UNIVERSITY OF OKLAHOMA

GRADUATE COLLEGE

DERIVING OPERATIONALLY RELEVANT TORNADO PROBABILITIES FROM CONVECTION-ALLOWING ENSEMBLES

A DISSERTATION

SUBMITTED TO THE GRADUATE FACULTY

in partial fulfillment of the requirements for the

Degree of

DOCTOR OF PHILOSOPHY

By

BURKELY LYNN GALLO
Norman, Oklahoma
2017

DERIVING OPERATIONALLY RELEVANT TORNADO PROBABILITIES FROM CONVECTION-ALLOWING ENSEMBLES

A DISSERTATION APPROVED FOR THE
SCHOOL OF METEOROLOGY

BY

_____
Dr. Adam Clark, Co-Chair

_____
Dr. Xuguang Wang, Co-Chair

_____
Dr. Harold Brooks

_____
Dr. Michael Richman

_____
Dr. Steven Cavallo

_____
Dr. Deborah Trytten

Acknowledgements

First, I would like to thank my advisor, Dr. Adam Clark, for all of the opportunities he's given me as a part of this project. Through his mentoring, I have learned how to be a better scientist, communicate more effectively (and occasionally more concisely!), and bring together a huge variety of people within the field of meteorology. I also owe great thanks to Dr. Harold Brooks, who has provided much fuel for thought on what constitutes a good and useful forecast. Through my conversations with him I have gained a better understanding of how weather and people intersect, and I believe it has made my work far more applicable than it would otherwise be. I would also like to thank the rest of my committee: Dr. Xuguang Wang, Dr. Michael Richman, Dr. Steven Cavallo, and Dr. Deborah Trytten. Their open-door policies always made me feel welcome whenever I needed to untangle a particularly knotty problem. Finally, thanks go to Christie Upchurch for helping me manage the ins and outs of the degree process, and for making life much easier through her efforts and advice.

Funding for this work came partially from a NSF Graduate Research Fellowship under Grant No. DGE-1102691, Project #A00-4125. Funding also was provided by NOAA/Office of Oceanic and Atmospheric Research under NOAA-University of Oklahoma Cooperative Agreement #NA16OAR4320115, U.S. Department of Commerce.

Throughout this research, I have been able to work with top-notch forecasters and science support branch scientists at the Storm Prediction Center, which helped me maintain an operational perspective and develop tools to help operational forecasters in

their mission to protect lives and property. Special thanks go to Bryan Smith and Rich Thompson, for being eager to brainstorm new ways to blend observations and numerical weather prediction and for teaching me what forecasters look for in forecasting severe convective weather. Their feedback and input helped shape the trajectory of this work. In addition, I would like to thank Dr. Israel Jirak, for his support through this project, providing advocacy for this work and supporting the R2O–O2R framework. Also, thanks go to Andy Dean and Dr. Chris Melick for their wizardry in supplying data.

This work was inspired by, and often took place within, the Spring Forecasting Experiment (SFE) in NOAA's Hazardous Weather Testbed. I would like to thank the people who have made the SFEs possible, including the original designers of the SFEs. I would also like to thank the participants in SFEs 2015, 2016, and 2017, who provided feedback on the work herein and helped to improve it. It has been an honor to be able to talk to forecasters, model developers, and researchers from around the world through the SFEs, and the five weeks of the experiment remain the most exciting and intellectually energizing out of each year. I have also been able to learn about all aspects of severe convective forecasting during the SFEs, by watching master forecasters at work. Thanks go to Dave Imy and Dr. Greg Carbin for teaching me so much about how the atmosphere works on a day-to-day basis. Finally, thanks also go to Kent Knopfmeier, Robert Hepper, and Jack Kain, each of whom were instrumental in the SFEs.

I would also not have made it to this point without the support of my Penn State cohort, who have become my sounding board for all things professional as well as some

of my best friends. Alicia Klees, Dr. Alex Anderson-Frey, Dr. Livia Souza Friere Grion, and Scott Sieron, thank you for all of your support and proofreading, and for motivating me to become a better scientist and student of life. Your dedication to the science of meteorology and to helping your peers and students succeed has been a source of inspiration to me for years.

To my University of Oklahoma cohort, thank you for taking me in as one of your own when I was new to the university and the state – you made me feel welcome in an overwhelmingly large place! Going through graduate school with Dr. Katie Wilson, Veronica Fall, Liz Digangi and Hristo Chipilski kept me motivated and sane from the start of classwork to the end of my dissertation writing process. Thanks also go to Greg Blumberg, Ryan Lagerquist, Matt Flournoy, Kenzie Krocak, and many others for your friendship.

I have also been fortunate to have many non-meteorology or non-academia friends to keep me grounded in "the real world" and remind me that there is so much to enjoy outside the pursuits within this dissertation. To Clare, Laura, Rachel, Dan, Sean, Matt, and Ryan, thank you for the endless encouragement and empathy, the visits and the phone calls, and the support you provided always. To Dr. Nedra Kearney-Vakulick and Vaughn Valkulick, thank you for being my family-away-from-family and always being there for me.

I also wouldn't be where I am without the support of my wonderful family, which owes its very existence to academia. Having family who understand the trials and tribulations of higher education has been a huge help throughout this journey. Thanks go first to my parents, Drs. Mark and Meghan Twiest, for always encouraging me to

pursue my interests and taking me around the country to experience all the U.S. has to offer, not to mention helping with several cross-country moves! I would also like to thank my parents for dealing with my early childhood fear of severe thunderstorms and tornadoes – it worked out pretty well in the end! Thanks also go to my in-laws, Jan and Lorrie Gallo, who have become second parents to me over the last decade. Thanks also go to my brother Joost, who always has my back despite being an ocean away, and my sister-in-law and brother-in-law Kara and Tyson Bugis, who have also encouraged me constantly from the very start of this process. A very special thanks goes to my grandmother Joni and late grandfather Jack Mahoney, for always making sure that I was maintaining a work-life balance and constantly reminding me that I was loved, and to my grandparents Gilbert and Linda Twiest, who helped inspire me to do the difficult things in life with a positive attitude, making the best of every opportunity.

Finally, I have to extend thanks beyond thanks to my wonderful husband Jed, who has provided constant love and support from near and far. He has been privy to the highs and the lows of this process, celebrating the highs and buoying me at the lows. His patience, understanding, and sacrifices have enabled me to pursue all of my dreams, and I can never express how much that means to me.

Table of Contents

List of Tables

List of Figures

Abstract

Hourly maximum fields of simulated storm diagnostics from experimental versions of convection-allowing models (CAMs) provide valuable information regarding severe weather potential. The focus of this work is to extract operationally relevant tornado probabilities from the CAM-based Weather Research and Forecasting (WRF) ensemble initialized daily at the National Severe Storms Laboratory (NSSL-WRF). Probabilities are derived in three main ways: by using updraft helicity (UH), UH filtered by model-derived environmental parameters, and through combining UH and model-derived environmental parameters with observed climatological tornado frequencies. Contrasting these methods compares a binary threshold exceedance approach and a probabilistic paradigm. Rather than using a specific threshold of UH as a proxy for tornadogenesis and relying on the ensemble to generate probabilities, the probabilistic approach treats each point as having a certain probability of producing a tornado, depending on the surrounding environmental conditions. Additionally, the ensemble-generated forecasts using both approaches are compared with the 0600 UTC Storm Prediction Center (SPC)'s tornado probabilities, to determine whether ensemble forecasts approach the skill of expert forecasters. While the methods derived using the threshold approach overforecast tornado probability magnitude, the probabilities that incorporate climatological frequency information perform much more reliably, particularly when the storm timing was considered.

For the probabilistic forecasts to be operationally relevant, cooperation with forecasters is critical in their development. A database of right-moving supercells developed by SPC forecasters was used to generate the climatological frequencies on

which three sets of probabilities are based. Through NOAA's Hazardous Weather Testbed Spring Forecasting Experiment (SFE), subjective daily evaluation of the probabilistic forecasts provided feedback during SFE 2015 that led to the development of the climatological frequency probabilities. When the climatological frequency probabilities were evaluated in SFE 2017, the prevalence of false alarm from nocturnal mesoscale convective systems led to the incorporation of timing information which reduces that false alarm. Therefore, the forecast probabilities are targeting the right-moving supercells, reflecting the underlying climatological frequencies.

Chapter 1: Introduction

Severe convective hazards such as hail, thunderstorm winds, and tornadoes are a common threat in the United States, with its unique geography particularly conducive to the formation of persistently rotating thunderstorms, called supercells. The defining characteristic of these storms is the rotating updraft at the core of the storm, which allows for dynamic pressure perturbations to enhance the lift generated by the realization of convective available potential energy (CAPE). The rotation of the storm is initiated by shear between the mid-troposphere and the surface, which also helps to displace precipitation from the updraft. The combination of the displaced precipitation and the dynamic pressure perturbations leads to long-lasting, quasi-steady-state storms, with the potential to leave a swath of damage in their wake. Supercells also produce a majority of tornadoes compared to other individual storm morphologies, including 89% of EF2+ and 97% of EF3+ tornadoes (Smith et al. 2012). Tornadoes cause tens to hundreds of deaths and hundreds of millions of dollars in damages each year (Simmons et al. 2013), motivating studies on how best to forecast tornadoes and their parent storms and improve forecasters' capability to protect lives and property.

The first work to address the question of tornado forecasts and verification was Finley (1884), which highlights the local effects of tornado prediction and calls for minimizing false alarm in forecasts. The small-scale nature of tornadoes inherently makes their prediction difficult; not only are tornadoes rare events, but the processes leading to tornado formation are highly localized and difficult to observe. However, field campaigns and idealized numerical modelling experiments have allowed researchers to determine environmental characteristics of tornadic storms and discern

favorable ingredients for tornadogenesis. By applying ingredients-based methods to tornado forecasting following Doswell et al. (1996) and taking advantage of increased numerical weather prediction (NWP) capability, operational forecasts of potentially tornadic environments have improved since the 1970s (Hitchens and Brooks 2012).

With the increase in computer power over the last 40 years, the ability to run finer grid-resolution NWP models over larger areas has allowed for the depiction of finer-scale atmospheric phenomena. Models with grid spacings as small as 3- and 4-km are now run for the entire contiguous United States, successfully simulating convective-scale phenomena. Though these models are not run at a fine enough grid spacing to explicitly resolve features of convective overturning such as the entrainment process [which would require a horizontal grid spacing of ~100 m according to Bryan et al. (2003)], turning off the convective parameterization at horizontal resolutions of 3–4 km reproduces much of the mesoscale structure and evolution of linear convective systems, including depiction of the cold pool (Weisman et al. 1997). The capacity to reproduce realistic convective systems within these NWP models aids operational forecasters attempting to determine convective storm characteristics (Weisman et al. 2008; Kain et al. 2008). The simulated reflectivity from convection within these models often resembles actual radar reflectivity, giving forecasters insight into convective occurrence, evolution, and mode. In addition to typical NWP parameters such as temperature and pressure, convection-allowing models (CAMs) allow for storm-based metrics, including metrics diagnosing storm rotation. However, a rotating simulated storm does not necessarily indicate a tornado threat. Storm-scale dynamics often determine whether or not a tornado will occur, and simulated mesocyclones (as in

reality) will not always be tornadic. Therefore, midlevel rotation alone is not expected to be a perfect indicator of whether or not a tornado is expected to form and additional information is required.

