports having more influence than a day with fewer reports. Thus, days with a higher number of reports would have a greater influence when analyzing the data creating a bias in the dataset. Another advantage of using the max value of the first report would give insight into the environment for when convection was first initiated. The downside of this method that was it does not accurately portray the overall environmental characteristics. Using one report only gives a small insight into the environment of a PCO.

There were two statistical tests used to analyze the dataset, the two-sample difference of means test, and the one-way analysis of variance (ANOVA) test. Since there were only two parameter datasets for tornadoes, the two-sample difference of means test was used to determine if two population means (overforecast and good skill) are equal or different. The ANOVA test was used to see if there was variance in the mean of the three parameter datasets of hail and wind. The Tukey test, a post-hoc test, was used to find out which specific group's means (compared with each other) are different. The test compares all pairs of parameter means. To visually assess the datasets, parameters were plotted onto boxplots to help characterize the environments of each category and hazard.

4. Results

The following paragraphs discuss the results of the environmental analysis of the three hazards. Each hazard section examines the verification result, parameter distribution, boxplots, and 2D density scatter plots. Each category's environment was compared to see if there any differences or similarities.

4.1 Tornado

Out of the nearly 12-year analysis, 444 days met the 10-report requirement (see Table 4.1). From the dataset, 53 days met the criteria for overforecast, 42 for good skill, and only eight for poor skill. Since the poor skill category had only eight days, no data was collected for this category. Out of the 444 days, there were nine days in which the HSS was below zero. From the study, there were 664 days where the SPC issued a tornado PCO but verified with no reports (see section 5.4 a more detailed discussion). The SPC tended to be overly cautious when forecasting tornadic convection compared to the other two hazards. However, it was expected that the SPC would forecast on the safer side since tornadoes cause more damage and fatalities compared to the two hazards. On the other hand, the SPC rarely missed a forecast, as there were only three days with more than 10 reports and no issued PCO.

Table 4.1. Tornado verification statistics.

Description	Tornado
Number of days with 10 reports or more	444
Number of days with an overforecast	53
Number of days with good skill	42
Number of days with poor skill	8
Number of days with a PCO but no reports	664
Number of days with no PCO but more than 10 reports	3
Number of days with a HSS below 0	9

4.1.1 Statistical analysis on tornado parameters

There were eleven parameters tested to see if there was a difference in means between the tornado overforecast and good skill datasets. Out of the five parameters that make up the STPeff composite parameter, only MLLCL was statistically significant at the 1% level for distinguishing between tornado overforecast and good skill (see Table 4.2). The mean for tornado good skill was much higher compared to overforecast (See Table 4.3). Mixed-layer CIN was close to being statistically significant at the 5% level, with a p-value of 0.067; the MLCIN mean for tornado good skill was higher compared to overforecast. The other three parameters that make up the STPeff parameter (see equation 3), MLCAPE, ESRH, and ESHR, were not statistically significant in differentiating between tornado overforecast and good skill. Mixed-layer CAPE had a lower p-value at 0.277 compared to ESHR and ESRH. Curiously, ESRH and ESHR had the first and third highest p-value out of the eleven parameters. It is interesting to note that ESHR and ESRH, parameters that have been documented in the past as being useful for forecasting tornadic convection but demonstrated no difference in the means between overforecast and good skill. Lastly, the STPeff parameter, p-value of 0.075, was close to being statistically significant and distinguishing between tornado overforecast and good skill environments, with the STPeff mean being higher for tornado good skill compared to overforecast (Tables 4.3 and 4.4).

$$STPeff = \left(\frac{\text{MLCAPE}}{1500\,J/kg}\right) * \left(\frac{2000-\text{MLLCL}}{1000\text{ m}}\right) * \left(\frac{\text{ESRH}}{150\text{ m}^2\text{s}^2}\right) * \left(\frac{\text{ESHR}}{20\,m/s}\right) * \left(\frac{200\ +\ \text{MLCIN}}{150\,J/kg}\right) \quad (3)$$

Table 4.2. Tornado parameter two-sample difference of means test between overforecast and good skill. Parameters that were close to the 5% significance level and their mean were italicized, and parameters that were statistically significant at the 1% level and there mean were bolded.

Parameters	Overforecast mean	Good skill mean	p-value
SHR6 (m/s)	48.543	50.287	0.489
ESHR (m/s)	37.619	36.958	0.631
ESRH (m^2/s^2)	420.439	424.067	0.908
SRH1 (m^2/s^2)	481.779	459.454	0.564
MLCIN (J/kg)	*-34.121*	*-21.154*	*0.067*
MLLCL (m)	**752.490**	**928.897**	**< 0.001**
MLCAPE (J/kg)	2153.863	2424.882	0.277
STPeff	*2.387*	*3.026*	*0.075*
VGP3	0.331	0.334	0.869
LR3 (C/km)	8.851	9.321	0.134
MLCAPE3 (J/kg)	183.175	169.359	0.243

Deep-level shear and MLCAPE are both common parameters used when forecasting organized convection. However, neither parameter could distinguish the means between tornado overforecast and good skill. Comparing the p-values of parameters, SRH1 had a much lower p-value compared to ESRH. Analyzing the overforecast and good skill parameter tornado distribution, the values for SRH1 for good skill were lower compared to overforecast. Effective SRH had similar distributions between overforecast and good skill. Next, comparing STPeff and VGP3, there was a notable difference in the p-value between the two parameters: STPeff clearly outperformed VGP3 in differentiating between tornado good skill and overforecast. The low-level instability parameters, LR3 and MLCAPE3, failed to distinguish between the two categories of good skill and overforecast. However, looking at the p-value, there were some differences in the means between good skill and overforecast. Good skill was characterized by moderate MLCAPE3 values, while overforecast tornado events had higher MLCAPE3 distribution compared to good skill. Tornado good skill days where characterized by higher LR3 values compared to overforecast.

Table 4.3. Overforecast tornado parameter percentiles.

Parameters	25th percentile	50th percentile	75th percentile
SHR6 (m/s)	38.021	51.989	57.609
ESHR (m/s)	32.395	38.646	41.763
ESRH (m^2/s^2)	327.150	422.600	489.690
SRH1 (m^2/s^2)	371.270	468.610	613.790
MLCAPE (J/kg)	1017.440	2040.190	2841.880
MLCAPE3 (J/kg)	152.200	184.540	219.210
MLCIN (J/kg)	-49.620	-12.220	-4.135
MLLCL (m)	549.640	660.550	799.650
LR3 (C/km)	7.440	9.100	10.280
STPeff	1.460	2.070	3.250
VGP3	0.270	0.330	0.380

Table 4.4. Good skill tornado parameter percentiles.

Parameters	25th percentile	50th percentile	75th percentile
SHR6 (m/s)	41.271	49.773	60.995
ESHR (m/s)	32.644	37.329	41.177
ESRH (m^2/s^2)	297.988	374.90	507.460
SRH1 (m^2/s^2)	321.593	434.680	582.833
MLCAPE (J/kg)	1641.590	2192.310	3057.455
MLCAPE3 (J/kg)	130.740	179.065	207.293
MLCIN (J/kg)	-28.570	-11.470	-5.580
MLLCL (m)	584.940	773.575	1132.557
LR3 (C/km)	7.948	9.900	10.448
STPeff	1.758	2.480	3.880
VGP3	0.290	0.320	0.388

4.1.2 Tornado boxplots

Boxplots were used to compare categorical differences with parameters that tested statistically significant. Mixed-layer LCL was the only parameter to be statistically significant out of the group (Fig. 4.1). Good skill was characterized by higher MLLCL heights compared to the lower MLLCL heights for overforecast. Lower MLLCL could lead to overforecast bias when forecasting for tornadic convection, influencing a larger drawn tornado PCO. The noticeable difference between the two categories was that 50% of tornado events occurred when MLLCL heights were greater than 773.575 m compared to the overforecast median of 660.550, and the 75th percentile for good skill was higher compared to overforecast. The last difference between the two categories was the interquartile range (IQR). Good skill had a much larger IQR compared to overforecast.

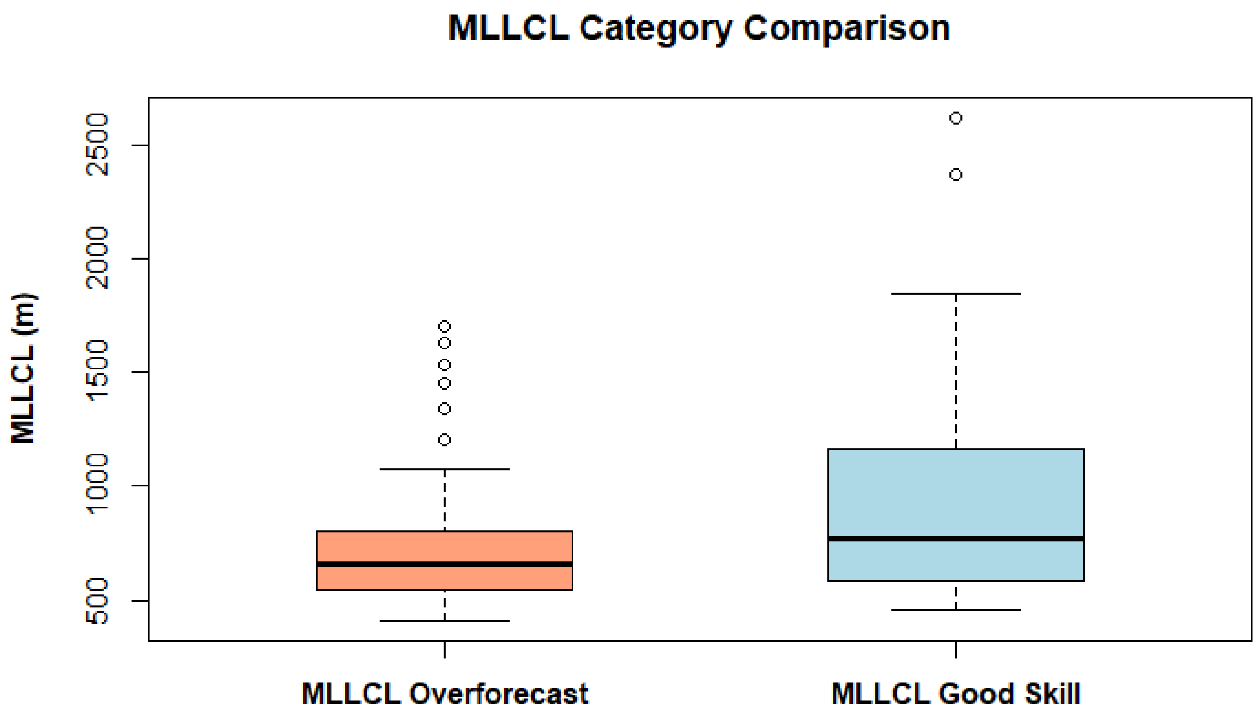

Fig. 4.1. Boxplot comparing tornado overforecast (left) and good skill (right) environment using the parameter MLLCL.

In the MLCIN boxplot (Fig. 4.2), tornado reports on good skill days were produced in lower range of values in comparison to overforecast. The range between the 25th and 75th percentile for good was -28.750 J/kg to -5.580 J/kg MLCIN vs overforecasts range -49.62 to -4.135 J/kg MLCIN. Tornado events on overforecast days occurred over a larger IQR, with a few outliers near -150 J/kg. An interesting result was the 75th percentile and median for both categories were the same but the 25th percentile differed.

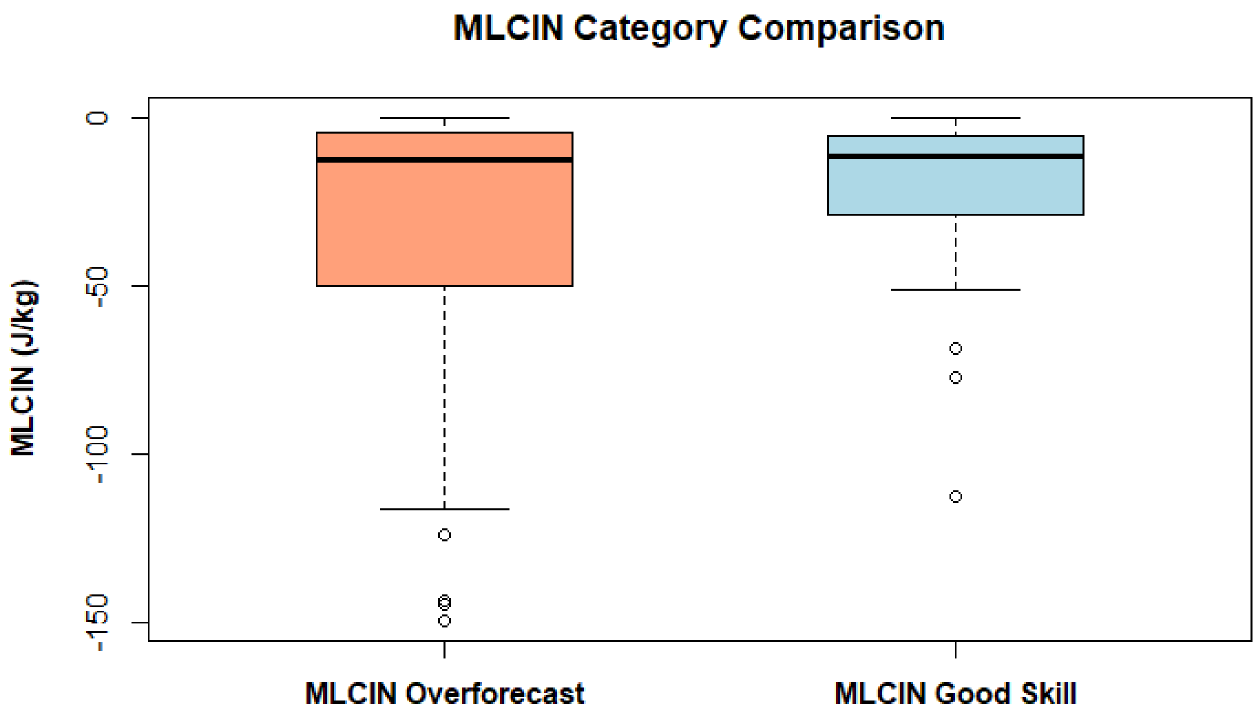

Fig. 4.2. Boxplot comparing tornado overforecast (left) and good skill (right) environment using the parameter MLCIN.