Output from a single forecast model provides one scenario of how a day's weather may unfold, but individual models imperfectly depict the atmosphere. Even the initial state of the atmosphere is never perfectly known, due to limited observing capabilities. A common solution is to create ensembles of NWP models, providing a range of solutions by using multiple initial conditions, differing parameterizations of small-scale atmospheric processes, and different methods of incorporating observations. By presenting a number of solutions, the eventual outcome will ideally fall within the envelope of the ensemble solutions. An operational convection-allowing ensemble became available on 1 November 2017, but experimental convection-allowing ensembles have been used by severe weather forecasters when available for years. From these ensembles, rather than having a deterministic *yes* or *no* forecast as to whether severe convection will occur at a given point, a probability of severe convection occurring can be generated. These probabilities take into account the uncertainty inherent in severe convective forecasting.

The central question of this dissertation is how to best utilize guidance from convection-allowing NWP ensembles to generate probabilistic tornado forecasts, which will in turn aid operational forecasters. Multiple methods are developed and tested objectively, using statistical metrics, and subjectively, getting feedback from researchers and forecasters. Testing and verification take place in a daily, operational setting, as well as aggregated across spring seasons, as the methods' usefulness to

forecasters are contingent upon both their daily performance and their seasonal performance. Isolating the tornado threat from the severe convective threat is a difficult challenge, but combining high-resolution CAM ensembles with prior studies of environmental parameters provides an opportunity to attack this challenge with the newest tools available.

1.1 Research Background

NWP forecasts focused on severe convective storms began with idealized studies of convective dynamics on relatively coarse grids (e.g., Steiner 1973). As computer power has increased, idealized models have been run at finer and finer grid spacing, including simulations run down to 30-m horizontal grid spacing that are capable of simulating tornadic supercells (Orf et al. 2017). Running NWP with such small grid spacing requires large computational resources, produces terabytes of data, and has limited domain constraints. However, lessons from idealized simulations can often give insight to forecasting processes. For example, Weisman and Klemp (1982, 1984) developed parameter studies that linked the Richardson number and convective mode, helping forecasters anticipate particular hazards associated with different modes. Operationally, the first experiment to help determine real-time storm mode was the Storm Type Operational Research Model Test Including Predictability Evaluation (STORMTIPE; Brooks et al. 1993; Wicker et al. 1997), which used an environmental sounding in an idealized simulation to determine storm mode. Presently, the highest-resolution operational model that the National Centers for Environmental Prediction (NCEP) runs (the High-Resolution Rapid Refresh [HRRR] model), has 3-km grid

spacing (Benjamin et al. 2016) and runs hourly, providing specific information on initiation and evolution of severe convective storms, as well as storm mode (Kain et al. 2008; Clark et al. 2012a). The widespread increase in computing power has led to multiple agencies running different experimental convection-allowing models, with the output available online [example agencies include the National Severe Storms Laboratory (NSSL), Texas Tech, the Earth Systems Research Laboratory, and the Center for the Analysis and Prediction of Storms (CAPS), among others].

While deterministic NWP forecasts provide realistic scenarios of how convection may occur on a given day, the large uncertainty inherent in small-scale prediction encourages an ensemble approach to realize multiple outcomes. Operational ensemble forecasting at coarse grid spacing began at NCEP and the European Centre for Medium-Range Weather Forecasts (ECMWF) in December of 1992 (see Kalnay 2003 for a thorough review of historical ensemble configuration techniques). Kalnay (2003) states that ensemble forecasting has three basic goals: (1) to improve forecasts via ensemble averaging, (2) to provide an indication of the reliability of the forecast, and (3) to provide a quantitative basis for probabilistic forecasting. All three of these basic goals can be applied to the severe convective forecasting problem specifically. CAPS developed the first CAM ensemble system in 2007 for the annual Spring Forecasting Experiment (SFE), testing multiple methods for generating different members (Xue et al. 2007). Different ensemble configurations were then contributed by CAPS to subsequent SFEs, and other agencies were contributing ensembles by the 2014 SFE. The current operational high-resolution ensemble, the High Resolution Ensemble Forecast, version 2 (HREFv2) is based on the Storm-Scale Ensemble of Opportunity

(SSEO; Jirak et al. 2012a), assembled by the Storm Prediction Center (SPC) from deterministic CAMs developed by NSSL and the Environmental Modeling Center (EMC). Since the member CAMs are available of the SSEO are available daily (though not all are operational), this grouping served as a "poor man's ensemble" and provided a good starting point for an operationalized CAM ensemble, which became available on 1 November 2017.

SFEs have served as a testing ground for CAM ensembles since 2007, bringing operational forecasters, researchers, and model developers together. Formal SFEs began in 2000 and continue to this day, although informal collaboration occurred prior to the implementation of the formal programs (Kain et al. 2003). These experiments test cutting-edge NWP models and post-processing techniques, as well as give forecasters the opportunity to provide feedback to researchers to aid in developing useful tools. A thorough overview of the SFEs, particularly the 2015 SFE, is presented in Chapter 2. Since there was a desire for work within this dissertation to be operationally based, obtaining feedback on the products developed herein was critical to ensure that forecaster concerns were addressed and guidance was generated that forecasters could trust, making its operational use more likely.

Although cutting-edge technology is tested each year in the SFEs, operational forecasters also play an essential role in the forecast process by incorporating the latest NWP with their knowledge of the atmosphere to generate the best possible forecast. While the specific conditions leading to tornadogenesis by a particular supercell are often extremely small-scale (e.g., interaction with small-scale boundaries; Markowski et al. 1998; Rasmussen et al. 2000), large-scale environmental characteristics conducive to

supercellular storms and subsequent tornadogenesis are less subject to large uncertainty than individual storm attributes, making them more easily anticipated and modeled by NWP output. Thus, forecasters often use an ingredients-based approach (Doswell et al. 1996) to forecasting tornadoes, assessing where environmental conditions conducive to supercells and subsequent tornadogenesis may occur. Brooks et al. (2003) found high CAPE and strong 0–6 km shear in proximity soundings to supercellular convection, two ingredients that can be depicted by both coarse-resolution and fine-resolution NWP.

However, identifying conditions favorable to supercells is insufficient for tornado forecasting — Trapp et al. (2005) found that ~26% of storms with mesocyclones produced tornadoes, and Thompson et al. (2017) found that number to be just 18%. Rasmussen and Blanchard (1998) identified two further parameters that typically differ between tornadic supercells and non-tornadic supercells: 0–3 km storm-relative helicity (SRH), and the lifted condensation level (LCL) height. Tornadic supercells tended to have higher SRH and lower LCLs than non-tornadic supercells, although some overlap occurred between the distributions. These two parameters, in addition to the 0–6 km bulk shear and the CAPE from a 100 mb mixed-layer parcel, were combined by Thompson et al. (2003) into the significant tornado parameter (STP). Using the STP, supercells producing a significant tornado (defined therein as producing F2 or greater damage) had a statistically significantly larger STP value than non-tornadic supercells. The STP was formulated so that a value of 1 best discriminated the two types, but Thompson et al. (2003) noted the importance of convective mode prediction in forecasting, wanting to preclude any use of this metric as a “magic number”. The STP was later improved upon by modifying the relative weights of each

parameter, using effective-layer wind shear parameters to better reflect the storm inflow, and adding a convective inhibition (CIN) term to limit areal false alarms (Thompson et al. 2012). While other composite parameters have been tested (Hart and Korotky 1991; Rasmussen 2003; Craven and Brooks 2004), the STP remains a key tool for forecasters looking to summarize the environmental ingredients conducive to tornadogenesis.

While convection-parameterizing and CAM ensembles can both simulate environmental parameters conducive to tornadogenesis, CAM ensembles explicitly depict convection from which storm-based diagnostics can be computed, including a metric determining the rotational characteristics of a simulated storm: updraft helicity (UH). Formulated as the vertical vorticity times the updraft speed integrated over a layer, 2–5 km UH was determined to be a reliable indicator of mesocyclone-scale rotation (Kain et al. 2010), and therefore a successful indicator of supercells (Carley et al. 2011; Naylor et al. 2012). UH fields soon became used throughout the literature to identify areas of general severe convective threats in deterministic and ensemble frameworks (Sobash et al. 2011; Schwartz et al. 2015a; Sobash et al. 2016a; Loken et al. 2017) and were extended to individual hazards forecasting (Clark et al. 2012b; Sobash et al. 2016b, Gagne et al. 2017). The ensemble can then provide probabilistic hazard information, as well as multiple possible realizations of convection, which could lend confidence in forecasting convective mode.

As the prevalence of CAM ensembles increases, ever more information is being provided to operational forecasters – six CAM ensembles were available and evaluated in real-time for the 2015 SFE, for example. Since forecasters are working within strict

time constraints for product issuance and often do not have the time to consider each member of each ensemble, post-processing the CAM ensemble output summarizes relevant information for the forecasters, supplanting the need to look separately at storm-scale and environmental fields. How best to post-process this information is the crux of this dissertation, which aims to determine which metric or combination of metrics provides the best forecast, what impact incorporating empirical climatologies into forecasts has, and how the timing of specific parameters may influence the forecasts. Together, these questions determine how to formulate a reliable first-guess product for operational forecasters, distilling the flood of information to a manageable, reliable, and useful graphic.

1.2 Research Hypotheses

Four hypotheses were designed to explore how convection-allowing ensembles may be used to create skillful tornado probabilities. These hypotheses were tested through typical forecast verification metrics, but also through real-time evaluation by researchers and forecasters in NOAA's Hazardous Weather Testbed during annual SFEs. These hypotheses all share the core principle that additional information from convection-allowing models can add to the storm-scale attributes provided by these ensembles to generate tornado probabilities, rather than probabilities of severe convective hazards as a whole. The first hypothesis is that *adding high-resolution information to constrain tornado probabilities to areas that are environmentally favorable to tornadogenesis will result in more skillful probabilities than solely using 2–5 km UH.* Using environmental information to constrain the probabilities eliminates

areas that have, for example, high cloud bases or in which storms are drawing their inflow from above the surface layer.

The second hypothesis tested is that *incorporating observed tornado frequencies given a right-moving supercell will provide more accurate and reliable probabilities than those generated solely using model-derived information.* The method used to test this hypothesis treats each grid point as though it has a probability of generating a tornado given some environmental information, rather than relying on fixed thresholds of UH and environmental information. In addition, the probability of a tornado given a value of STP is rooted in observed tornado frequencies, giving the probabilities a foundation in observed storm characteristics.

The third hypothesis is that *tornado probabilities generated using a convection-allowing ensemble can be used operationally as first-guess tornado forecasts and have similar verification statistics to initial probabilistic tornado forecasts issued by the Storm Prediction Center (SPC) at 0600 UTC.* This hypothesis addresses the operational nature of the probabilities and helps determine the usefulness of multiple methods of tornado probability formation by directly comparing model-generated forecasts to operationally issued forecasts to determine the strengths and weaknesses of the model-generated forecasts. If the model-generated forecasts are useful as starting points for operational forecasters, it may help reduce the burden on forecasters caused by large amounts of high-resolution data provided by convection-allowing ensembles.

The fourth and final hypothesis explored by this dissertation is that *incorporating temporal information regarding UH occurrence will reduce areas of false alarm linked to nocturnal mesoscale convective systems (MCSs), which often*

produce UH in NWP but do not often produce tornadoes. This hypothesis arises from observations during the 2015 and 2017 SFEs, when the tornado probabilities were tested in real-time. Broad swaths of false alarm were linked to nocturnal systems, which are less likely to produce tornadoes than systems occurring earlier in the day due to a decrease in CAPE and an increase in CIN as the surface layer becomes decoupled from the free atmosphere. If nocturnal UH can be weighted less than diurnal UH when generating daylong forecast tornado probabilities, the forecasts are hypothesized to be more useful to operational forecasters by producing fewer false alarms.

Taken together, these hypotheses advance the usage of convection-allowing ensembles to make tornado forecasts. As tornadoes are particularly high-impact events with a large impact on society, having accurate probabilistic forecasts on the daylong convective outlook scale can allow forecasters to focus on more rapidly evolving, shorter-term scenarios that are more difficult to capture with convection-allowing models.

1.3 Dissertation Organization

Chapter 2 of this dissertation is a paper providing an overview of the 2015 Spring Forecasting Experiment (SFE), *Breaking New Ground in Severe Weather Prediction: The 2015 NOAA/Hazardous Weather Testbed Spring Forecasting Experiment*. This paper was published by *Weather and Forecasting* in August of 2017. Additionally, Chapter 2 describes how convection-allowing ensembles are contributing to new forecast products. Since the goal of this dissertation is to provide operationally

relevant forecast probabilities, establishing an overview of the real-time experiment in which they are tested offers operational context for the remainder of the dissertation.

A paper investigating the first hypothesis described above is assigned to Chapter 3, *Forecasting Tornadoes using Convection-Permitting Ensembles*, which was published by *Weather and Forecasting* in February of 2016. A third paper, *Blended Probabilistic Tornado Forecasts: Combining Climatological Frequencies with NSSL-WRF Ensemble Forecasts* investigates hypotheses two and three, is assigned to Chapter 4, and has been conditionally accepted by *Weather and Forecasting*. Finally, a fourth paper, *The Impact of Updraft Helicity Timing on Ensemble-Derived Tornado Probabilities*, will explore the final hypothesis and is assigned Chapter 5. This paper will be submitted to *Weather and Forecasting*. Chapter 6 will consist of general conclusions and propose directions for future research.

Chapter 2: Breaking New Ground in Severe Weather Prediction: The 2015 NOAA/Hazardous Weather Testbed Spring Forecasting Experiment

A paper published in *Weather and Forecasting*

Burkely T. Gallo[1], Adam J. Clark[2], Israel Jirak[3], John S. Kain[2], Steven J. Weiss[3], Michael Coniglio[2], Kent Knopfmeier[4,2], James Correia Jr.[4,2],Christopher J. Melick[8],Christopher D. Karstens[4,2], Eswar Iyer[1], Andrew R. Dean[3], Ming Xue[1,5], Fanyou Kong[5], Youngsun Jung[5], Feifei Shen[5], Kevin W. Thomas[5], Keith Brewster[5], Derek Stratman[1,5], Gregory W. Carbin[6], William Line[4,3], Rebecca Adams-Selin[7], and Steve Willington[9]

[1]School of Meteorology, University of Oklahoma, Norman, Oklahoma

[2]NOAA/OAR National Severe Storms Laboratory, Norman, Oklahoma

[3]NOAA/Storm Prediction Center, Norman, Oklahoma

[4]Cooperative Institute for Mesoscale Meteorological Studies, University of Oklahoma, Norman, Oklahoma

[5]Center for Analysis and Prediction of Storms, Norman, Oklahoma

[6]NOAA/Weather Prediction Center, College Park, Maryland

[7]Atmospheric and Environmental Research, Inc., Lexington, Massachusetts

[8]557th Weather Wing/16th Weather Squadron, Offutt AFB, Nebraska

[9]Met Office, Exeter, Devon, UK

Abstract

Led by NOAA's Storm Prediction Center and National Severe Storms Laboratory, annual Spring Forecasting Experiments (SFEs) in the Hazardous Weather Testbed test and evaluate cutting-edge technologies and concepts for improving severe weather prediction through intensive real-time forecasting and evaluation activities. Experimental forecast guidance is provided through collaborations with several United States government and academic institutions, and the United Kingdom Met Office. The purpose of this article is to summarize activities, insights, and preliminary findings from recent SFEs, emphasizing SFE 2015. Several innovative aspects of recent experiments are discussed, including (1) use of convection-allowing model (CAM) ensembles with advanced ensemble data assimilation, (2) generation of severe weather outlooks valid at time periods shorter than those issued operationally (e.g., 1 to 4 h), (3) use of CAMs to issue outlooks beyond the Day 1 period, (4) increased participant interaction through software allowing participants to create individual severe weather outlooks, and (5) tests of newly developed storm-attribute based diagnostics for predicting tornadoes and hail size. Additionally, plans for future experiments will be discussed, including creation of a Community Leveraged Unified Ensemble (CLUE) system, which will test various strategies for CAM ensemble design using carefully designed sets of ensemble members contributed by different agencies to drive evidence-based decision making for near-future operational systems.

2.1 Introduction

Annual Spring Forecasting Experiments (SFEs) conducted in the National Oceanic and Atmospheric Administration (NOAA)'s Hazardous Weather Testbed (HWT) provide opportunities for testing new tools and techniques in forecasting severe thunderstorms. Jointly run by the National Severe Storms Laboratory (NSSL) and the Storm Prediction Center (SPC), SFEs provide a two-way research-to-operations/operations-to-research pathway for enhanced understanding and problem-solving regarding severe thunderstorm forecasting. The real-time SFE takes place during the spring severe weather season, providing realistic operational pressure for participants as each day provides a unique set of conditions regarding severe weather potential.

Formal SFEs began in 2000; Kain et al. (2003) emphasizes that collaboration is the crux of SFEs, noting that "the interaction between forecasters and numerical modelers was the most rewarding part of (the) Spring Program". This collaboration has created greater forecaster understanding of numerical models and greater researcher understanding of operational challenges (Kain et al. 2003). Clark et al. (2012a) further emphasizes SFE's collaborative aspects, detailing the extension of severe thunderstorm forecasts issued during SFE 2010 to aviation and heavy precipitation interests.

While SFEs involve real-time forecasting, daily evaluation exercises are another key aspect of SFEs (Clark et al. 2012a). Evaluating cutting-edge techniques such as experimental severe weather guidance derived from convection-allowing models (CAMs) allows participants to grasp strengths and weaknesses of each technique and assess readiness for operational adoption. Subjective evaluations illustrate the

impressions participants have, while objective evaluations often take place after SFEs when time permits a thorough examination of the large volume of data (e.g., Johnson et al. 2013, Smith et al. 2014, Surcel et al. 2014, Duda et al. 2014).

Since 2007, the Center for Analysis and Prediction of Storms (CAPS) at the University of Oklahoma has provided a real-time CONUS forecast at 4-km grid spacing from a multi-model Storm-Scale Ensemble Forecast system (SSEF) to the SFE (Kong et al. 2015 and references therein). This system was reduced to 3-km grid spacing for SFE 2015. SFE 2015 also included five other unique CAM ensembles. Multiple organizations contributed NWP forecasts, including the Environmental Modeling Center (EMC), Earth Science Research Lab's Global Systems Division (ESRL/GSD), NSSL, CAPS, the National Center for Atmospheric Research (NCAR), and the 557th Weather Wing (formerly the Air Force Weather Agency [AWFA]). Experimental deterministic guidance also featured during SFE 2015, particularly three versions of the Unified Model (UM; Davies et al. 2005) from the United Kingdom Met Office and the Model for Prediction Across Scales (MPAS; Skamarock et al. 2012) from NCAR.

SFE 2015 pursued a number of goals consistent with the visions of both the Forecasting a Continuum of Environmental Threats (FACETs; Rothfusz et al. 2014) and Warn-on Forecast (WoF; Stensrud et al. 2009) initiatives. These programs aim to generate probabilistic hazard information (PHI), to go beyond the current binary paradigm of products such as watches, warnings, and advisories. Under a probabilistic paradigm, forecasters can give users more specific, understandable information that they can use to take action based on their individual needs. Developing probabilistic guidance to support this new paradigm requires cooperation between the operational

forecasting and research communities, making the SFEs optimal for exploration of probabilistic forecasts. SFE 2015's goals fall into two categories consistent with the visions of FACETs and WoF: (1) Operational Product and Service Improvements and (2) Applied Science Activities. The Operational Product and Service Improvements goals focused on model guidance-driven forecast *generation* by participants, while Applied Science Activities focused on the *evaluation* of new forecasting tools and forecast types, including new numerical guidance and post-processing techniques. Numerical guidance characterization supported both types of goals by determining how to incorporate guidance into the forecasts and evaluating model output fields such as simulated reflectivity and hail size estimates.

Introduced in SFE 2014 and continued in SFE 2015 is the incorporation of individual participant forecasts, essentially forming an "ensemble" of participant forecasts (Coniglio et al. 2014). Prior SFEs solely issued group forecasts, reaching a consensus on the placement of the day's probability contours. While group discussion and consensus forming remained an integral part of the Day 1 full period forecasting process, individuals then created higher time frequency forecasts. These forecasts tested the feasibility of operationally issuing more forecasts, each covering a shorter time window, and the subsequent increase in forecaster workload. Individuals' forecasts also illustrated a variety of forecasting approaches, with differing reliance on observations, model guidance, and prior forecaster experience.

Also new to SFE 2015 are the evaluation capabilities of participants using laptops with internet connectivity. Previously, evaluations were also consensus-based. However, laptop usage enabled approximately five independent forecaster ratings per

day for each evaluation. Although the SFE leaders had documented previous experiments' discussions, enabling individuals to comment on products provided a more complete record of opinions, suggestions, and reflections on each product's operational potential than in previous experiments.