Analyzing the STPeff boxplot, tornado events on forecasts with good skill had higher STPeff values compared to overforecast (Fig. 4.3). The median, also, was higher for good skill at 2.480 compared to the overforecast median of 2.070. The higher MLCAPE and lower MLCIN values could have attributed to higher STPeff values for good skill than overforecast. Since ESHR and ESRH were similar, they did not attribute to the differences in the STPeff distribution.

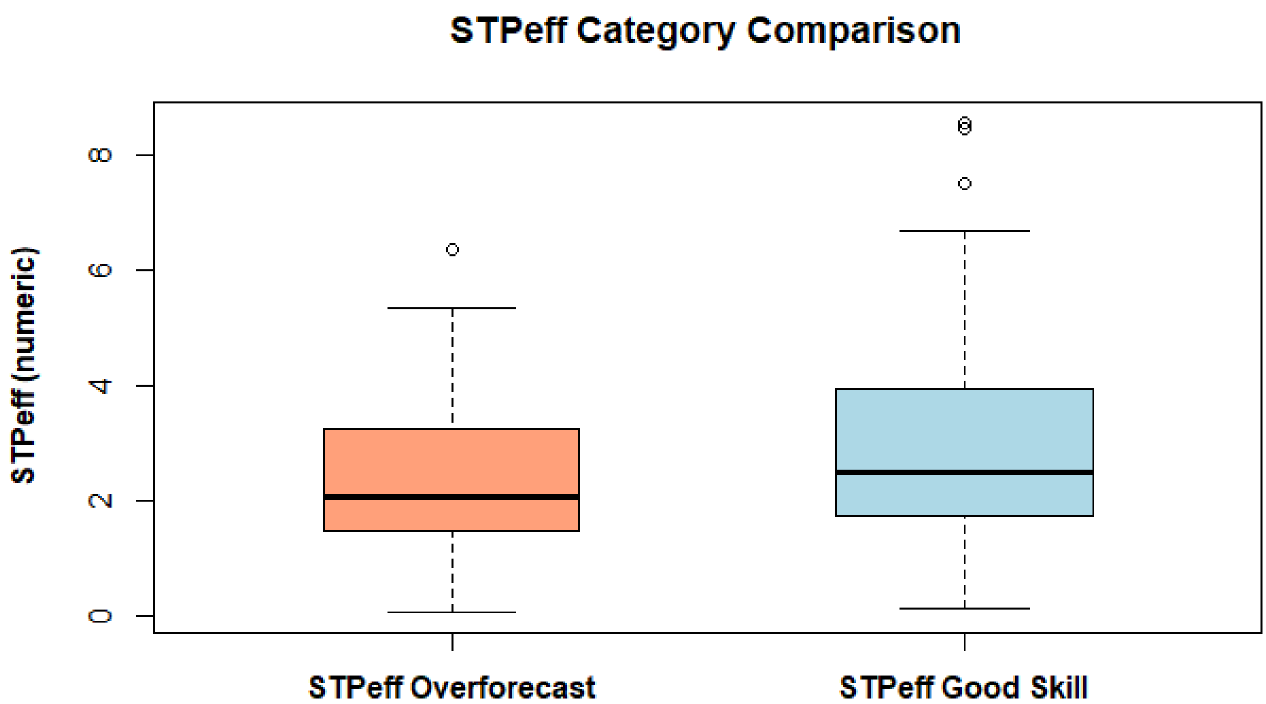

Fig. 4.3. Boxplot comparing tornado overforecast (left) and good skill (right) environment using the parameter STPeff.

4.2 Hail

From the 12-year study period, there were 1964 days (Table 9) that met the 10-report requirement. Out of the 1964 days, 270 were deemed as an overforecast, 165 days as a good skill forecast, and 55 days with poor skill. In comparison to the tornado PCO, the SPC issued 209 hail PCOs that verified without a report, far fewer days compared to tornadoes. However, there were 43 days where no hail PCO was issued and there were more than 10 reports, 40 more days compared to tornadoes. Lastly, there were the18 days with a forecast HSS below 0.

There were intriguing results looking more into the verification statistics of hail (See Table 4.5). From 2008–2013 there were 176 days considered as an overforecast. However, from 2014–2019 there were 94 cases. From 2008–2013, the outlooks tended to be large, sometimes covering two-thirds of the U.S. As a result, the PCO was then evaluated as an overforecast because of the size. This would result in poor skill and a high bias. Analyzing the poor skill PCO, there were 20 days with poor skill between 2008–2013, and 35 days between 2014–2019. For good skill, the PCO placement was better, but more importantly, the size was smaller overall. Comparing the 6-year splits, there were 65 days with good skill between 2008–2013, and 100 days between 2014–2019.

Table 4.5. Hail verification statistics.

Description	Hail
Number of days with 10 reports or more	1964
Number of days with an overforecast	270
Number of days with skilled forecasts	165
Number of days with poor skill	55
Number of days with a PCO but no reports	209
Number of days with no PCO but more than 10 reports	43
Number of days with a HSS below 0	18

4.2.1 Statistical analysis on hail parameters

Several parameters were statistically significant in distinguishing the variance between the three categories pairs. The only parameter that did not was the LR75. Distribution values for LR75 were all similar between each category; with a surprising p-value of 0.995 between hail good vs overforecast. The LR75 mean for the three hail categories were identical with little variance. Analyzing Table 4.6, PWAT was successful in differentiating between each hail categorical pair at the 5% level. It was also statistically significant at 1% for hail good skill vs overforecast and poor skill vs good skill. The mean PWAT value for hail good skill was considerably lower compared to overforecast.

Table 4.6. Hail ANOVA and Tukey test results. The p-values that were statistically significant at the 1% level were bolded, and p-values that statistically significant at the 5% level were italicized.

Parameter	Good skill-Overforecast p-value	Poor skill-Overforecast p-value	Poor skill-Good skill p-value
LR75 (C/km)	0.995	0.464	0.537
PWAT (inches)	**< 0.001**	*0.019*	**< 0.010**
MUCAPE (J/kg)	**< 0.001**	**< 0.001**	0.891
FMUCAPE (J/kg)	**< 0.001**	**< 0.001**	0.444
MLCAPE (J/kg)	**< 0.001**	**< 0.001**	0.078
SHR6 (m/s)	**< 0.001**	0.251	0.498
SSP (m^3/s^3)	**< 0.010**	**< 0.001**	**< 0.010**

Table 4.7. Hail parameter means for overforecast, good skill, and poor skill.

Parameter	Overforecast mean	Good skill mean	Poor skill mean
LR75 (C/km)	8.723	8.729	8.832
PWAT (inches)	1.921	1.591	1.772
MUCAPE (J/kg)	3847.729	2816.214	2921.121
FMUCAPE (J/kg)	993.485	839.174	773.006
MLCAPE (J/kg)	1196.137	888.718	661.518
SHR6 (m/s)	41.785	46.791	44.652
SSP (m^3/s^3)	45788.830	38310.300	24991.700

Most unstable CAPE could distinguish at the 1% level between two of the hail categorical pairs, good skill vs overforecast and poor skill vs overforecast (see Table 4.6) This parameter could not distinguish between poor hail skill vs good skill, with a large p-value of 0.891. Looking at Table 4.7, there was no difference in the MUCAPE mean value between hail good skill and poor skill. This was also the same result for FMUCAPE, as it was statistically significant with p-values less than 0.001 for the same hail categorical pairs good skill vs overforecast and poor skill vs overforecast. As with MUCAPE, FMUCAPE failed to distinguish between poor skill vs good skill with a p-value of 0.444. The datasets for both MUCAPE and FMUCAPE were similar between hail good skill and poor skill (see Table 4.8, 4.9, and 4.10). Mixed-layer CAPE had the same results as for MUCAPE and MLCAPE. Interestingly, MLCAPE had a much lower p-value of 0.078 for the poor skill vs good skill hail categorical pairing compared to MUCAPE and MLCAPE. The MLCAPE mean value for hail good skill was higher compared to poor skill; this is also true when comparing the MLCAPE mean value between hail overforecast and poor skill.

Next, there was a difference in the SHR6 variance for the good skill vs overforecast hail categorical pairing, with a p-value less than 0.001. The SHR6 mean value for hail good skill was 5 m/s higher compared to overforecast. The parameter was unable to distinguish poor skill vs overforecast pair with a p-value of 0.251, and the poor skill vs good skill pair with a higher p-value of 0.498. The SSP parameter is the product of MLCAPE and SHR6. It was the only parameter that was successful in differing between all three categorical pairs being statistically significant at 1%.

$$SSP = MLCAPE * SHR6 \tag{4}$$

Table 4.8. Overforecast hail parameter percentiles.

Parameters	25th percentile	50th percentile	75th percentile
LR75 (C/Km)	8.230	8.785	9.200
PWAT (inches)	1.660	1.950	2.180
MUCAPE (J/kg)	2759.280	3681.565	4778.160
FMUCAPE (J/kg)	715.900	961.410	1218.258
MLCAPE (J/kg)	627.357	1104.589	1619.933
SHR6 (m/s)	33.009	39.072	48.975
SSP (m^3/s^3)	28021.250	39456.500	59262.000

Table 4.9. Good skill hail parameter percentiles.

Parameters	25th percentile	50th percentile	75th percentile
LR75 (C/Km)	8.310	8.710	9.090
PWAT (inches)	1.310	1.570	1.820
MUCAPE (J/kg)	1860.380	2622.120	3719.500
FMUCAPE (J/kg)	648.720	821.120	1009.120
MLCAPE (J/kg)	441.699	749.920	1245.917
SHR6 (m/s)	38.803	45.452	55.598
SSP (m^3/s^3)	20231.000	35626.000	50842.000

Table 4.10. Poor skill hail parameter percentiles.

Parameters	25th percentile	50th percentile	75th percentile
LR75 (C/Km)	8.540	8.850	9.215
PWAT (inches)	1.355	1.830	2.135
MUCAPE (J/kg)	1819.125	2774.000	3726.000
FMUCAPE (J/kg)	520.815	759.470	952.485
MLCAPE (J/kg)	306.176	660.234	884.024
SHR6 (m/s)	31.850	42.062	54.621
SSP (m^3/s^3)	14638.500	25191.000	33984.000

4.2.2 Hail boxplots

There was a considerable difference in the PWAT distribution between hail good skill and the other categories (Fig. 4.4). For good skill, hail events occurred in lower PWAT values, with 75% of the events occurring when PWAT was less than 1.820 inches. Overforecast was characterized by higher PWAT amounts, with the median at 1.950 inches compared to good skill's 1.570 inches. Hail events occurred over large IQR for poor skill. In addition, poor skill's median was higher compared to good skill and similar compared to overforecasts median. Comparing poor skill vs overforecast, overforecast typically had hail events with higher PWAT greater than 1.660 inches. Poor skill hail events had high PWAT values with the median at 1.830 inches and a large IQR.