This paper provides a broad overview of SFE 2015 and its innovations, which advance the two-way research-to-operations/operations-to-research pathway inherent to SFEs. Section 2.2.1 of this paper describes the numerical weather prediction (NWP) systems utilized throughout SFE 2015, and Section 2.2.2 elaborates upon the daily activities of the SFE. Section 2.3 highlights preliminary results from the SFE, including subjective and objective evaluations. Finally, Section 2.4 provides a summary and evaluation of SFE 2015, along with plans and directions for future SFEs.

2.2 Experiment Description

2.2.1 Experimental Numerical Guidance

SFE 2015 focused on experimental probabilistic forecast generation informed by a suite of experimental NWP forecasts. Four of the six experimental ensembles extended into the Day 2 period, allowing for exploration of longer-range CAM forecasts. All models detailed below produced hourly maximum fields (Kain et al. 2010) of explicit storm attributes such as simulated reflectivity and updraft helicity (UH) for forecasting and evaluation purposes.

a. NSSL-WRF and NSSL-WRF Ensemble

SPC forecasters have used output from an experimental 4-km grid spacing Weather Research and Forecasting (WRF; Skamarock et al. 2008) Advanced Research

WRF (ARW) model produced by the NSSL (Kain et al. 2010) since the fall of 2006. Currently, this model runs twice daily at 0000 UTC and 1200 UTC over a full-CONUS domain, with forecasts to 36 hours (Table 2.1, Ensemble Member cn [control]). Nine additional 4km WRF-ARW members are run at 0000 UTC to 36 hours by varying the initial conditions and lateral boundary conditions of the control, to compose the 10-member NSSL-WRF ensemble (Table 2.1; Gallo et al. 2016). These members use the 0000 UTC National Centers for Environmental Prediction (NCEP) Global Forecast System (GFS) analysis or the 3-h Short-Range Ensemble Forecast (SREF; Du et al. 2014) system forecasts initialized at 2100 UTC for initial conditions and corresponding GFS or SREF member forecasts as lateral boundary conditions. Physics parameterizations amongst all members are identical.

b. CAPS Storm-Scale Ensemble Forecast Systems

The Center for the Analysis and Prediction of Storms (CAPS) provided two ensembles to SFE 2015. The 20-member SSEF system included 12 members that accounted for as many sources of forecast error as possible (e.g., initial conditions, boundary conditions, multi-physics; Table 2.2). These members were used to generate probabilities of severe convective hazards. The eight remaining members tested physics sensitivities. WSR-88D data was used for data assimilation along with available surface and upper air observations using the Advanced Regional Prediction System (ARPS) 3DVAR/cloud-analysis system (Xue et al. 2003; Hu et al. 2006) to produce the control member. The 0000 UTC North American Mesoscale (NAM) model analysis on a 12-km grid was used as a background for the analysis, and NAM forecasts provided boundary conditions. Perturbed members applied initial condition and boundary condition

perturbations drawn from the SREF to the control analyses and forecasts. The CAPS forecasts were run with 3-km grid spacing and extended to 60 h, supporting Day 2 forecasts.

A separate 12-member ensemble of 60-h forecasts was also produced on the same 3-km domain as the prior SSEF system (Table 2.3) using XSEDE supercomputing facilities (Towns et al. 2014). Rather than 3DVAR, the ensemble Kalman filter (EnKF; Evensen 1994, 2003) data assimilation method was used, specifically the CAPS EnKF DA system (Xue et al. 2006; Wang et al. 2013) that has been directly interfaced with the WRF model. Specifically, 40-member ensemble forecasts were launched from NAM analysis plus SREF perturbations at 1800 UTC, and run to 2300 UTC. The configuration of this ensemble involved both initial perturbations and mixed physics options, to provide a variety of input for the EnKF analysis. Each member used the WRF single-moment six-class (WSM6; Hong and Lim 2006) microphysics with different intercept parameter settings for rain and graupel, and the density of graupel, and included relatively small random perturbations (0.5 K for potential temperature and 5% for relative humidity) with recursive filtering of approximately 20-km horizontal correlations scales. EnKF cycling utilizing radar data was performed every 15 minutes from 2300 UTC to 0000 UTC, using the 40-member ensemble as background. Besides radar data, only Meteorological Assimilation Data Ingest System (MADIS; Miller et al. 2005, 2007) surface observations, profiler, and radiosondes were assimilated at 2300 UTC and/or 0000 UTC. A 12-member ensemble forecast to 60 h followed, using the last EnKF analyses at 0000 UTC (Table 2.3).

c. SPC Storm-Scale Ensemble of Opportunity

The SPC Storm-Scale Ensemble of Opportunity (SSEO; Jirak et al. 2012a) is a 7-member, multi-model, multi-physics ensemble consisting of deterministic CAMs available year-round to the SPC (Table 2.4). Individual members include one model produced by NSSL, and six members produced by EMC. The ensemble has been utilized in SPC operations since 2011 as a practical alternative to a formal storm-scale ensemble (Jirak et al. 2012a), which is planned for implementation in the next few years (Dimego, G., personal communication). Forecasts are initialized from the operational NAM with no additional data assimilation and are generated twice daily to 36 hours, starting at 0000 UTC and 1200 UTC. These members differ slightly in grid spacing (3.6 km to 4.2 km), vertical levels, and length, with 36-h forecasts, 48-h forecasts, and 60-h forecasts. Microphysics schemes of the members include WSM6, Ferrier (Ferrier 1994), and Ferrier-Aligo (Aligo et al. 2014).

d. Air Force Weather Agency 4-km Ensemble

The U.S. Air Force 557th Weather Wing at Offutt Air Force Base (USAF) ran a real-time 10-member, 4-km WRF-ARW model ensemble (AFWA; Kuchera et al. 2014) over the CONUS for SFE 2015 to 60 h (Table 2.5). Forecasts were initialized twice daily, at 0000 UTC and 1200 UTC, using 6- or 12-h forecasts from three global models: the Met Office UM, the NCEP GFS, and the Canadian Meteorological Center Global Environmental Multiscale (GEM) Model. Member microphysics and boundary layer parameterizations varied, and no data assimilation was performed during initialization.

e. NCAR EnKF-based Ensemble

In SFE 2015, NCAR provided a new 10-member, 3-km grid spacing ensemble with a CONUS domain (Schwartz et al. 2015b). EnKF data assimilation occurred every

6 h with 15-km grid spacing using the following observational sources: Aircraft Communications Addressing and Reporting System (ACARS), MADIS surface observations, METARs and radiosondes, NCEP MARINE, Cooperative Institute for Meteorological Satellite Studies (CIMSS) cloud-track winds (Menzel 2001), and the Oklahoma Mesonet stations. From this mesoscale background, ten downscaled 3-km forecasts were initialized daily at 0000 UTC using consistent physics with the data assimilation system, sans cumulus parameterization. The first ten members of the analysis were selected after random shuffling between analyses, and therefore differed daily. Each selection of ten members was equally representative of the ensemble mean analysis and perturbations, and unique lateral boundary condition perturbations were member-dependent, but used random draws from global background error covaraiances. (Schwartz et al. 2014). Both the data assimilation scheme and the forecasts used Thompson microphysics (Thompson et al. 2008), Rapid Radiative Transfer Model (RRTM; Mlawer et al. 1997) for Global Climate Models (RRTMG; Iacono et al. 2008), Mellor-Yamada-Janjić (MYJ; Mellor and Yamada 1982; Janjić 1994, 2002) planetary boundary layer (PBL) parameterization, and the Noah land surface model (Chen and Dudhia 2001). The analysis system contained 50 members of constant physics that were continuously cycled using the ensemble adjustment Kalman filter (EAKF; Anderson 2001, 2003) within NCAR's Data Assimilation Research Testbed (DART; Anderson et al. 2009) software. The analyses provided initial conditions for the daily forecasts, which were run to 48 h. Both the analyses and the forecasts had 40 vertical levels.

f. UKMET Convection-Allowing Model Runs

The Met Office provided three nested, limited-area high-resolution versions of the UM to SFE 2015: two at 2.2-km grid spacing, and one at 1.1-km grid spacing. The operational 2.2-km version incorporated the UM specifications currently run in the Met Office's operational 1.5-km grid length, UK-centered model (McBeath et al. 2014; Mittermaier 2014). The operational 2.2 km provided for SFE 2015 had 70 vertical levels across a domain ranging from just west of the Rocky Mountains to the western border of Maine. Initial and lateral boundary conditions were taken from the 0000 UTC 17-km global version of the UM without additional data assimilation, and forecasts extended to 48 hours.

A unique aspect of the UM models was the configuration of the turbulence parameterization. The operational run used a 3D turbulent mixing scheme consisting of a locally scale-dependent blending of Smagorinsky (Smagorinsky 1963) and boundary layer mixing schemes, wherein stochastic perturbations were made to the low-level resolved scale temperature field in conditionally unstable regimes to encourage the transition from subgrid to resolved scale flows (Clark et al. 2015). This turbulent mixing scheme differs from that of WRF, which utilizes 3D Smagorinsky turbulence closure to determine eddy viscosities in the absence of a PBL scheme (Skamarock et al. 2008). The operational 2.2-km run had single moment microphysics (Wilson and Ballard 1999), and diagnosed partial cloudiness assuming a triangular moisture distribution whose width is a function of height only.

The parallel version of the 2.2-km UM used an experimental parameterization of partial cloudiness, expanding upon the prognostic scheme used in the Met Office global UM. The parallel scheme includes an additional parameterization of subgrid moisture

variability linked to the boundary layer turbulence. This version was also run to 48 hours, and was otherwise identical to the operational 2.2-km version of the UM.

Finally, the 1.1-km horizontal resolution UM centered on Oklahoma ran over a 1300 km by 1800 km domain nested within the 2.2 km model. The initial and lateral boundary conditions were taken from hour 3 of the 0000 UTC 2.2 km run to reduce spinup time, and run to 33 hours. The 1.1-km run was otherwise identical to the 2.2-km operational run, thereby testing the horizontal resolution effects.

g. Model for Prediction Across Scales (MPAS)

Another new deterministic modeling system provided to SFE 2015 was the MPAS, which produced daily 0000 UTC initialized forecasts at 3-km grid spacing over the CONUS. Forecasts from MPAS extended to 120 h (5 days), allowing for a unique glimpse into the long-range capabilities of convection-allowing models. The MPAS horizontal mesh is based on Spherical Centroidal Voronoi Tessellations (SCVTs; Satoh et al. 2008), allowing for quasi-uniform discretization of the sphere and local refinement with smoothly varying mesh spacing between regions with differing resolutions. Smoothly varying mesh eliminates major problems regarding transitions between differing resolutions of nests (Skamarock et al. 2012). MPAS has 55 vertical levels, and the "scale-aware" physics allows for the output of explicit storm attributes for those regions at convection-allowing resolution. Physics parameterizations include the MYJ PBL scheme and the WSM6 microphysics.

h. Parallel Operational CAMs

During SFE 2015, SPC had access to parallel versions of NAM and the High-Resolution Rapid Refresh (HRRR; Alexander et al. 2010), containing improvements

over the operational versions of these models (Table 2.6). The parallel versions were candidates for operational implementation by NCEP. Parallel high-resolution window (HRW) WRF-ARW and Nonhydrostatic Multiscale Model on the B grid (NMMB; Janjić and Gall 2012) runs included slight changes such as increasing vertical levels from 40 to 50, updating the WRF version used, and modifying the microphysics scheme in the WRF-ARW to decrease the amount of falling graupel. The parallel HRRR included changes in the physics to improve an afternoon warm, dry bias in the operational HRRR that had resulted in overpredicting convective initiation (Alexander et al. 2015). These changes included updating the microphysics to the Thompson-Eidhammer scheme (Thompson and Eidhammer 2014) and modifying the MYNN PBL scheme. Changes to the NAM Nest included reducing the grid spacing to 3 km in the parallel version, as opposed to the 4 km operational version. The parent NAM providing the boundary conditions was also updated.