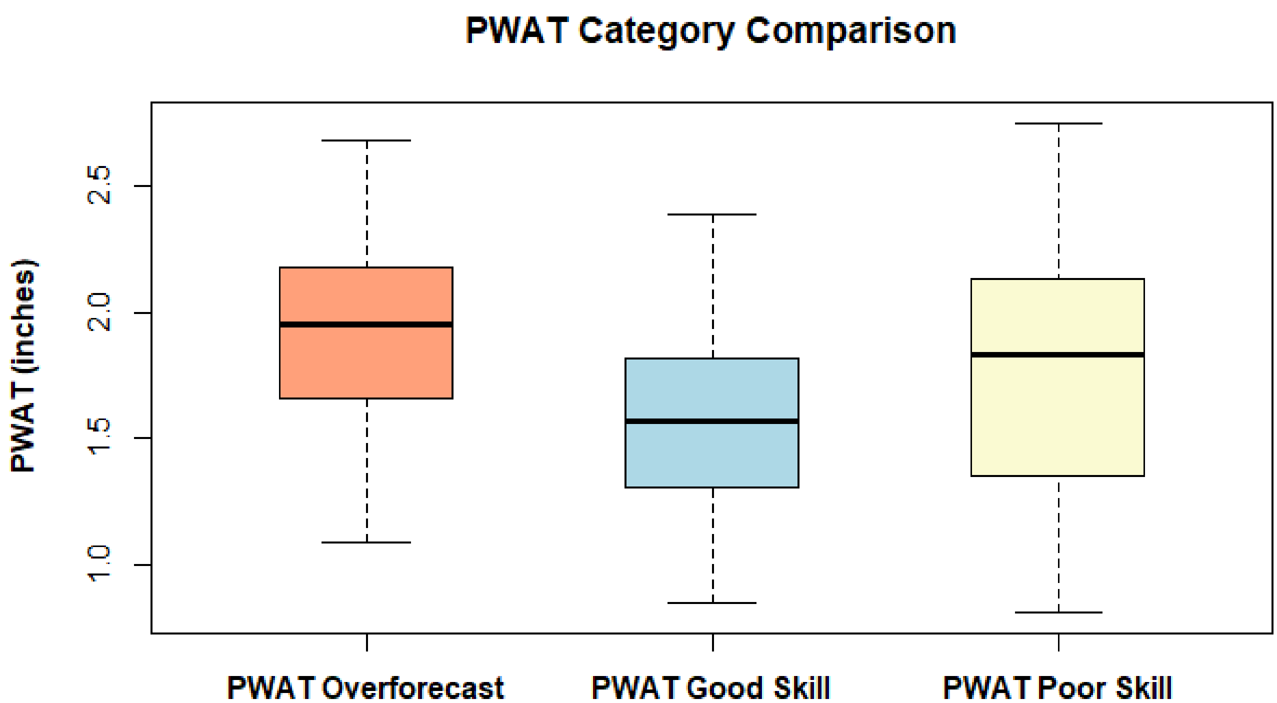

Fig. 4.4. Boxplot comparing hail overforecast (left), good skill (middle), poor skill (right) environment using the parameter PWAT.

There was a notable similarity in the distribution of MUCAPE comparing good skill and poor skill (see Fig. 4.5). Both categories had identical IQR and median. When comparing the two categories to overforecast, overforecast MUCAPE values were higher compared to good skill and poor skill. The median difference between overforecast and good skill was 1059.445, and the difference between overforecast and poor skill was 907.565. Thus, there are major differences in the environments of overforecast compared to good/poor skill, but poor skill and good skill had similar characteristics in the distribution.

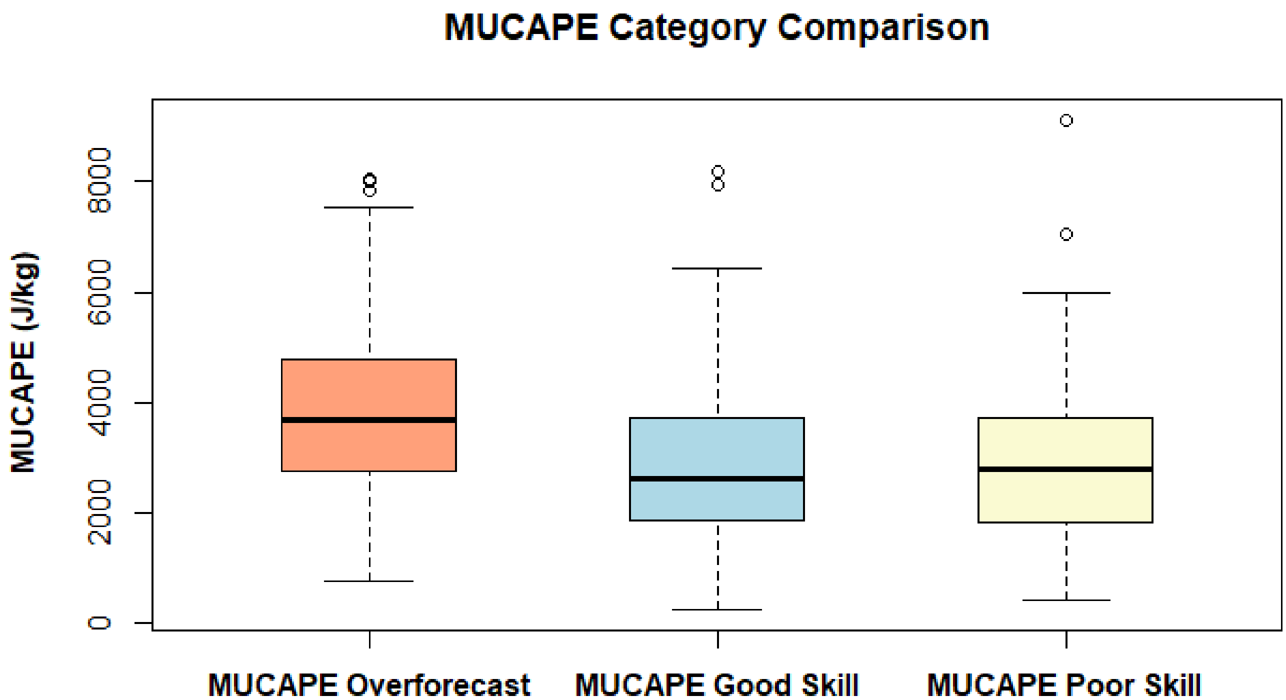

Fig. 4.5. Boxplot comparing hail overforecast (left), good skill (middle), poor skill (right) environment using the parameter MUCAPE.

The FMUCAPE median for good skill and poor skill were similar (Fig. 4.6), and overall had an identical distribution of FMUCAPE values. Like MUCAPE, the noteworthy differences were between good skill vs overforecast and poor skill vs overforecast. Overforecast FMUCAPE distribution was higher compared to poor skill. With a median difference of 201.94 J/kg. Furthermore, overforecast FMUCAPE values were also higher compared to good skill, with a median difference of 140.29 J/kg.

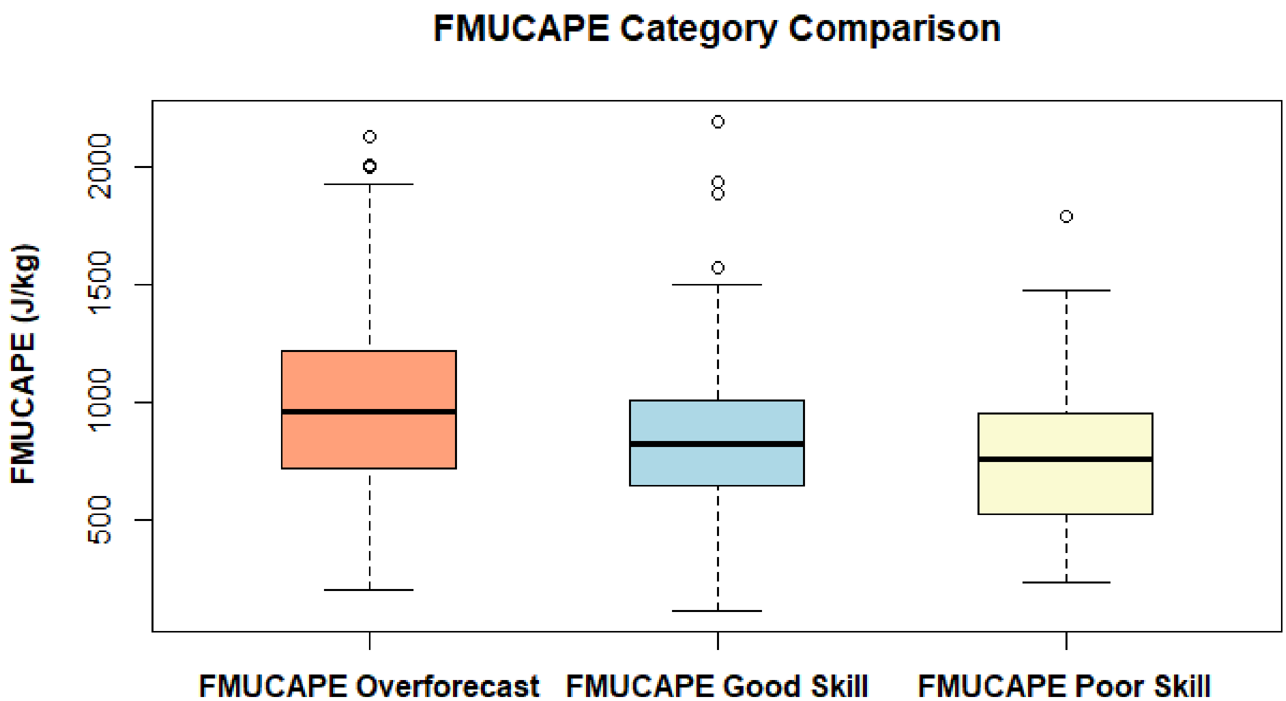

Fig. 4.6. Boxplot comparing hail overforecast (left), good skill (middle), poor skill (right) environment using the parameter FMUCAPE.

Just at the two previous thermodynamic parameters, MLCAPE had the same distribution characteristics between the three categories (see Fig. 4.7. There were higher MLCAPE values for overforecast compared to good skill and poor skill. The median difference between overforecast and good skill was 354.669 MLCAPE, and between overforecast and poor skill was 444.355 J/kg. Furthermore, MLCAPE good skill values were higher compared to poor skill. Next, there were some differences between the three categories for SHR6. Hail events for good skill had higher SHR6 values compared to overforecast; with a median difference of 6.473 m/s (Fig. 4.8). There was a large IQR for poor skill when compared to overforecast and good skill.

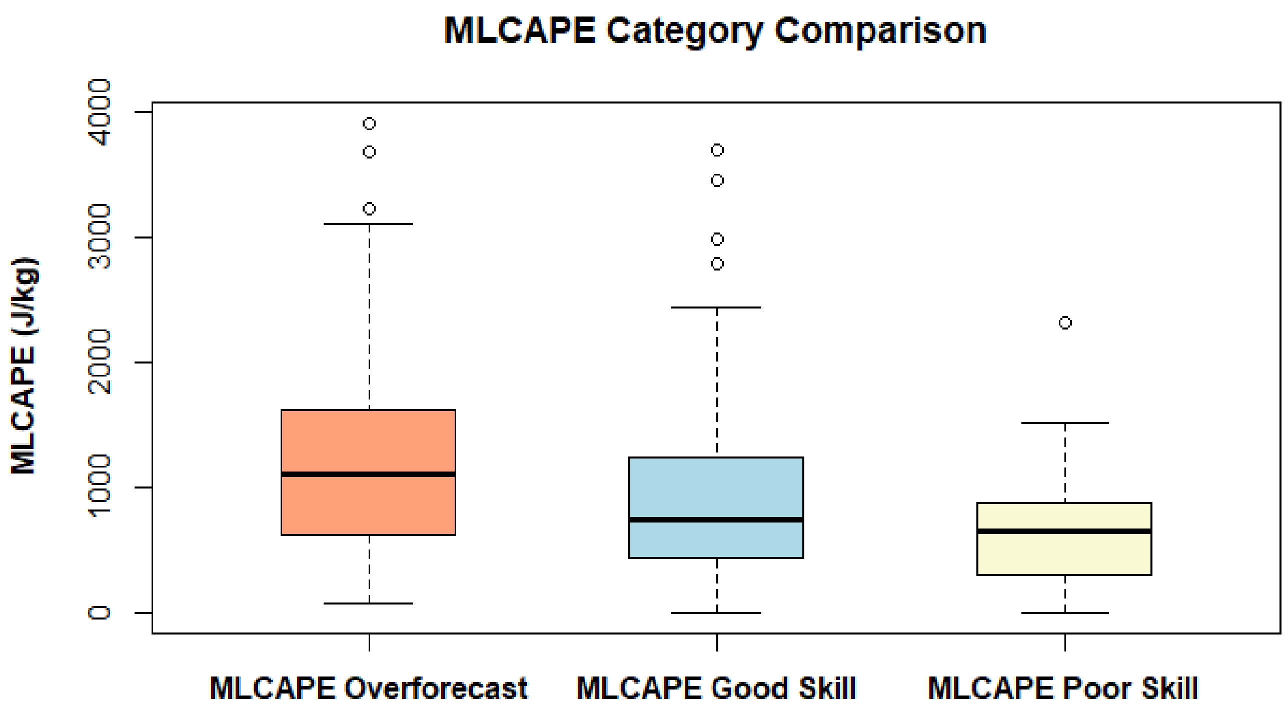

Fig. 4.7. Boxplot comparing hail overforecast (left), good skill (middle), poor skill (right) environment using the parameter MLCAPE.

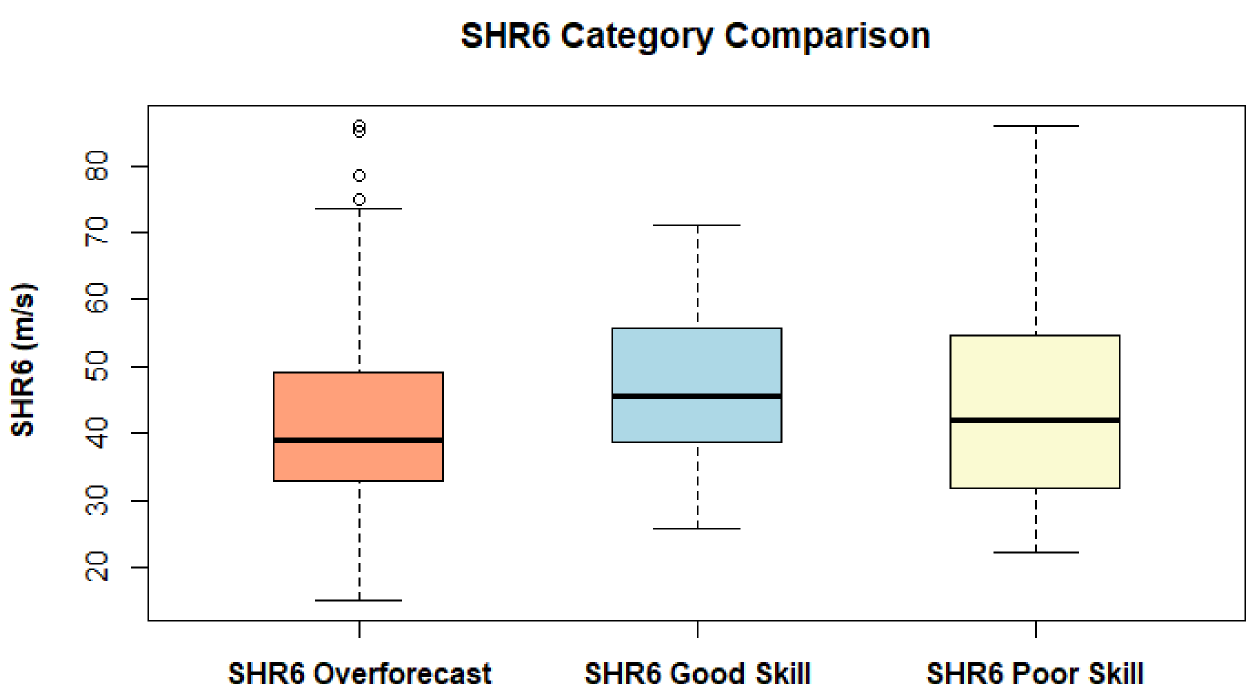

Fig. 4.8. Boxplot comparing hail overforecast (left), good skill (middle), poor skill (right) environment using the parameter SHR6.

The significant severe parameter was the only parameter that could distinguish between all three category pairs (Fig. 4.9). The median was similar between overforecast and good skill. However, SSP overforecast values were higher compared to good skill. This was due to the higher MLCAPE for overforecast. There was also a notable difference between the distribution between overforecast and good skill. The SSP overforecast values were much higher compared to poor skill. This was due to poor skill having moderate SHR6 and low MLCAPE values. Lastly, the good skill was characterized by higher SSP values compared to poor skill.

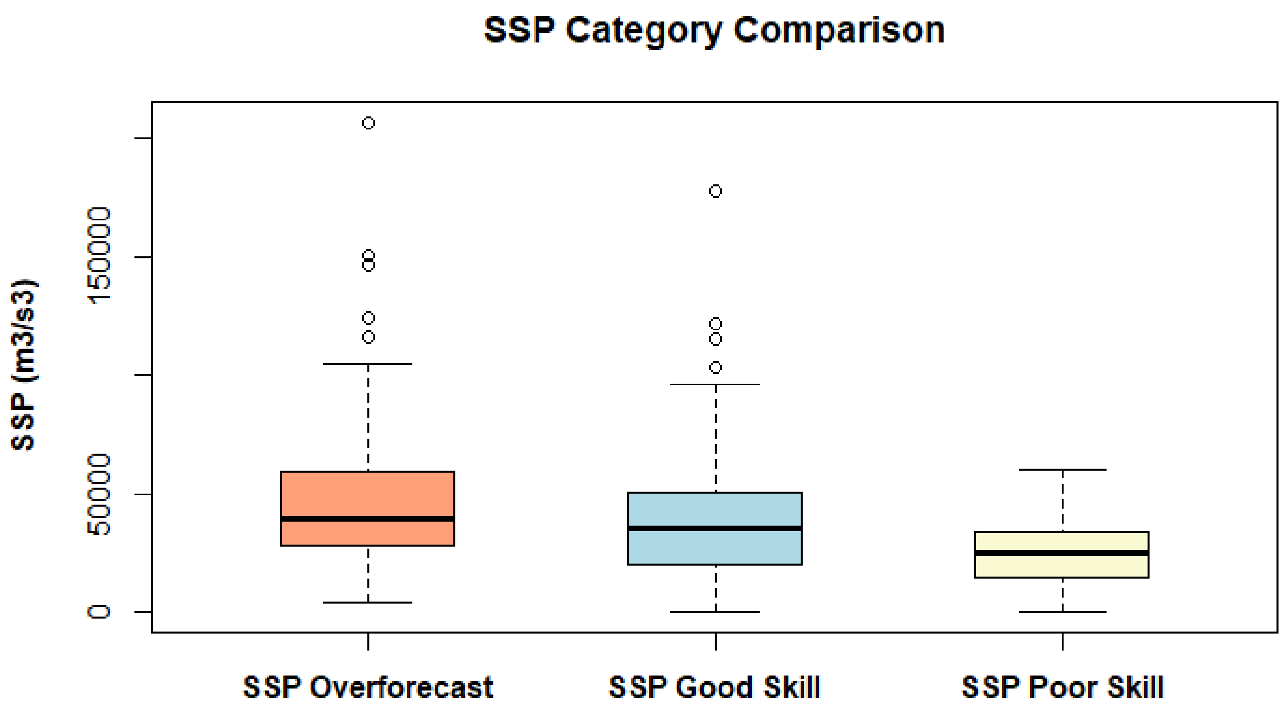

Fig. 4.9. Boxplot comparing hail overforecast (left), good skill (middle), poor skill (right) environment using the parameter SSP.

4.3 Wind

The SPC had 271 days with a PCO issued but verified with no reports. Furthermore, there were 32 days with no issued PCO but more than 10 reports (see Table 4.11). Out of the three hazards, wind had the highest count of days where HSS was below zero with 29 days. From the 12-year study period, there were 267 days had 10 or more reports that met the requirements.

There were 171 days that had good skill, and 49 days with poor skill. Overall, wind and hail had similar verification results looking at the three categories.

From 2008–2013 there were 153 days considered as an overforecast. However, from 2014–2019 there were 114 cases. In the nature of drawing the wind PCO, the outlooks tended to cover vast parts of the U.S. and verified with few reports. As a result, the PCO was evaluated as an overforecast event because of the large size. This would result in poor skill and a high bias. Overforecast wind PCOs would have two areas of interest. As in some cases, one area would have numerous wind events reported, while the other area would have few to none, leading to an overforecast. Thus, there was either a philosophy change in forecasting due to the introduction of the marginal and enhanced risk late in 2014 or it constrained forecasters, meaning they were restricted within the categorical risk. There were 58 days with good skill between 2008–2013, and 113 days between 2014–2019. Furthermore, good skill PCOs were smaller in size compared to the size of overforecast PCOs. Lastly, poor skill had the opposite 6-year splits compared to good skill PCO. There were 14 days with poor skill between 2008–2013, and 35 days between 2014–2019.

Table 4.11. Wind verification statistics.

Description	Wind
Number of days with 10 reports or more	2089
Number of days with an overforecast	267
Number of days with skilled forecasts	171
Number of days with poor skill	49
Number of days with a PCO but no reports	271
Number of days with no PCO but more than 10 reports	32
Number of days with a HSS below 0	29

4.3.1 Statistical analysis on wind parameters and the three categories

There were only two parameters, MCSM and DCP, that were statistically significant at the 5% level for all three wind categorical comparisons (see table 4.12). The two categorical comparisons that were statistically significant at 1% for MCSM were wind good skill vs overforecast and poor skill vs good skill. Despite SHR1 for good vs overforecast and poor skill vs overforecast failing at the 5% level, there was a difference in the mean variance between poor skill vs good skill with a p-value of 0.022. MCSM forecasts the probability of a mesoscale complex system (MCS). There are five parameters were used in developing the probability of an MCS event: MUCAPE, 3-12 km mean wind speed (m/s), 3-8 km lapse rate (degrees C/km), SHR1, and wind shear in the 6-10 km layer (Coniglio and Corfidi 2006). Although MUCAPE did poorly in distinguishing between all three categorical pairs, SHR1 only between poor skill and good, the other three parameters could be useful in distinguishing the environments between the three categories (see Tables 4.14, 4.15, 4.16).

Table 4.12. Wind ANOVA and Tukey test results. The p-values that were statistically significant at the 1% level were bolded, and p-values that statistically significant at the 5% level were italicized.

Parameter	Good skill-Overforecast p-value	Poor skill-Overforecast p-value	Poor skill-Good skill p-value
MCSM (%)	< **0.010**	*0.044*	< **0.001**
DCP (numeric)	< **0.001**	< **0.01**	< **0.001**
DCAPE (J/kg)	0.971	0.836	0.913
MUCAPE (J/kg)	0.491	0.933	0.559
SHR1 (m/s)	0.203	0.203	*0.022*
SHR6 (m/s)	0.185	0.852	0.256
SHR8 (m/s)	0.819	0.731	0.522

Table 4.13. Wind parameter means for overforecast, good skill, and poor skill.

Parameter	Overforecast mean	Good skill mean	Poor skill mean
MCSM (%)	88.767	93.432	83.544
DCP (numeric)	3.960	5.107	2.639
DCAPE (J/kg)	1313.103	1322.098	1348.286
MUCAPE (J/kg)	3198.149	3364.017	3115.934
SHR1 (m/s)	23.643	24.695	21.986
SHR6 (m/s)	45.438	47.688	44.344
SHR8 (m/s)	55.619	56.492	53.885

The DCP was successful at the in distinguishing between all three wind categorical pairs (see Equation 5, Table 4.12 , and 4.13). This can explicitly be seen between good skill and poor skill. Analyzing components of the DCP parameter, DCAPE and MUCAPE could not distinguish the three categorical pairs (Evans and Doswell 2001). The distribution of DCAPE was similar for all three wind category pairs. The next parameter, SHR6, had lower p-values with the comparison between good vs overforecast and poor skill vs good skill but did poorly differentiating between poor skill vs overforecast. Thus, 0-6 km mean wind could distinguish the characteristics between the three categories. Lastly, there was no variance using the SHR8 between the three pairs, with each category having similar distributions.

$$\mathrm{DCP} = \left(\frac{\mathrm{DCAPE}}{980\ \mathrm{J/kg}}\right) * \left(\frac{\mathrm{MUCAPE}}{2000\ \mathrm{J/kg}}\right) * \left(\frac{\mathrm{SHR6}}{10.289\ \mathrm{m/s}}\right) * \left(\frac{\text{0-6 km mean wind}}{8.231\ \mathrm{m/s}}\right) \tag{5}$$

Table 4.14. Overforecast wind parameter percentiles.

Parameters	25th percentile	50th percentile	75th percentile
MCSM (%)	85.260	95.310	98.420
DCP	2.205	3.410	5.115
DCAPE (J/kg)	1047.775	1249.280	1555.375
MUCAPE (J/kg)	2236.405	3177.000	4107.435
SHR1 (m/s)	19.091	22.416	27.849
SHR6 (m/s)	34.959	43.469	55.536
SHR8 (m/s)	44.802	54.457	65.613

Table 4.15. Good skill wind parameter percentiles.

Parameters	25th percentile	50th percentile	75th percentile
MCSM (%)	93.045	96.990	98.825
DCP	2.480	4.210	6.665
DCAPE (J/kg)	1053.520	1343.520	1598.645
MUCAPE (J/kg)	2086.685	3215.750	4578.000
SHR1 (m/s)	20.209	24.837	28.568
SHR6 (m/s)	39.287	46.808	56.921
SHR8 (m/s)	46.603	56.734	66.049

Table 4.16. Poor skill wind parameter percentiles.

Parameters	25th percentile	50th percentile	75th percentile
MCSM (%)	72.840	94.580	98.380
DCP	1.260	1.930	3.170
DCAPE (J/kg)	847.810	1509.590	1784.500
MUCAPE (J/kg)	1534.190	3207.880	4356.250
SHR1 (m/s)	16.318	20.278	27.149
SHR6 (m/s)	30.377	45.713	58.008
SHR8 (m/s)	35.794	56.064	69.276

4.3.2 Boxplots

Since MCSM, DCP and SHR1 showed to be useful in distinguishing between the three categories, boxplots were used to further analyze the differences. Looking at the MCSM boxplot (Fig. 4.10), the IQR for good skill is compact compared to the larger IQR for overforecast and poor skill. This distinction was particularly true when comparing the IQR between good skill and poor skill. The max, 75th percentile, and median for all three categories were identical. The difference became apparent comparing the 25th percentile of the three categories. Good skill's 25th percentile was in the 90%, while for overforecast it was in the 80%, and 70% for poor skill. Thus, good skill wind environments can be characterized by high MCSM values greater than 80%. Wind events on overforecast days occurred over a large MCSM range, with several outliers below 70%.