2.2.2 Daily Activities

SFE 2015 was conducted weekdays from 4 May through 5 June 2015, excepting the Memorial Day holiday on 25 May 2015, for a total of 24 days. Each day, participants completed the same activities, separated broadly into experimental forecasts and evaluations.

a. Experimental forecasts

Daily activities were split between two "desks", led by SPC forecasters. Each desk focused on different experimental forecasts and evaluations, and participants rotated through desks during the week to gain exposure to all experimental products. Besides generating forecasts, participants at each desk evaluated prior forecasts and

experimental numerical guidance. Activities took place at roughly the same time each day (Table 2.7), and mainly occurred over regions of the United States which had the greatest potential for severe weather during a given day.

Participants at the "individual hazards" desk issued daily probabilistic forecasts of severe hail, damaging wind, and tornadoes within 25 miles (40 km) of a point, consistent with the SPC's definition of a severe convective hazard, valid from 1600 UTC to 1200 UTC the following day. Meanwhile, participants at the "total severe" desk forecasted the risk of any severe hazard following the SPC's operational Day 2 Convective Outlook format, valid over the Day 1 time period. Participants at both desks then refined their Day 1 forecasts into higher temporal resolution forecasts, with the individual hazards desk issuing hail, wind, and tornado forecasts for two 4-h periods: 1800-2200 UTC and 2200-0200 UTC. Individual hazard forecasters could use temporally disaggregated first-guess probabilities generated from the full-period hazard outlook to constrain and scale the magnitude and spatial extent of the SSEO neighborhood probabilities of proxy variables (i.e., UH for tornadoes, updraft speed for hail, 10 m wind speed for wind), ensuring consistency among the 24-h and 4-h forecasts (Jirak et al. 2012b).

At the total severe desk, probabilistic forecasts were manually stratified by participants and the desk lead forecaster into 1 h periods valid starting at 1800-0000 UTC. The 2100-0000 UTC forecasts were updated each afternoon, with two additional hourly forecasts issued from 0100-0200 UTC and 0200-0300 UTC. This approach was first attempted in 2014, and continued in 2015. Reliability diagrams computed post-SFE 2014, when hourly forecasts were issued from 1800-0300 UTC (Coniglio et al. 2014;

Fig. 2.1), showed that when verified on a 40-km grid (~20 km neighborhood), participants and the desk lead forecaster issued reliable hourly probabilistic forecasts, but overforecasted severe weather when verified on a 20-km grid (~10km neighborhood). These hourly forecasts were verified by gridding local storm reports (LSRs) and grid points of NSSL Multi-Radar Multi-Sensor maximum estimated size of hail (MESH; Witt et al. 1998) ≥ 29 mm (following Cintineo et al. 2012), aggregated over the nine hourly periods initially forecast (Fig. 2.1a) and the six afternoon update hours (Fig. 2.1b).

Hourly probabilistic forecasts were tested with the goal of introducing probabilistic severe weather forecasts on time scales that are currently addressed only as needed operationally (e.g., severe thunderstorm/tornado watches). Breaking down a full-period outlook into hourly probabilities also tested seamlessly merging probabilistic severe weather outlooks to probabilistic severe weather warnings, consistent with the visions of FACETs (Rothfusz et al. 2014) and WoF (Stensrud et al. 2009).

All participants individually generated the hourly forecasts using a web-based PHI tool (Karstens et al. 2014, 2015) to draw hazard probability contours (Fig. 2.2). Five laptops were available at each desk. If there were more participants than laptops at a desk, some participants worked in pairs to generate forecasts. Individual forecasts (Fig. 2.2a-e) were later compared to those issued by the desk lead (Fig. 2.2f). Participants subjectively evaluated the previous day's short-term forecast issued by the desk lead on a 1-10 scale, with 10 being the highest rating, compared to a "practically perfect" forecast (Brooks et al. 1998; Hitchens et al. 2013), which is analogous to probabilities a forecaster would issue with prior perfect knowledge of the LSR

distribution (Fig. 2.2g). While preliminary LSRs provided the largest component of ground truth, MESH, watches, warnings, and observed composite reflectivity were also considered.

The individual hazard desk's Day 2 outlooks explored the feasibility of issuing individual hazard forecasts beyond Day 1, utilizing experimental extended CAM guidance. Currently, individual hazard forecasts are limited to Day 1 in SPC operations. The total severe desk also generated Day 2 forecasts, as is done operationally by SPC, but informed by experimental CAM guidance. Day 3 forecasts were occasionally issued by the total severe desk, depending on time constraints and the anticipated severity of Day 3. MPAS often heavily informed these extended forecasts, particularly because two prior runs encompassed a Day 3 outlook, allowing consideration of run-to-run consistency.

The final forecasting activity of each day was an update to the earlier, participant-drawn forecasts informed by group discussion and updated data. Individual hazard participants updated their 2200-0200 UTC period, and the total severe participants updated their hourly forecasts from 2100-0000 UTC. Total severe participants also issued new hourly probabilities for 0000-0200 UTC.

While issuing forecasts, participants had access to high-temporal resolution satellite imagery. 1-min visible and infrared satellite imagery from GOES-14 was made available experimentally to participants during SFE 2015 from 18 May – 11 June. This special 1-min imagery, known as Super Rapid Scan Operations for GOES-R (SRSOR), helps to prepare users for the very-high-temporal-resolution sampling capability of the GOES-R Advanced Baseline Imagery (Line et al. 2016). SFE 2015 participants

primarily utilized the 1-min satellite imagery to identify and track boundaries, assess cumulus cloud trends, and diagnose areas of convective initiation.

b. Evaluations

In addition to forecasting activities, each participant performed multiple evaluations of the previous day's forecasts and model guidance. Participants rated the desk lead's forecasts and numerical guidance on a scale of 1 (Very Poor) to 10 (Very Good) and commented on particular strengths and weaknesses. This evaluation subjectively assessed the skill of the first-guess guidance and the human-generated forecasts for all periods (i.e., each hourly forecast at the total severe desk was assigned a rating). Model evaluations focused on the accuracy of the forecasts in predicting severe convective threats (including considerations such as the mode and timing of convective initiation) by comparing forecasts of hourly maximum fields (e.g., UH) relative to LSRs, maximum MESH, and radar observations across the previous day's domain. For ensembles extending to the Day 2 period, participants compared Day 2 guidance to Day 1 guidance, to examine if the ensembles improved with shorter lead times.

New experimental fields were also evaluated, such as hail guidance available in the WRF-ARW (Adams-Selin et al. 2014), tornado probabilities generated from the NSSL-WRF ensemble (Gallo et al. 2016), and pre-convective, model-generated environmental soundings from the UM and the NSSL-WRF. In SFE 2015, the WRF-HAILCAST algorithm was implemented in the CAPS ensembles to predict hail size (Adams-Selin and Zeigler 2016). This algorithm is a modified version of the coupled cloud and hail model found in Brimelow et al. (2002) and Jewell and Brimelow (2009), which forecast the maximum expected hail diameter at the surface using a profile of

nearby atmospheric temperature, moisture, and winds. The WRF-HAILCAST model uses WRF-generated convective cloud and updraft attributes coupled with a physical model of hail growth to determine hail growth from five predetermined initial embryo sizes. Another hail size diagnostic, derived directly from the microphysical parameterizations and developed by G. Thompson (Skamarock et al. 2008), was new to SFE 2015 and was output by the NCAR ensemble.

During SFE 2015, probabilistic tornado forecasts were generated from the NSSL-WRF ensemble using 2–5 km UH $\geq 75 m^2 s^{-2}$ as a proxy for tornadoes, with varying environmental constraints on probability generation. The environmental constraints required the probabilities to reflect UH only at grid points where certain environmental criteria were met in the previous hour: Lifted Condensation Level < 1500m, ratio of Surface-Based CAPE to Most Unstable CAPE ≥ 0.75 (Clark et al. 2012b), and Significant Tornado Parameter ≥ 1 (Thompson et al. 2003). Gallo et al. (2016) elaborates on the probability generation details. Tornado reports from the LSR database were overlaid on the forecast probabilities for subjective evaluation, which considered the entire CONUS.

Introduced in SFE 2014 and enhanced in SFE 2015 were three-dimensional animations of CAM output (Clyne et al. 2007). The 3D images were generated from a 600x600 km sub-domain of the CAPS control forecast chosen daily based on prior forecasts and 2D output fields. Selected 3D animations were shown to participants during the daily weather briefing, allowing a deeper investigation of the processes that lead to potential severe weather threats. These four-dimensional depictions showed features such as local UH (calculated at each volume rendered in the visualization;

Brewster et al. 2016), near-surface radar reflectivity, and near-surface wind vectors (Fig. 2.3). Deep columns of UH indicated supercellular storms, while animation of the images showed the longevity of such columns: long-lived UH columns often indicated heightened tornado risk.

The experimental forecasts for individual severe hazards were objectively evaluated in near real-time for SFE 2015, a continuation of efforts which had started in SFE 2014 (Melick et al. 2014). For the probabilistic hail forecasts, side-by-side spatial plots and corresponding forecast verification metrics for both LSR and MESH were provided daily, allowing participants to test the usefulness of alternative verifying data sources. For the current work, comparisons of MESH and LSR observational datasets for hail verification were made using the area under the receiver operating characteristic curve (ROC curve; Mason 1982) estimated using a triangular approach, which measures the ability of a forecast to discriminate between events (i.e., hail occurrence) and nonevents (i.e., no hail occurrence). ROC area values range from 0 to 1, with 1 indicating perfect discrimination, and 0.5 indicating no forecast skill.

The experimental, probabilistic hail forecasts for Day 1 and Day 2 full periods and the 4 h periods of 1800 UTC – 2200 UTC and 2200 UTC – 0200 UTC were verified using practically perfect forecasts, formed from the LSRs by applying a two-dimensional Gaussian smoother (Brooks et al. 2003) to reports within 40km of a 40 km-by-40 km grid box. For effective comparison against LSRs, similar practically perfect forecasts for MESH were produced by applying the same smoother to a separate set of derived severe hail events created by determining if MESH $\geq$ 29 mm (Cintineo et al. 2012) at each grid point. To avoid inclusion of spurious hourly MESH tracks, the

presence of at least one cloud-to-ground lightning flash detected by the National Lightning Detection Network (Cummins et al. 1998) within a 40-km radius of influence (ROI) was also required. A 40-km ROI neighborhood maximum was then applied to the final analyses. These quality control measures are similar in nature to those outlined in Melick et al. (2014). The components of the POD and the POFD were aggregated over the subdomains which had the highest severe weather potential for the given day across the experiment. In addition to the objective verification, participants commented on using MESH compared to LSRs for verifying probabilistic severe hail forecasts.

In addition to evaluation of severe convective hazards, objective evaluation of the ensemble mean quantitative precipitation forecasts (QPFs) also took place during SFE 2015. The ensemble means were computed using the probability matching technique (Ebert 2001) over a domain encompassing approximately the eastern two-thirds of the CONUS. This technique assumes that the best spatial representation of the precipitation field is given by the ensemble mean, and that the best probability density function of rain rates is given by the ensemble member QPFs of all *n* ensemble members.

Objective evaluation of these mean fields used the equitable threat score (ETS; Schaefer 1990) for four quantitative precipitation forecast (QPF) thresholds. This analysis encompassed five of the six ensembles within the experiment. The ETS measures the fraction of observed and/or forecast events that were correctly predicted, adjusted for correct yes forecasts associated with random chance. The ETS was calculated using contingency table elements computed every 3 h (from forecast hour 3 through forecast hour 36) from each grid point in the ensemble mean analysis domain,

using NCEP Stage IV precipitation data as truth. Forecasts and observations were regridded to a common 4 km grid prior to evaluation. An ETS of 1 is perfect, and a negative score represents no forecast skill. Probabilities of exceeding each threshold were computed by using the ratio of members that exceeded the specified threshold to the total number of members. These forecasts were evaluated using the ROC area, with probability thresholds ranging from 0.05-0.95 in increments of 0.05.

2.3 Preliminary Findings and Results

2.3.1 Evaluation of Short-Term Severe Forecasts

a. 1-h Total Severe Forecasts

Participants generally rated the hourly total severe forecasts highly (Fig. 2.4), with the updated afternoon forecasts garnering higher ratings than corresponding preliminary morning forecasts. These ratings encompass all individual hourly forecast ratings, and therefore include timing, placement, and magnitude error. Afternoon updates allowed forecasters to shift both the magnitude and the location of the probabilities, which produced mixed subjective results in SFE 2015. As stated by a 4 May participant: *"21-22Z improved from morning due to pulling the probabilities southward. However, an increase in probs was not appropriate.*" Though generally the afternoon updates occurred closer to the event, participants had difficulty forecasting on days when the convective mode was not yet apparent: "*Shorter lead time no help in anticipating messy storm evolution."* (4 May). On other days, there was some evidence that the convective mode was more apparent by the time of update issuance: "*Definitely an improvement from earlier. Convective mode was forecasted more accurately…*" (7

May). According to participants, the variability within the ensemble of participant forecasts mostly came from varying probability magnitudes, rather than varying locations. Some participants mention the difficulty of calibrating themselves to issue appropriate one-hour forecast probabilities as a potential cause for the variability. Also, the afternoon updates to the forecasts often narrowed the envelope of participant forecasts, as ongoing convection often removed the convective initiation forecast problem.

The mode forecasting problem was perhaps partially illustrated by the widening of the inter-quartile range (IQR) of the forecast ratings during the afternoon updates (Fig. 2.4). Difficulty in convective mode forecasting increases ratings' variability, as it is difficult to discern to the hour when and if individual supercells will grow upscale into an organized mesoscale convective system (MCS). SFE 2015 also encompassed many days with complex, mixed-mode convection, leading to difficulty of forecasting on an hourly basis. A 4 May participant reflected: "*There were also questions early about whether or not convection would occur across the entire frontal boundary, and this question did not seem fully resolved by the afternoon update*". Ultimately, overall afternoon forecast improvement was also subjectively noted: "*The afternoon updates were able to trim false alarm areas and refine the major regions for higher probabilities*" (3 June).

b. 4-h Individual Hazard Forecasts

Participants rated the preliminary 4 h individual hazard forecasts and the disaggregated first-guess hazard probabilities for 1800 UTC – 2200 UTC and 2200 UTC – 0200 UTC. During the earlier period, experimental forecasts and the first-guess

guidance were often rated similarly, with a median rating difference of 0 for tornadoes and wind, and +1 for hail on a scale from -3 to +3 (Fig. 2.5). While the evening period experimental forecasts improved upon the earlier, first-guess guidance, most of these ratings reflected marginal improvement (i.e., 0 to +1). Participant comments also supported only marginal improvement, partially due to having relatively little updated model information available: "*It was difficult to justify substantial updates to the afternoon forecast given a modicum of new information (i.e., the new information we had, small in nature compared to the larger set of data from the 0000 UTC cycle), did not warrant changes*" (26 May).

2.3.2 Comparison of Convection-Allowing Ensembles

SFE 2015 provided the unique opportunity to compare multiple CAM ensemble designs of varying complexity. 3-h ETS scores of QPF for each ensemble across the experiment were positive, indicating that all ensembles showed positive forecast skill at each threshold and hour. The lowest QPF threshold (Fig. 2.6a) overall had the highest ETS scores, with the SSEF 3DVAR performing better than all of the other ensembles at all forecast hours, though the difference typically only showed significance for the first few hours, and then again at approximately 24 h from initialization. At the highest precipitation threshold (Fig. 2.6d), ETS score difference among the ensembles was largest in the first twelve hours of the forecast period, and had essentially vanished by forecast hour 18. The ROC areas at each threshold (Fig. 2.7) show a similar trend at all precipitation thresholds, although the dominance of the SSEF 3DVAR is less pronounced. Interestingly, however, these ROC area differences between the SSEF 3DVAR and the other ensembles were often significant, particularly at the lower

thresholds. At the 0.10-in (Fig. 2.7a) and 0.25-in (Fig. 2.7b) exceedance thresholds, all ensembles (with the exception of the NCAR ensemble 3-h forecast) maintain skillful ROC areas. At higher thresholds the ensembles were less skillful, with the NCAR and SSEF EnKF ensembles having ROC areas less than 0.7 for most forecast hours when considering at least 0.50 in of precipitation (Fig. 2.7c) and only a handful of forecast hours for each ensemble system having skillful ROC areas at the 0.75 in threshold (Fig. 2.7d). EnKF analyzed reflectivity was noted to be too low, suggesting that there may have been an error in the EnKF configuration. Additionally, differing ensemble background and data assimilation may have affected the score; for example, only limited sets of conventional observations were assimilated in the SSEF EnKF compared to other ensembles. ROC areas tended to decrease later in the forecast period at low thresholds, and had a slight decrease in the middle of the forecast period at higher thresholds. Overall, the SSEF 3DVAR generally scored highest in the objective QPF metrics.

Subjectively, the participants' ratings of the Day 1 ensemble forecasts hourly maximum fields were again rather similar between ensembles (Fig. 2.8) excepting the SSEF EnKF, which was clearly the lowest-rated ensemble. The top-performing ensembles had a mean rating above six for the SFE, indicating that they provided useful severe weather guidance more often than not. As one participant commented, "*Mostly agreeing forecasts which all did reasonably well. Some modest discrimination based on amount of false alarm"* (14 May). Of the six CAM ensembles, the NSSL ensemble had a slightly higher mean and median rating than the other ensembles, which was significantly higher than the SSEF, SSEF EnKF, and the AFWA ensembles as

determined by a paired-sample t-test. The AFWA and NCAR ensembles had lower mean ratings than the SSEO, NSSL, and SSEF, but the difference did not reach significance. The only other significant difference between the mean ratings was that the SSEF EnKF was rated significantly lower than the NSSL, AFWA, and SSEO ensembles.

The Day 2 period (forecast hours 36-60) was less frequently objectively evaluated than the Day 1 period due to computational and data constraints, but the preliminary subjective results provide some insights. The AFWA and NCAR ensembles were more likely to have Day 2 forecasts rated similar to or better than their Day 1 ratings compared to the SSEF 3DVAR or the SSEF EnKF, as illustrated by the AFWA ensemble on 21 May 2015 (Fig. 2.9). For this case, the Day 1 forecasts (Fig. 2.9b-d) placed the majority of the UH-based ensemble neighborhood severe probabilities too far north and offshore, away from the verifying LSRs, whereas the Day 2 forecasts (Fig. 2.9e-g) encompass all LSRs. However, specificity of the Day 2 probabilities was also occasionally problematic : "*One issue with the longer range forecasts is that areas seem more joined rather than separate, which is reasonable (expected) but still makes it not as good as the day 1*" (3 June). Another participant stated that "*at least for this date the ensemble sets not assimilating radar data do better from the Day 2 forecast over the Day 1 forecast. I'm guessing this would be more likely for cases in which convection is ongoing and the non-radar assimilating ensembles serve more utility as a medium 24-48 range forecast.*" (14 May). Overall, the extended CAM ensembles provided useful Day 2 severe weather guidance, although poor depiction of Day 1 convection can detract from the Day 2 forecasts.

2.3.3 Comparison and Evaluation of Convection-Allowing Deterministic Models

a. Parallel Operational CAMs

The parallel versions of both the NAM nest and the HRRR showed subjective improvements over the operational versions, while the parallel and operational NAM runs were given similar subjective ratings (not shown). The parallel HRRR showed a reduction in the warm, dry, afternoon bias compared to the operational HRRR, resulting in improved convective initiation forecasts (e.g., Fig. 2.10). The parallel HRRR became operational on 23 August 2016, displacing the operational version used during the SFE 2015 timeframe, and the parallel NAM nest became operational on 8 September 2015.

b. Met Office UM

Participants compared the operational UM to the NSSL-WRF daily in SFE 2015. In addition to the 12-h to 36-h forecasts, the 1-h to 11-h forecasts were compared between the modelling systems to test which system better handled convective spin-up. Out of 133 responses, 55% rated the UM better than the NSSL-WRF, 23% rated the UM worse than the NSSL-WRF, and 22% said that they were the same in the first twelve hours of the forecast. These percentages were roughly the same when considering the 12- to 36-h period (132 total responses), with a slightly larger percentage (26% of responses) reporting that they were the same. Overall, the parallel UM (122 responses) was generally worse than (46%) or the same as (30%) the operational UM, and the 1.1-km UM (104 responses) was typically the same (43%) or worse (32%) than the 2.2-km.