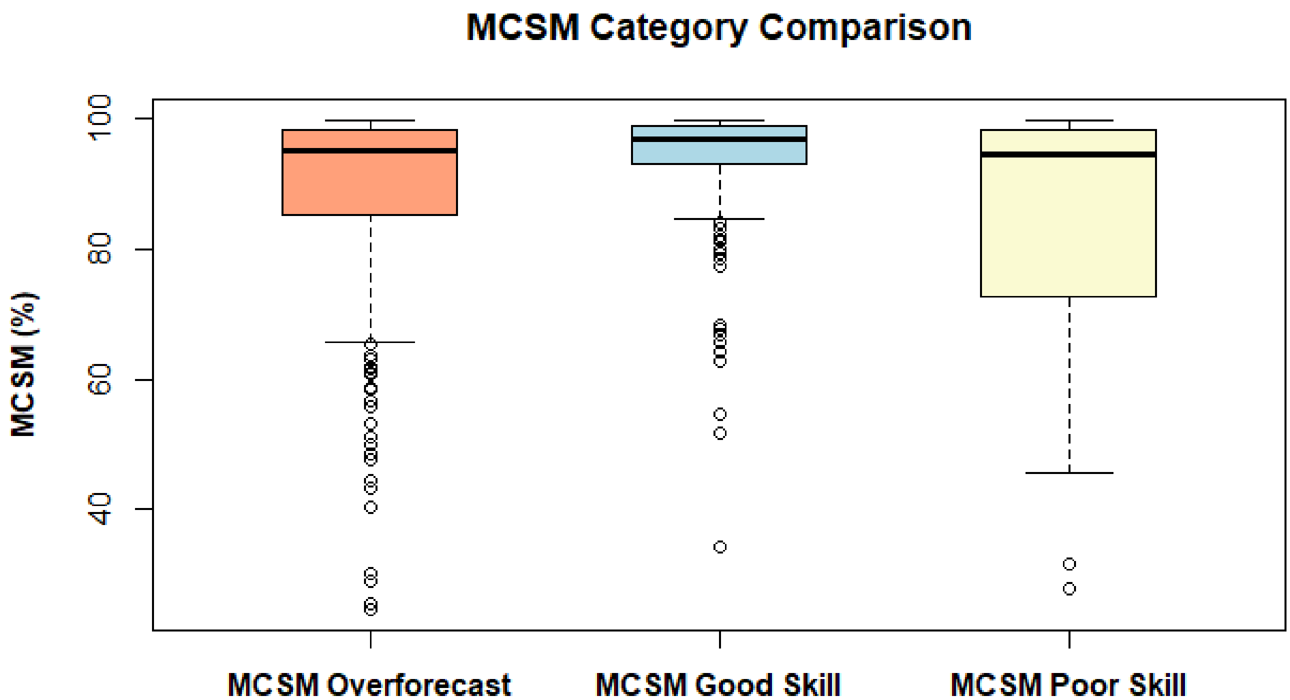

Fig. 4.10. Boxplot comparing wind overforecast (left), good skill (middle), poor skill (right) environment using the parameter MCSM.

In Fig. 4.11, the DCP median for good skill (4.210) was slightly higher compared to overforecast (3.410). The median for good skill was much greater compared to poor skill (1.930) with a difference in the median of 2.190. Comparing the three categories, wind events occurred over a larger IQR compared to overforecast, with both categories having several outliers greater than 10. Poor skill had many events having a DCP value less than 5. There was a distinct difference in the distribution between overcast and poor skill. With overforecast having moderate DCP values compared to the lower values for poor skill. Overall, good skill was characterized by higher DCP values greater than 5, overforecast between 2.205–5.115, and poor skill less than 3.170.

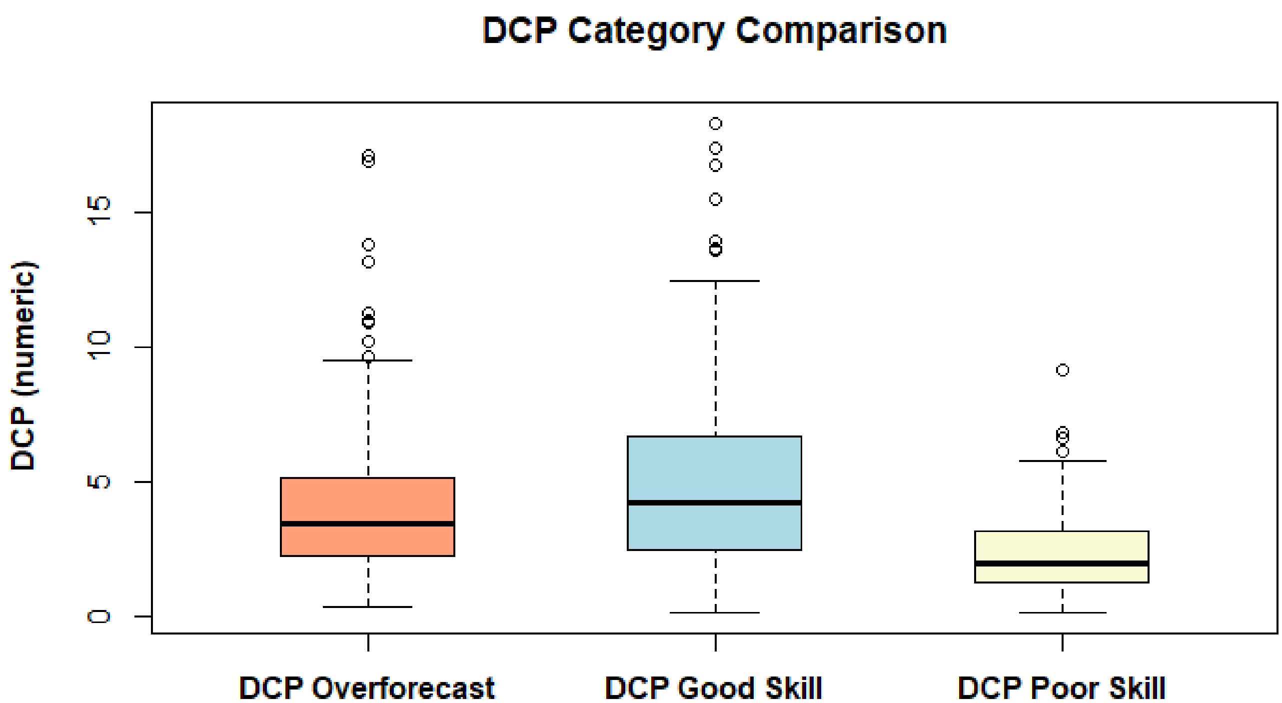

Fig. 4.11. Boxplot comparing hail overforecast (left), good skill (middle), poor skill (right) environment using the parameter DCP.

The last parameter analyzed for wind parameters was SHR1 (Fig. 4.12). Wind shear between 0-1 km was only statistically comparing poor skill and good skill. The median value for good skill was slightly higher compared to overforecast. As seen from the boxplot, there was a slight difference between overforecast and good skill. The median difference between good skill and poor skill was 4.559, with 25% of the wind events for poor skill occurring between 16.318–20.278 m/s, lower than the other two categories. The parameter distribution for overforecast and good skill was identical, with almost no median difference between overforecast and poor skill, and a distinct median difference between good skill and poor skill.

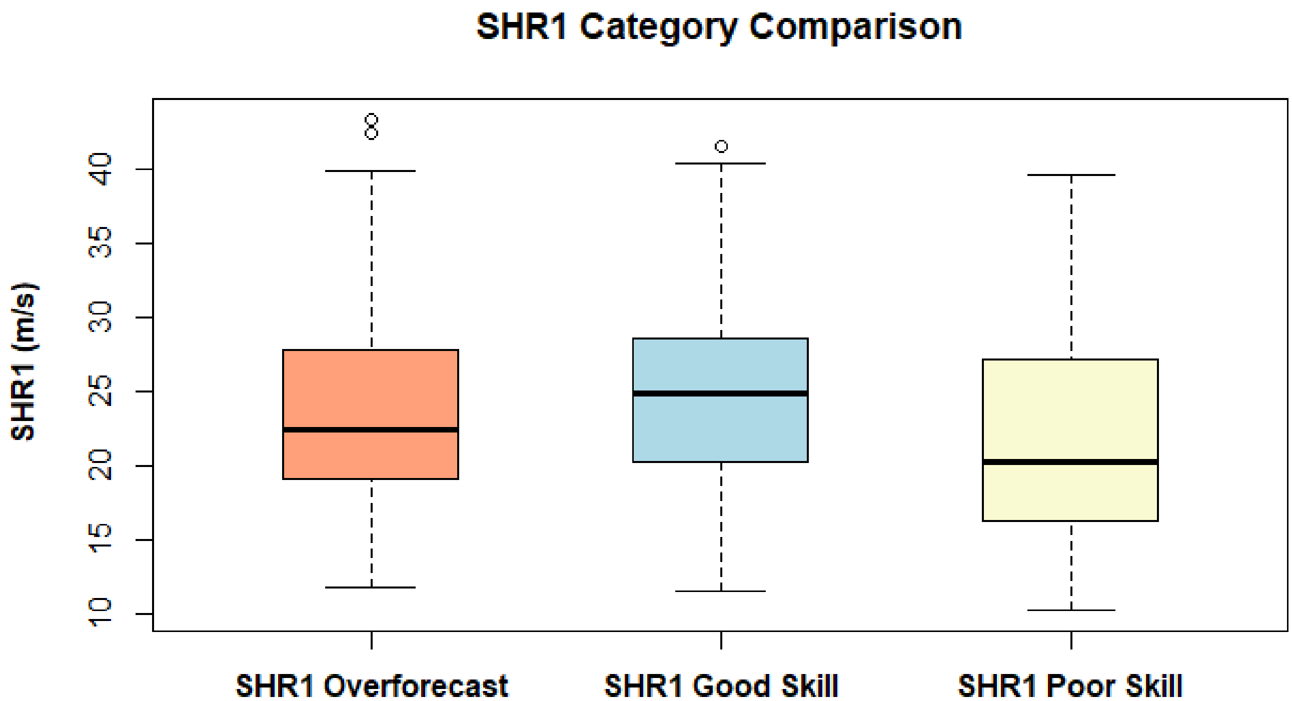

Fig. 4.12. Boxplot comparing hail overforecast (left), good skill (middle), poor skill (right) environment using the parameter SHR1.

5. Discussion and conclusion

The purpose of this study was to verify and analyze the SPC's PCO and their environmental characteristics. The SPC's Day One 1200 UTC PCO from March 26, 2008, through the year 2019 were verified using an 80.5 km grid and evaluated by a 2 by 2 contingency table. Bias and HSS were used to create three categories for each hazard using a 10-report threshold: good skill, overforecast, and poor skill. The three categories were developed using the 15th and 85th percentile of bias and HSS with respect to each hazard. To analyze the characteristics of each hazard category, SFOCA was used to collect the max value within a 11.1 by 11.1 km box created by using the first report of the day for each hazard's category. To test whether there were statistically difference or similarities between each hazard's category, the two-sample difference of means test was used for tornadoes, and the ANOVA and Tukey test were used for hail and wind. The following paragraphs go into detail about each hazard.

5.1 Tornadoes

The results show none of the shear and helicity parameters could differentiate the means between tornado good skill and overforecast. Effective SRH had the highest p-value (0.908) out of the list, and SRH1 had a lower p-value of 0.564. Shear from 0-6 km had the lowest p-value out of the shear/helicity parameters. The two parameters that make up the STPeff parameter, ESHR and ESRH, could not distinguish between tornado good skill and overforecast. Therefore, SRH1 or low-level helicity could be more useful in characterizing tornado PCO days with good skill compared to ESRH (Coffer et al. 2019). Overall, good skill and overforecast had similar environments of high deep-level shear and helicity.

The lower-level instability parameters were not able to distinguish the mean between tornado good skill and overforecast; but mixed-layer CAPE and lapse rates at the lower levels did have lower p-values compared to the shear and helicity parameters. Comparing STPeff and VGP3, STPeff was more useful in distinguishing between good skill and overforecast. Next, MLCAPE for good skill had slightly higher values compared to overforecast, but overall, the distribution was similar.

Mixed-layer LCL was the only parameter that was statistically significant at 1% and useful in distinguishing between good skill and overforecast. Previous research has found lower MLLCL heights are associated with significant or violent tornadoes (Craven and Brooks 2004; Hampshire et al. 2018). Lower MLLCL heights could influence a bias in the size of the tornado PCO and the forecasters as it would result in a numerically higher STPeff value. Good skill was characterized by higher MLLCL heights, which would lead to lower numerical STPeff values. Tornadoes that occurred in high MLLCL environments could be explained by steep LR3 lapse rates and high 0-3 km MLCAPE values (Davies 2006; Hampshire et al. 2018).

Mixed-layer CIN was close to being statistically significant at 5% and showed some ability to distinguish between overforecast and good skill. Good skill tornado events had lower MLCIN values compared to overforecast. With MLCIN being a component of the STPeff composite parameter, lower MLCIN values would result in a higher STPeff value. Higher MLCIN values could suppress convection, reduce the number of convective storms, resulting in a large tornado PCO with no observed reports. Lastly, the STPeff was useful in differentiating between good skill and overforecast. Good skill was characterized by higher values compared to the lower distribution of STPeff values for overforecast. The differences in the categorical STPeff distribution can be attributed to MLCAPE, MLCIN, and MLLCL values in both categories.

In conclusion, MLLCL can best distinguish between overforecast and good skill. Mixed-layer CIN and STPeff could both differentiate between overforecast and good skill. Despite

MLCAPE not being statistically significant at any level, good skill had slightly higher values compared to overforecast. More data samples can further test whether MLCAPE is useful in distinguishing the environments of overforecast and good skill. The same can also be said for MLCAPE3 and LR3.

5.2 Hail

There were six parameters that were skillful in in distinguishing between one or more categorical pairs. All three categories had similar LR75 distributions, particularly between hail good skill vs overforecast. The three thermodynamics could differentiate between the datasets of good skill vs overforecast and poor skill vs overforecast at the 1% level. However, the parameters could not distinguish between poor skill vs good skill. Although, MLCAPE was close to being statistically significant at the 5% level. Thus, the three thermodynamic parameters all had similar data distribution between good skill and poor skill. Overall, there were only two parameters that could distinguish between poor skill and good skill, PWAT and SSP.

Overforecast could be characterized by higher MUCAPE, FMUCAPE, and MLCAPE values compared to good skill and poor skill. Higher CAPE values indicate an overforecast bias. Furthermore, MUCAPE for hail was more useful in distinguishing between the categories compared to Wind. The higher MUCAPE, FMUCAPE, and MLCAPE values correspond to a larger hail PCO. Forecasters use CAPE to quickly approximate the strength of the updraft in a storm. Higher CAPE values can translate to a stronger updraft in a storm, and the stronger updraft would support hail growth. On the other hand, the results show that higher CAPE values could influence a bias in large PCO areas for hail. Further analyzing the differences, 50% of hail events for good skill occurred within a range of 1860.380–3719.5 J/kg MUCAPE, 648.72–1009.12 J/kg FMUCAPE, and 441.699–1245.917 J/kg MLCAPE.

In addition, deep-level shear could only distinguish between one category pair, good skill vs overforecast. Good skill was characterized by higher SHR6 distribution values compared to overforecast. The higher shear values support the updraft of a storm and aid in preventing the downdraft from overtaking the updraft. Deep-level shear is a key factor in the development of supercell thunderstorms. In addition, hail events are predominantly produced by supercell thunderstorms, and have a significant role in hail climatology (Rasmussen and Blanchard 1998; Thompson et al. 2003; Schneider and Dean 2008). Next, the SSP was the only parameter that could successfully distinguish between all three category pairs. Overforecast had higher SSP values compared to good skill, and SSP for good skill was higher compared to poor skill. The components of the SSP parameter, MLCAPE and SHR6, show why. Higher SSP values are attributed to moderate SHR6 and high MLCAPE values. Good skill was characterized by low MLCAPE but higher SHR6, resulting in good skill having lower SSP values than overforecast. Hail events that occurred on poor skill had deep-level shear with a large IQR, but MLCAPE values were much lower compared to the other two categories.

Precipitable water has rarely been discussed in the literature of hail environmental characteristics. Therefore, PWAT being statistically significant at 5% for all three categorical pairs was an interesting finding. Good skill was characterized by drier environments compared to overforecast that had higher PWAT amounts. The drier environment associated with good skill could be produced by low precipitation (LP) or classic supercells. Higher PWAT amounts would reduce the strength or completely eradicate the storm's updraft, lowering the probability of a hail event. This can explain why overforecast had many hail events with high PWAT amounts, and overforecast hail events can be associated with high precipitation supercells.

5.3 Wind

The two thermodynamic parameters, DCAPE and MUCAPE, could not distinguish between the three categories. The three categories' DCAPE and MUCAPE distribution was uniform, as each category pair had a p-value greater than 0.850. The three wind shear parameters were not useful in distinguishing between all three categories. However, SHR1 could distinguish between poor skill vs good skill with a p-value of 0.022. Good skill wind events compared occurred in higher SHR1 environments compared to the other two categories, but the good skill SHR1 distribution was similar to overforecast. In addition, the deep-level shear distribution between poor skill and overforecast was similar. Lastly, SHR8 was the least skilled in distinguishing between the three categorical pairs compared to the other two wind shear parameters.

There were only two parameters that could distinguish between all three categorical pairs at the 5% level. The derecho composite parameter was successful at distinguishing between all three category pairs being statistically significant at 1%. Good skill was characterized by moderate MUCAPE, SHR6, and DCAPE attributing to the higher DCP values compared to overforecast and poor skill. Since DCAPE, MUCAPE, and SHR6 were not able to differentiate between the three categorical pairs, 0–6 km mean wind and other variations of mean wind in different levels of the atmosphere could aid in distinguishing between the three categories. As the two thermodynamics and deep-level shear parameters for poor skill were similar to the other two categories, poor skill by default can be characterized by low 0–6 km mean wind and resulting in low DCP values.

The second parameter, MCSM, was skillful in distinguishing between all three categories being statistically significant at 5%. Good skill wind events occurred in high MCSM percentages, with 75% of events occurring when MCSM was greater than 93.045 %. Analyzing the parameters that were developed in making the MCSM parameter for good skill, SHR1 and MUCAPE were not useful. Thus, wind shear in the 6–10 km layer, 3–8 km lapse rate, and 3–12 mean wind speed parameters should be used to see if they could distinguish between the three categories.

There were multiple similarities between the three categories' environmental characteristics and the fewest parameters to distinguish the category's environments. Unlike tornadoes and hail, wind events can occur in multiple environments from various convection storm modes and synoptic conditions (Evans and Doswell 2001). A damaging wind event can occur from an RM supercell, microbursts, QCLS, derecho, mesoscale complex system, or a mesoscale convective complex. Linear convection like QCLS is the typical storm mode for prolific wind events, but not one single parameter or combination cannot completely distinguish between a convective environment that will produce wind or hail/tornado event ahead of time (Thompson et al. 2012; A. R. Dean 2021, personal communication). From the analysis of wind environmental characteristics, this is true.

5.4 Parameter implications

This section goes into detail about the how the parameters can be useful when forecasting an outlook. There are six parameters that could be useful when forecasting and placement of a tornado PCO. Hail had six parameters that forecasters could use when issuing a hail PCO. There only two parameters useful when forecasting a wind PCO.

5.4.1 Tornado

The five parameters that should slightly used cautiously in forecasting a tornado PCO are ESRH, ESHR, VGP3, SHR6, and SRH1. These parameters had no skill in distinguishing the mean between tornado good skill and overforecast. Environments with STPeff values greater than or equal to 3.026 are associated with forecasts with good skill. When forecasting the location of an outlook, forecasters could focus in on areas where the STPeff value is greater than 3. In tandem with the STPeff values, high MLCAPE (MLCAPE $\geq$ 2424.882 J/kg) and steep LR3 (LR3 $\geq$ 9.321 C/km) environments can aid forecasters. Mixed-layer CIN had some skill in differentiating between good skill and overforecast. Environments with lower MLCIN values, greater than -28.570 J/kg, could also be useful. Higher MLLCL heights were affiliated with forecasts with good skill. Environments with higher heights accompanied by moderate MLCAPE3 and steep LR3 can be beneficial when locating a tornado PCO and for low-level tornadogenesis (Davies 2006; Hampshire et al. 2018).

5.4.2 Hail

The lapse rate between 700-500 mb was not useful in differentiating between the three hail categories. When issuing a hail outlook, environments with MUCAPE values (MUCAPE $\geq$ 3681.565 J/kg) were linked to forecasts with an overforecast bias. This was also the same case for FMUCAPE $\geq$ 961.410 J/kg and MLCAPE $\geq$ 1104.589. Therefore, when issuing a hail PCO and predicting organized convection, forecasters should beware of environments with high CAPE values. In addition, areas with high PWAT amounts (PWAT $\geq$ 1.950 inches) could decrease the probability of hail growth by wiping out the storm's updraft since higher PWAT values were linked with an overforecast bias. In theory, the high SSP environments would be paired with convection producing severe and significant hail events; however, the higher values could lead to less skill when forecasting (Craven and Brooks 2004).

Therefore, forecasters should focus on areas with drier PWAT environments (PWAT $\leqq$ 1.820 inches). Environments with lower PWAT amounts suggest it is easier for hail growth and the updraft to persist in convection. The results indicate organized convection that produce a hail event are associated with substantially lower MUCAPE, MLCAPE, and FMUCAPE values compared to overforecast (Schneider and Dean 2008). Hail forecasts with good skill were associated when MUCAPE was between 1860.380–3719.5 J/kg, FMUCAPE 648.72–1009.12 J/kg FMUCAPE, and MLCAPE 441.699–1245.917 J/kg MLCAPE. Deep-level shear is a common parameter forecasting for organized convection, and high SHR values (SHR6 $\geq$ 45.452 m/s) environments were associated with good skill. Additionally, strong deep-level shear and modest MLCAPE values were skillful in characterizing hail forecasts with good skill. Thus, forecasters can use moderate SSP values when evaluating the location of a hail PCO.

5.4.3 Wind

Five out of the seven parameters for wind had similar distributions and means and were not useful when forecasting a wind PCO. Results from the study found DCP and MCSM values were lower for both wind overforecast and poor skill when compared to the good skill category. Environments with high DCP values (DCP $\geq$ 4.210) were associated with forecasts with good skill. The DCP could also be a useful parameter to forecast wind events from various storm modes like QCLS or an MCS. In addition to environments with high DCP values, forecasters can use high MCSM percentages (MCSM $\geq$ 93.432 %) when issuing the wind PCO.

5.5 SPC verification

On examination of the verification results for each severe weather hazard, there were interesting outcomes. The SPC when forecasting for tornadoes tended to be over-cautious compared to hail and wind. There were 664 days when a tornado PCO was issued but verified with no reports. This is of no surprise, as tornadoes cause more damage to property and have higher fatalities compared to the other two hazards. Analyzing the two time periods, the SPC had 381 days of when the SPC issued a tornado PCO but verified with no reports between 2008–2013, and from 2014 through 2019 there were 283 days. Next, there was not a vast difference in the splits for good skill and overforecast. There were 29 tornado overforecast days from 2008–2013 and 24 days between 2014–2019, and a slight decrease in the number of days. This was also the same result for analyzing the two time periods for the tornado good skill category. There were 18 days that verified as good skill in the first time period and 24 days in the second.

Evaluating the hail PCO, there were notable results. The SPC had fewer days, 209 for hail compared to 664 days for tornadoes, for when the SPC issued a hail PCO but verified with no reports. Looking into the two 6-year splits from 2008–2013 there were 112 cases, and from 2014–2019 there were 97. Furthermore, the lopsided six-year splits for hail overforecasts and good skill indicate a change in forecasting ideology. As from 2008–2013, there were 176 days considered as an overforecast for hail, and from 2014–2019 there were 94 cases. There were 65 days with good skill between 2008–2013, and 100 days between 2014–2019. Between 2008–2013, overforecast nearly tripled the amount of good skill forecast days. But between 2014–2019, their numbers were the same, with good skill having six more days.

The SPC had 271 days with a wind PCO issued but verified with no reports. Studying the two time periods, the SPC had 153 days of when a wind PCO was issued but verified with no reports between 2008–2013, and from 2014 through 2019 there were 118 days. Furthermore, from 2008–2013 there were 153 days considered as an overforecast, and from 2014–2019 there were 114 cases. There were 58 days with good skill between 2008–2013, and 113 days between 2014–2019. The number of good skill days nearly doubled in the second 6-year split.

During the 2008–2013 period, the SPC would periodically use the "see text" option to detail the threat of severe weather. In the same period, the SPC only had three threat risks: slight, moderate, and high. The slight risks could represent two different threat levels of a hazard: a low threat or a high threat. The "see text" option was used in the SPCs forecast to draw attention to certain areas where a threat of severe weather exists but did not warrant a slight risk to be issued (Hitchens et al. 2013). The lowest probability contour for wind, hail, and tornado was forecast in the vicinity of where the "see text" was placed.

Further analyzing the forecasts in the first 6-year split (2008–2013), the hail and wind outlooks tended to cover vast parts of the U.S. and typically verified with few reports. As a result, the PCO was evaluated as an overforecast event because of its large-sized area and this would result in poor skill. In some cases, overforecast wind and hail PCOs would have two areas of interest. One area would have numerous wind events reported, while the other area would have few or no reports leading to an overforecast. In late 2014, the SPC introduced the marginal and enhanced risks. The purpose of the risks was that the marginal risk would replace the "see text" option, and the enhanced/marginal would represent the low and high end of a slight risk.