Sounding comparisons between the NSSL-WRF and the operational UM (Fig. 2.11) often showed striking differences. Throughout SFE 2015, capping inversions in

the operational UM were consistently more sharply defined than in the NSSL-WRF, more closely matching the observational soundings and consistent with the examples shown in Kain et al. (2016). Out of 89 total participant responses, 60 expressed that the UM soundings were better than the NSSL-WRF, while 19 felt the two were the same. Only 10 responses rated the NSSL-WRF soundings better than the UM. The structure and sharpness of the strong capping inversions were subjectively noted by participants as much better depicted in the UM than the NSSL-WRF: "*UKMET is better. Depicts inversion temperature profile perfectly. This is the biggest difference.*" (2 June). Although the UKMET has nearly double the vertical levels of the NSSL-WRF, Kain et al. (2016) state that merely increasing the vertical resolution of the NSSL-WRF does not negate this tendency.

c. MPAS

While no formal evaluation of the MPAS forecasts took place, the guidance was examined on a daily basis and used during the forecasting process. Two cases where useful convective-scale guidance to Day 3 and beyond are presented here, as a preliminary indication of the usefulness of MPAS in forecasting severe convection at longer time scales than most current convection-allowing guidance. Both days provided similar synoptic patterns conducive to a severe weather outbreak across the southern plains, with the eventual outcome heavily dependent on the presence of morning convection, related to the strength of the capping inversion.

Several days in advance of 9 May 2015, the SPC Day 3 convective outlook outlined an area across Oklahoma and Kansas as having a moderate risk for severe storms. In reality, during the late morning of 9 May, strong forcing for ascent combined

with a weak capping inversion led to widespread convection and associated cloud cover across much of western Oklahoma and Kansas, inhibiting afternoon destabilization. The early convection led to minimal CAPE (<1000 J/kg) across much of Oklahoma and Kansas (Fig. 2.12a). Although severe storms did occur from Texas into western Kansas, because of the early storms the event as a whole ended up being less significant than what some earlier model guidance had suggested. While forecasting a synoptic-scale pattern favorable for widespread severe weather 3 days in advance of 9 May, the MPAS forecasts also indicated that widespread convection would develop early in the day on 9 May. The impact of this early convection manifested in reduced CAPE simulated across Oklahoma and Kansas (Fig. 2.12c). Thus, the scenario depicted by MPAS 3 days in advance was consistent with what occurred.

The second case with a favorable synoptic pattern for severe weather in which MPAS provided useful extended range guidance was on 16 May 2015. Similar to 9 May, the extent and intensity of the severe weather threat was uncertain, because it was not clear how much early convection would inhibit heating and destabilization in the warm sector. Despite a shallow layer of clouds, a lack of widespread early convection allowed enough destabilization (Fig. 2.12b) to support a significant severe weather event and several long-lived tornadic supercells across the Texas Panhandle, Oklahoma, and Missouri. The forecasts from MPAS 3 days in advance were consistent with this scenario, maintaining CAPE through early convection (Fig. 2.12d) and matching quite well the observed range of CAPE. Furthermore, the MPAS forecasts depicted intense supercells forming in the warm sector around 2100 UTC beginning with the 93 h, Day 4 forecast (Fig. 2.13e) and continuing through the Day 3 (Fig. 2.13d), Day 2 (Fig. 2.13c)

and Day 1 (Fig. 2.13b) forecasts. The location of the storms was initially too far east compared to observations (Fig. 2.13a). Additionally, the timing of upscale growth was also well-depicted as far as four days in advance (Fig. 2.13k), clearly showing the squall line over central Oklahoma at 0355 UTC (Fig. 2.13g). The overall forecast scenario corresponded well to the observations, particularly regarding the mode and timing of mode evolution and again would have provided useful extended-range convective scale guidance to forecasters.

2.3.4 Evaluation of New Diagnostics

a. Hail Diagnostics

Three days of WRF-HAILCAST were formally evaluated in SFE 2015, precluding robust conclusions. Compatibility issues resulted in the Thompson method only being available in the NCAR ensemble, and thus a direct comparison to the WRF-HAILCAST implemented in the SSEF system was impossible. However, participants unanimously agreed that across the three cases the hail size forecasts provided additional useful information relative to more commonly used hourly maximum fields such as UH, prompting the inclusion of the new hail diagnostics in future SFEs.

b. Tornado Diagnostics

The distributions of subjective ratings assigned to the 24 h tornado probabilities by the individual participants suggest that incorporating environmental information results in an improved forecast over solely using UH (Fig. 2.14). None of the environmental filters (LCL, CAPE, STP, or combined) clearly stood out as the best method; however, they all generally improved upon the UH-only guidance. Participants often noted that the incorporation of environmental information helped focus the area of

interest and reduce false alarm. However, they often felt that the probabilities were too high on a given day to directly translate into the current operational convective outlook categories (i.e., tornado probabilities of 30% on a day that SPC forecasters would not consider a "moderate" risk given the environment).

2.3.5 Hail Verification Comparisons

When participants evaluated MESH as verification for probabilistic severe hail forecasts, rather than LSRs, responses were generally positive. A participant said on 4 May: "*Assuming that MESH is reasonably representative of what actually occurred, it definitely helps fill in areas between local storm reports*." Many participants commented that the MESH provided verification in low population density areas such as eastern Colorado (Fig. 2.15a, b), where obtaining even a single report to verify a warning may be difficult. Participants "*liked the spatial and temporal details much better*" (7 May), and noted that in these locations when reports did occur, MESH often also diagnosed large hail (Fig. 2.15c, d). However, participants were unsure of directly comparing LSRs and MESH, stating: "*…Hard to say how well it does in verifying when not comparing hail sizes in MESH to actual LSR observed hail sizes…*". Ortega et al. (2009) performed a concentrated verification of MESH tracks, but a larger-scale verification database does not yet exist. Wilson et al. (2009) found that MESH performs best at values greater than 19 mm, which would include all severe hail, although they advise against using MESH alone as a form of synthetic verification; MESH has also been found to overforecast hail size (Wilson et al. 2009, Cintineo et al. 2012). Cintineo et al. (2012) find that Heidke skill scores are maximized in a comparison of MESH to high-resolution ground-truth reports of severe hail when a threshold of 29 mm is used.

Further, Melick et al. (2014) has suggested that MESH tracks can be useful as an independent dataset to supplement hail LSRs. Consequently, the positive response from participants recommends an objective look at MESH verification over the daily subdomains.

Objective verification of the experimental hail forecasts with practically perfect forecasts generated by MESH (17 cases) and LSR (23 cases) at different periods via ROC area (Fig 2.15e) showed that whether MESH or LSR verified the forecasts best was dependent on the time period examined. Looking at the full period Day 1 forecasts, LSRs had a higher POD and approximately the same POFD as the MESH, leading to a higher ROC area. Conversely, the Day 2 full period forecasts show both higher POD and higher POFD when verified using the MESH, rather than the LSRs. The four hour outlooks generally performed better than the daily outlooks in both verifications. This is particularly evident in the 22-02Z time frame, when convective initiation was less of a forecast problem. These results suggest that the hail forecasts are typically able to distinguish the area of hail. However, ROC areas do not take into consideration the reliability of the forecasts, which was a large factor in participants' subjective ratings of the verification methods. Indeed, participants noted the higher "practically perfect" probabilities were often generated using the MESH tracks (Fig. 2.15d) compared to the LSRs (Fig. 2.15b): "*Practically perfect probabilities from MESH seemed overestimated compared to the report probabilities*" (19 May). This may be because participants aren't used to seeing MESH-derived practically perfect probabilities. However, these higher probabilities did not seem to dampen the participants' enthusiasm for using MESH as a verification metric. One participant on 27 May stated, "*Even if a slight*

oververification [sic] *given its construction, the use of MESH for verification seems to be an improvement on this day*".

2.4 Summary and Discussion

Overall, SFE 2015 succeeded in testing new forecast products and modelling systems to address relevant issues in predicting hazardous convective weather. The sheer volume of daily numerical weather guidance examined throughout SFE 2015 was unprecedented, and the real-time, operational nature of the experiment emphasized the need for tools that forecasters can use to summarize large volumes of information when forecasting severe convective weather. The innovative nature of the experiment gave participants access to cutting-edge, operationally-relevant research from multiple institutions, evaluating six CAM ensembles, three deterministic Met Office CAMs, a deterministic CAM with forecasts extending out to five days (MPAS), parallel versions of current operational models, and new diagnostic techniques for hail size and tornado occurrence. The experiment found that parallel versions of the HRRR and the NAM Nest improved upon the current operational versions, providing strong evidence to support implementation of the experimental parallel modeling systems. Additionally, CAMs were found useful when issuing Day 2 forecasts, providing mode insight for medium-range severe convective forecasts. Day 2 forecasts occasionally rated more highly than the corresponding Day 1 forecasts, although participants noted that Day 2 forecasts started from ensembles assimilating radar data can be affected if the Day 1 convection is poorly handled, essentially relying on them as a medium-range forecast. The SFE also helped to determine that applying environmental filters to explicit UH

diagnostics improved guidance for probabilistic tornado forecasting compared to using UH only with the NSSL-WRF ensemble.

Increased participant interaction was a key component of SFE 2015. Using laptops in the experiment allowed participants to submit individual, rather than group consensus, evaluations and allowed for personalized feedback. The usage of the PHI tool on individual laptops allowed for more participant engagement, as they drew their own short-term forecasts. These short-term forecasts performed well objectively and subjectively, suggesting that moving these products into operations is feasible, fulfilling an Operational Product and Service Improvement goal. These forecasts benefited greatly from the availability of the CAM guidance, particularly the hourly forecasts of total severe. To make such reliable forecasts without CAM guidance would have been difficult.

Annual SFEs in the HWT have a long history of impacting National Weather Service operations, but oftentimes one has to consider a multi-year period to get a full measure of these impacts. For example, SFE 2010 contained one CAM ensemble provided by CAPS, and was just beginning to evaluate hourly maximum fields such as UH and simulated 1 km AGL reflectivity. These fields tested in SFE 2010 are now considered key output parameters in operational CAMs and are used worldwide, showing how the SFEs succeed in research-to-operations efforts. Since that SFE, grid spacing has decreased, and the number and availability of CAM ensembles has greatly increased. SFE 2015 allowed its participants to study the behavior of these ensembles, bolstering their knowledge of the latest forecasting techniques. SFE 2015 also provided researchers with knowledge of how the many NWP guidance options provided to

forecasters are perceived, in addition to information about how comparable these ensembles are at the height of the spring convective season.

SFE 2015 highlighted areas requiring future study through verification efforts in conjunction with the NOAA Applied Science Activities goals. Participant comments on using MESH in addition to LSRs for hail verification suggest that MESH tracks may be a good future verification source, albeit after a larger comparison database is compiled between MESH and LSRs. The tendency of hail guidance to either overforecast (WRF-HAILCAST) or underforecast (Thompson) hail sizes, and the overforecasting tendency of the tornado probabilities noted by participants highlights that more work is needed regarding individual hazard diagnostics. Future work focusing on individual hazard diagnostics is planned to compare the diagnostics between ensembles and to current SPC forecasts for individual hazards. Finally, the striking difference between the Met Office CAMs and the NSSL-WRF in representing strong vertical gradients in temperature and moisture near capping inversions demonstrates that work is still needed to hone the accuracy of vertical profiles.