It was found the probability contours (from 2008–2013) for each severe weather hazard were not restricted within the categorical slight risk but within the SPC's thunderstorm polygon (see example in Fig. 5.1a, 5.1b, 5.1c, and 5.1d). On many days, if the thunderstorm polygon covered extensive parts of the U.S., the probability polygons area for a hail and wind event would also be large and stretched out. After the addition of the two categorical risks, the lowest probability polygon for all severe weather hazards seemed to be restricted within the marginal risk. This can possibly explain why there was a large difference in the two 6-year splits for the following hazards. Hail having three times more overforecast days between 2008–2013, but then in 2014–2019 there six more days with good skill compared to overforecast. Next, the number of wind PCOs days with good skill were nearly doubling from 58 days between 2008–2013 to 113 days between 2014–2019. For tornadoes, the introduction of the risks could have aided in the large drop of days for when a tornado PCO was issued but verified with no reports, with 381 days between 2008–2013 and 283 days from 2014–2019. Thus, the outcome of the SPC verification results raises a question to be further researched. Did the introduction of the marginal and enhanced risk improve forecast skill for each severe weather hazard?

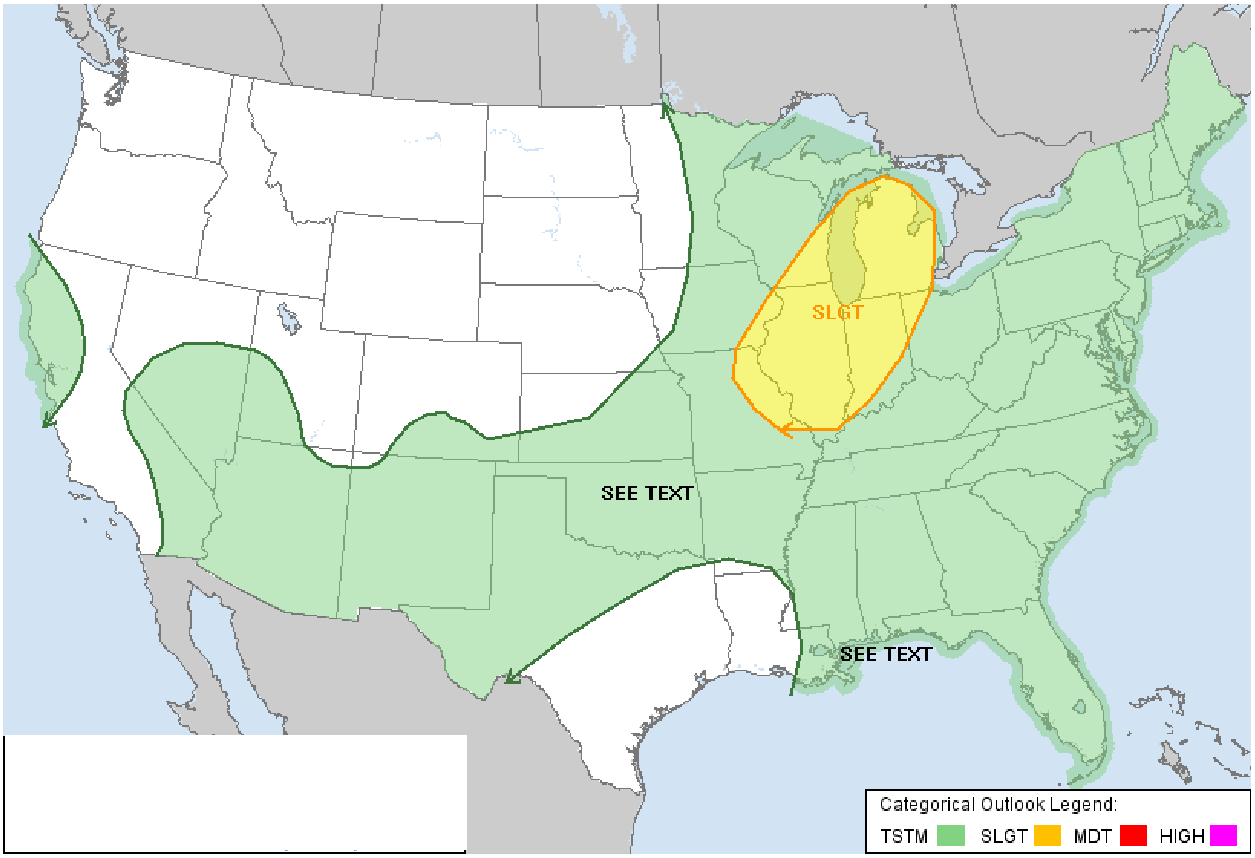

Fig. 5.1a. Image of the SPC's Day One 12Z categorical outlook on September 5th, 2012. This image was gathered and https://www.spc.noaa.gov/products/outlook/archive/2012/day1otlk_20120905_1200.html.

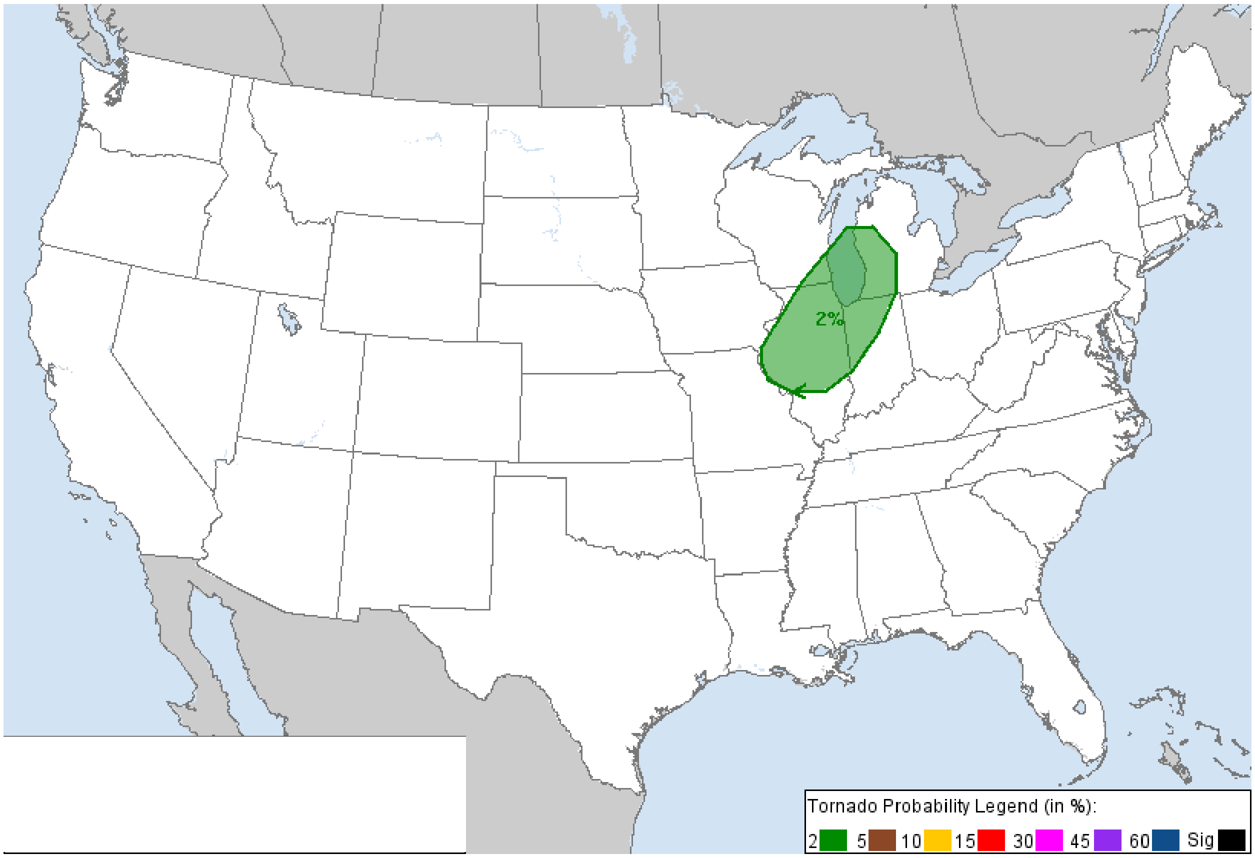

Fig. 5.1b. Image of the SPC's Day One tornado PCO on September 5th, 2012. This image was gathered and https://www.spc.noaa.gov/products/outlook/archive/2012/day1otlk_20120905_1200.html.

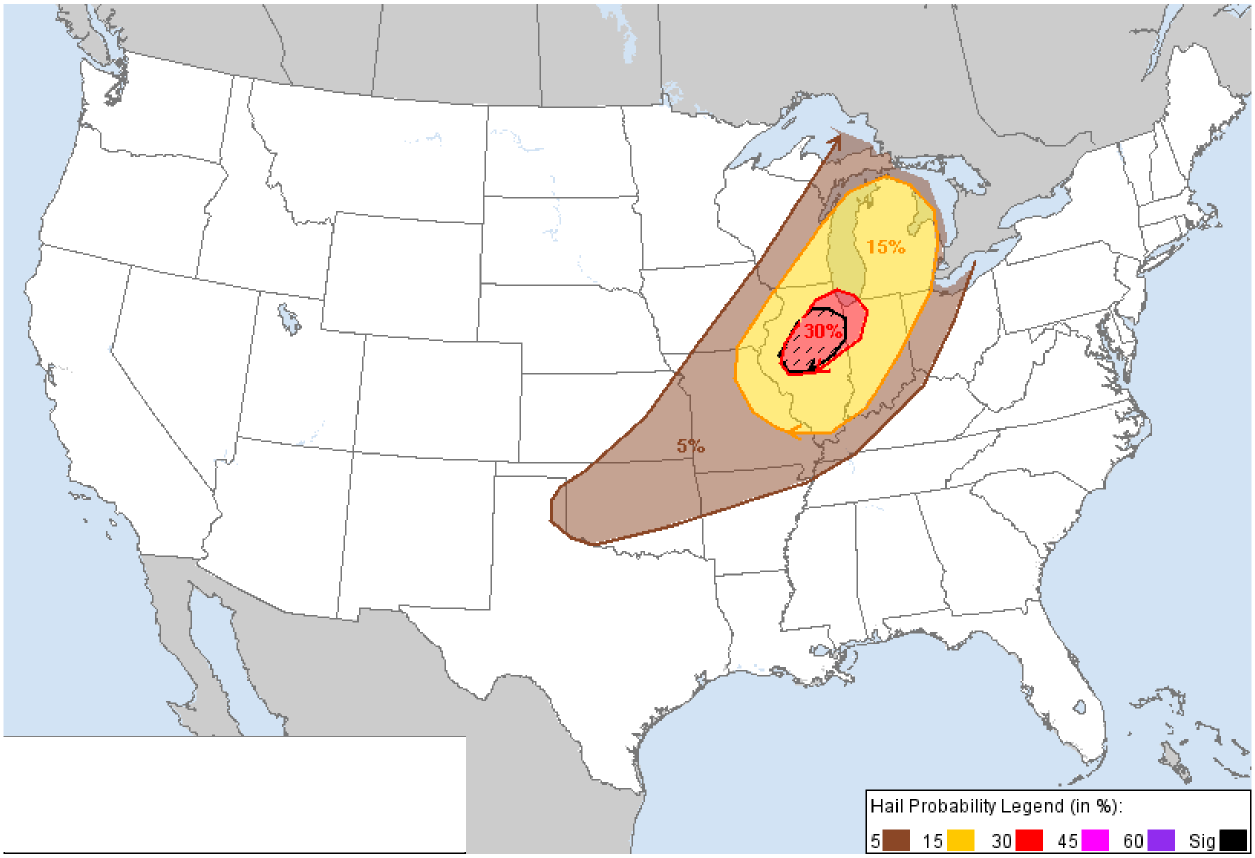

Fig. 5.1c. Image of the SPC's Day One hail PCO on September 5th, 2012. This image was gathered and https://www.spc.noaa.gov/products/outlook/archive/2012/day1otlk_20120905_1200.html.

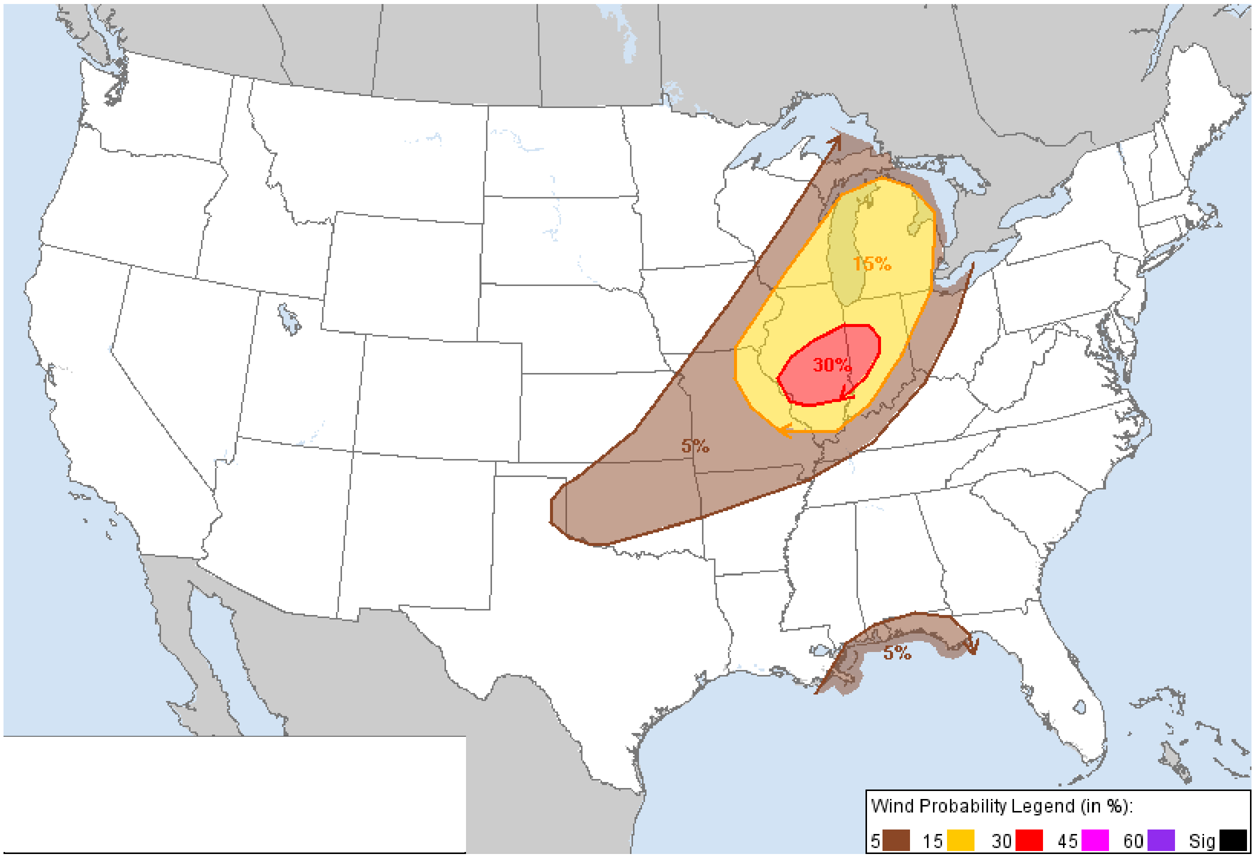

Fig. 5.1d. Image of the SPC's Day One wind PCO on September 5th, 2012. This image was gathered and https://www.spc.noaa.gov/products/outlook/archive/2012/day1otlk_20120905_1200.html.

5.6 Limitations in the research

While the results can provide useful insights in the environmental characteristic of forecast, there were limitations found during the conduction of this research project. The first limitation was focusing on individual days and parameters. Using the 15th and 85th percentiles allowed the study to focus on the extreme forecast that possess skill or no skill. However, this limits the number of days used and specific days were picked throughout the study period. The parameters picked in the study were based on frequency in forecasting, and their skill in forecasting documented in the literature. Using only a single parameter could constrict the analysis of the environment. Also, a short list of parameters for each hazard would also limit the evaluation of the environmental characteristics. The second limitation found was using the environmental parameter value associated the first report of day and using the man value only. Using only one report does accurately portray the total environmental characteristics of a forecast. The archived reanalysis data from SFCOA uses a 40-km resolution, and this is coarse when looking mesoscale events. So, the max value of the first report could be collected from a different airmass.

5.7 Future work

The expansion of the dataset would better produce conclusive results, especially for the tornado dataset. Having more data could aid distinguishing between environmental characteristics of the hazard's categories and similarities or differences. First, this could be done by lowering the report threshold to five or creating the categories using either the 20/80 or 25/75 percentiles. This could also aid in answering whether the introduction of the marginal and enhanced risks improved forecast skill for each severe weather hazard. Second, using all reports per day to calculate the daily parameter mean, or using all the reports within the dataset could aid in analyzing the total convection environment. Furthermore, a larger set of parameters for all three hazards could aid in distinguishing between the environments and would lead to clearer results. Expanding the parameter dataset would include synoptic or upper-air parameters such as upper-level jet or frontogenesis. Additionally, adding composite parameters such as the significant hail parameter, SHERBE, and VTP could also be added. The methods applied in this in the research can be used for future projects. This paper was a steppingstone to further analyze severe weather convection. An example of future study would be analyzing forecast skill and bias based on storm mode, like the study Thompson et al. (2012) conducted. This could lead to a better understanding between SPC forecast skill/bias and storm modes, and between forecasts and hazard environmental characteristics. This specifically could be useful to further understand the environmental characteristics of damaging wind events. Lastly, future work will partake in analyzing significant severe weather hazards environments and forecast skill/bias.

6. References

Anderson-Frey, A. K., Y. P. Richardson, A. R. Dean, R. L. Thompson, and B. T. Smith, 2016: Investigation of near-storm environments for tornado events and warnings. Wea. Forecasting, **31**, 1771–1790, doi:10.1175/WAF-D-16-0046.1.

——, ——, ——, ——, and ——, 2019: Characteristics of tornado events and warnings in the southeastern United States. Wea. Forecasting, **34**, 1017–1034, doi:10.1175/WAF-D-18-0211.1.

Bothwell, P. D., J. A. Hart, and R. L. Thompson, 2002: An integrated three dimensional objective analysis scheme in use at the Storm Prediction Center. Preprints, 21st Conf. on Severe Local Storms/19th Conf. on Weather Analysis and Forecasting/15th Conf. on Numerical Weather Prediction, San Antonio, TX, Amer. Meteor. Soc., JP3.1. [Available online at https://ams.confex.com/ams/pdfpapers/47482.pdf.]

Bunkers, M. J., J. S. Johnson, L. J. Czepyha, J. M. Grzywacz, B. A. Klimowski, and M. R. Hjelmfelt, 2006: An observational examination of long-lived supercells. Part II: Environmental conditions and forecasting. Wea. Forecasting, **21**, 689–714, doi:10.1175/WAF952.1.

Coffer, B. E., M. D. Parker, R. L. Thompson, B. T. Smith, and R. E. Jewell, 2019: Using near-ground storm relative helicity in supercell tornado forecasting. Wea. Forecasting, **34**, 1417–1435, doi:10.1175/WAF-D-19-0115.1.

Coniglio, M.C., and S.F. Corfidi, 2006: Forecasting the speed and longevity of severe mesoscale convective systems. Preprints, Symp. on the Challenges of Severe Convective Storms, Atlanta, GA, Amer. Meteor. Soc., [Available online at https://ams.confex.com/ams/Annual2006/techprogram/paper_104815.htm].

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.

Craven, J. P., and H. E. Brooks, 2004: Baseline climatology of sounding derived parameters associated with deep moist convection. Natl. Wea. Dig., **28**, 13–24. [Available online at https://www.nssl.noaa.gov/users/brooks/public_html/papers/cravenbrooksnwa.pdf].

Davies, J M., 2006: Tornadoes in environments with helicity and/or high LCL heights. Wea. Forecasting, 21, 579–594, doi:10.1175/WAF928.1.

Dean, A. R., and R. S. Schneider, 2008: Forecast challenges at the NWS Storm Prediction Center relating to the frequency of favorable severe storm environments. Preprints, 24th Conf. on Severe Local Storms, Savannah, GA, Amer. Meteor. Soc., 9A.2. [Available online at https://ams.confex.com/ams/pdfpapers/141743.pdf.]

——, 2012: An examination of tornado environments, events, and impacts from 2003–2012. Preprints, 26th Conf. on Severe Local Storms, Nashville, TN, Amer. Meteor. Soc., 60. [Available online at https://ams.confex.com/ams/26SLS/webprogram/Paper211580.html.]

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.

——, and D. M. Schultz, 2006: On the use of indices and parameters in forecasting severe storms. Electronic J. Severe Storms Meteor., **1**, 1–22. [Available online at https://ejssm.org/ojs/index.php/ejssm/article/viewFile/11/10.]

Evans, J. S., and C. A. Doswell, 2001: Examination of derecho environments using proximity soundings. Wea. Forecasting, **16**, 329–342, doi:10.1175/1520-0434(2001)016<0329:EODEUP>2.0.CO;2.

Grams, J. S., R. L. Thompson, D. V. Snively, J. A. Prentice, G. M. Hodges, and L. J. Reames, 2012: A Climatology and comparison of parameters for significant tornado events in the United States. Wea. Forecasting, **27**, 106–123, doi:10.1175/WAF-D-11-00008.1.

Guyer, J. L., and A. R. Dean, 2010: Tornadoes within weak CAPE environments across the continental United States. Preprints, 25th Conf. on Severe Local Storms, Denver, CO, Amer. Meteor. Soc., 1.5. [Available online at https://ams.confex.com/ams/pdfpapers/175725.pdf.]

Hampshire, N. L., R. M. Mosier, T.M. Ryan, and D.E. Cavanaugh, 2018: Relationship of low-level instability and tornado damage rating based on observed soundings. J. Operational Meteor., **6**, 1-12, doi:10.15191/nwajom.2018.0601.

Herman, G. R., E. R. Nielsen, and 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.

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 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.

——, 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 ——, 2017: Determining criteria for missed events to evaluate significant severe convective outlooks. Wea. Forecasting, **32**, 1321–1328, doi:10.1175/WAF-D-16-0170.1.

Johnson, A. W. and K. E. Sugden, 2014: Evaluation of sounding-derived thermodynamic and wind-related parameters associated with large hail events. Electronic J. Severe Storms Meteor., **9**. [Available online at https://ejssm.org/ojs/index.php/ejssm/article/viewArticle/137].

Kay, M. P., and H. E. Brooks, 2000: Verification of probabilistic severe storm forecasts at the SPC. Preprints, 20th Conf. on Severe Local Storms, Orlando, FL, Amer. Meteor. Soc, 9.3.

Kuchera, E. L., and M. D. Parker, 2006: Severe convective wind environments. Wea. Forecasting, **21**, 595–612, doi:10.1175/WAF931.1.

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%3C0281:WIAGFA%3E2.0.CO;2.

Rasmussen, E. N., and D. O. Blanchard, 1998: A baseline climatology of sounding-derived supercell and tornado forecast parameters. Wea. Forecasting, **13**, 1148–1164, doi:10.1175/1520-0434(1998)013<1148:ABCOSD>2.0.CO;2.

——, 2003: Refined supercell and tornado forecast parameters. Wea. Forecasting, **18**, 530–535, doi:10.1175/1520-0434(2003)18<530:RSATFP>2.0.CO;2.

Schneider, R. S., A. R. Dean, S. J. Weiss, and P. D. Bothwell, 2006: Analysis of estimated environments for 2004 and 2005 severe convective storm reports. Preprints, 23rd Conf. on Severe Local Storms, St. Louis, MO, Amer. Meteor. Soc., 3.5. [Available online at https://ams.confex.com/ams/pdfpapers/115246.pdf.]

——, and ——, 2008: A comprehensive 5-year severe storm environment climatology for the continental United States. Preprints, 24th Conf. on Severe Local Storms, Savannah, GA, Amer. Meteor. Soc., 16A.4. [Available online at https://ams.confex.com/ams/pdfpapers/141748.pdf.]

Sherburn, K. D., and M. D. Parker, 2014: Climatology and ingredients of significant severe convection in high-shear, low-CAPE environments. Wea. Forecasting, **29**, 854–877, doi:10.1175/WAF-D-13-00041.1.

——, ——, J. R. King, and G. M. Lackmann, 2016: composite environments of severe and nonsevere high-shear, low-CAPE convective events. Wea. Forecasting, **31**, 1899–1927, doi:10.1175/WAF-D-16-0086.1.

Smith, B. T., R. L. Thompson, J. S. Grams, C. Broyles, and H. E. Brooks, 2012: convective modes for significant severe thunderstorms in the contiguous United States. Part I: Storm classification and climatology. Wea. Forecasting, **27**, 1114–1135, doi:10.1175/WAF-D-11-00115.1.

Thompson, R. L., R. Edwards, J. A. Hart, K. L. Elmore, and P. Markowski, 2003: Close proximity soundings within supercell environments obtained from the Rapid Update Cycle. Wea. Forecasting, **18**, 1243–1261, doi:10.1175/1520-0434(2003)018<1243:CPSWSE>2.0.CO;2.

——, C. M. Mead, and R. Edwards, 2007: Effective storm-relative helicity and bulk shear in supercell thunderstorm environments. Wea. Forecasting, **22**, 102–115, doi:10.1175/WAF969.1.

——, B. T. Smith, J. S. Grams, A. R. Dean, and C. Broyles, 2012: Convective modes for significant severe thunderstorms in the contiguous United States. Part II: Supercell and QLCS tornado environments. Wea. Forecasting, **27**, 1136–1154, doi:10.1175/WAF-D-11-00116.1.

Wilks, D. S., 2011: Statistical Methods in the Atmospheric Sciences. 3rd ed. Elsevier, 676 pp.