With SFE 2015 complete, future SFEs can build off the lessons learned therein. Surprisingly, though the six ensembles in SFE 2015 were configured differently, the ensembles' performance according to both objective and subjective measures was quite similar. This result led to a focus in SFE 2016 on uncovering how differences in ensemble configuration affect model performance with regards to severe convective weather using the recently developed Community Leveraged Unified Ensemble (CLUE; Clark et al. 2016). The CLUE consisted of 65 members provided by a number of institutions, all of which had the same domain, grid-spacing, and output fields. These

members were divided into a number of sub-experiments for directly comparing configuration strategies (i.e., multi-core vs. single core, multi-physics vs. single physics, 3DVAR vs. EnKF, ensemble size sensitivity). By minimizing as many differences as possible between the members, it is hoped that CLUE will help inform key ensemble configuration decisions, providing valuable guidance for operational CAM ensemble design.

Acknowledgments

First, the authors would like to thank all of the participants and contributors to the annual Spring Forecasting Experiments, whose work, insight, and excitement make the SFEs possible. This material is based upon work supported by the National Science Foundation Graduate Research Fellowship under Grant No. DGE-1102691, Project #A00-4125. AJC, KK, JC, CJM, CDK and WL were provided support by NOAA/Office of Oceanic and Atmospheric Research under NOAA-University of Oklahoma Cooperative Agreement #NA11OAR4320072, US Department of Commerce. AJC also received support from a Presidential Early Career Award for Scientists and Engineers. This work used the Extreme Science and Engineering Discovery Environment (XSEDE), which is supported by National Science Foundation grant number ACI-1053575. MX, KB, and FK received support from NOAA CSTAR grant NWSPO-2010-201696. Thanks also go to Scott Rentschler for providing information regarding the AFWA ensemble. Finally, the authors would like to thank two anonymous reviewers and Philip N. Schumacher for their careful consideration and comments, which greatly improved the clarity of the manuscript.

Tables

Table 2.1 NSSL-WRF ensemble specifications. All members use the WRF single-moment microphysics (WSM6; Hong and Lim 2006), the Mellor-Yamada-Janjić (MYJ; Mellor and Yamada 1982; Janjić 1994, 2002) planetary boundary layer (PBL) scheme, and the Noah (Chen and Dudhia 2001) land surface model (LSM). For radiation, all members use the Rapid Radiative Transfer Model (RRTM; (Mlawer et al. 1997; Iacono et al. 2008) longwave radiation and Dudhia (Dudhia 1989) shortwave radiation schemes.

Ensemble Member	Vertical Levels	Initial Conditions	Lateral Boundary Conditions	Microphysics	PBL
Cn	35	00Z NAM	00Z NAM	WSM6	MYJ
2	35	00Z GFS	00Z GFS	WSM6	MYJ
3	35	21Z em_ctl	21Z em_ctl	WSM6	MYJ
4	35	21Z nmb_ctl	21Z nmb_ctl	WSM6	MYJ
5	35	21Z nmb_p1	21Z nmb_p1	WSM6	MYJ
6	35	21Z nmm_ctl	21Z nmm_ctl	WSM6	MYJ
7	35	21Z nmm_n1	21Z nmm_n1	WSM6	MYJ
8	35	21Z nmm_p1	21Z nmm_p1	WSM6	MYJ
9	35	21Z nmb_n1	21Z nmb_n1	WSM6	MYJ
10	35	21Z nmb_p2	21Z nmb_p2	WSM6	MYJ

Table 2.2 SSEF ensemble specifications. All members use RRTMG radiation schemes. Microphysics schemes used include Thompson (Thompson et al. 2004b), Predicted Particle Properties (P3; Morrison and Milbrandt 2015), Milbrandt and Yau (M-Y; Milbrandt and Yau 2005), and Morrison (Morrison and Pinto 2005, 2006). Member 18 uses microphysics with two-category ice; all other P3 members use one-category ice. Planetary boundary layer schemes not previously defined include Yonsei University (YSU; Hong et al. 2006), Thompson-modified YSU (YSU-T), and Mellor-Yamada-Nakanishi-Niino (MYNN; Nakanishi and Niino 2004, 2006). Member 16 (Thompson ICLOUD=3) accounts for the sub-grid scale clouds in the Global RRTM (RRTMG) radiation scheme based on research by G. Thompson. Italicized members compose the HWT baseline SSEF.

Ensemble Member	Vertical Levels	Initial Conditions	Lateral Boundary Conditions	Microphysics	PBL
Cn	51	*00 UTC ARPSa*	*00 UTC NAMf*	*Thompson*	*MYJ*
c0	51	00 UTC ARPSa	00 UTC NAMf	Thompson	MYJ
m3	51	*cn + nmmb-p2_pert*	*21 UTC SREF nmmb-p2*	*P3*	*MYNN*
m4	51	*cn + nmmb-n2_pert*	*21 UTC SREF nmmb-n2*	*M-Y*	*YSU*
m5	51	*cn + nmm-p1_pert*	*21 UTC SREF nmm-p1*	*Morrison*	*MYNN*
m6	51	*cn + nmmb-n1_pert*	*21 UTC SREF nmmb-n1*	*M-Y*	*MYJ*
m7	51	*cn +nmmb-p1_pert*	*21 UTC SREF nmmb-p1*	*P3*	*YSU*
m8	51	*cn + em-n1_pert*	*21 UTC SREF em-n1*	*P3*	*MYJ*
m9	51	*cn + em-p2_pert*	*21 UTC SREF em-p2*	*M-Y*	*MYNN*
m10	51	*cn + nmmb-n3_pert*	*21 UTC SREF nmmb-n3*	*Morrison*	*YSU*
m11	51	*cn + nmmb-p3_pert*	*21 UTC SREF nmmb-p3*	*Thompson*	*YSU*
m12	51	*cn + nmm-n3_pert*	*21 UTC SREF nmm-n3*	*Thompson*	*MYNN*
m13	51	*cn + nmm-p2_pert*	*21 UTC SREF nmm-p2*	*Morrison*	*MJ*
m14	51	00 UTC ARPSa	00 UTC NAMf	Thompson	MYNN
m15	51	00 UTC ARPSa	00 UTC NAMf	Thompson	YSU-T
m16	51	00 UTC ARPSa	00 UTC NAMf	Thompson ICLOUD=3	YSU-T
m17	51	00 UTC ARPSa	00 UTC NAMf	M-Y	MYJ
m18	51	00 UTC ARPSa	00 UTC NAMf	P3-cat2	MYJ
m19	51	00 UTC ARPSa	00 UTC NAMf	P3	MYJ
m20	51	00 UTC ARPSa	00 UTC NAMf	Morrison	MYJ

Table 2.3 SSEF EnKF ensemble specifications.

Ensemble Member	Vertical Levels	Initial Conditions	Lateral Boundary Conditions	Microphysics	PBL
enkf_cn	51	enk_m1a	00 UTC NAMf	Thompson	MYJ
enkf_m6	51	enk_m2a	21 UTC SREF nmmb-n1	M-Y	MYJ
enkf_m9	51	enk_m6a	21 UTC SREF em-p2	M-Y	MYNN
enkf_m10	51	enk_m8a	21 UTC SREF nmmb-n3	Morrison	YSU
enkf_m5	51	enk_m10a	21 UTC SREF nmm-p1	Morrison	MYNN
enkf_m4	51	enk_m12a	21 UTC SREF nmmb-n2	M-Y	YSU
enkf_m3	51	enk_m17a	21 UTC SREF nmmb-p2	P3	MYNN
enkf_m8	51	enk_m23a	21 UTC SREF em-n1	P3	MYJ
enkf_m7	51	enk_m26a	21 UTC SREF nmmb-p1	P3	YSU
enkf_m12	51	enk_m37a	21 UTC SREF nmm-n3	Thompson	MYNN
enkf_m11	51	enk_m39a	21 UTC SREF nmmb-p3	Thompson	YSU
enkf_mn_thom	51	enfamean_thom	00 UTC NAMf	Thompson	MYJ
enkf_mn_wsm6	51	enfamean_wdm6	00 UTC NAMf	WSM6	MYJ
enkf_3dvar_thom	51	3dvar_thom	00 UTC NAMf	Thompson	MYJ
enkf_3dvar_wsm6	51	3dvar_wdm6	00 UTC NAMf	WSM6	MYJ

Table 2.4 SSEO specifications as of 12 August 2014

Ensemble Member	Vertical Levels	Initial Conditions	Lateral Boundary Conditions	Microphysics	PBL	Grid Spacing
NSSL WRF-ARW	35	NAM	NAM	WSM6	MYJ	4 km
EMC HRW WRF-ARW	40	RAP	GFS	WSM6	YSU	4.2 km
EMC HRW WRF-ARW; 12-h time lag	40	RAP	GFS	WSM6	YSU	4.2 km
EMC HRW NMMB	40	RAP	GFS	Ferrier updated	MYJ	3.6 km
EMC HRW NMMB; 12-h time lag	40	RAP	GFS	Ferrier updated	MYJ	3.6 km
EMC CONUS WRF-NMM	35	NAM	NAM	Ferrier	MYJ	4 km
EMC CONUS NAM NEST	60	NAM	NAM	Ferrier-Aligo	MYJ	4 km

Table 2.5 AFWA ensemble specifications. Land initial conditions for each member are from the NASA Goddard Space Flight Center Land Information System (LIS; Kumar et al. 2006, 2007). The PBL schemes include the BouLac (Bougeault and Lacarrère 1989) and the updated asymmetric convective model (ACM2). Members use either the Noah or the Rapid Update Cycle (RUC; Smirnova et al. 1997, 2000, 2015) land surface model. Microphysics schemes include the WRF double-moment microphysics (WDM6; Lim and Hong 2010) and the WRF 5-class single-moment microphysics (WSM5; Hong et al. 2004).

Ensemble Member	Vertical Levels	Initial Conditions	Lateral Boundary Conditions	Microphysics	PBL
1	27	UM	UM	WSM5	YSU
2	27	GFS	GFS	Morrison	BouLac
3	24	GEM	GEM	WDM6	YSU
4	21	GEM	GEM	Ferrier	BouLac
5	21	UM	UM	WDM6	ACM2
6	24	GFS	GFS	Thompson	ACM2
7	24	GEM	GEM	Morrison	YSU
8	24	GFS	GFS	Ferrier	YSU
9	27	UM	UM	Thompson	ACM2
10	21	GFS	GFS	WSM5	ACM2

Table 2.6 Parallel Operational CAM specifications. Initial and lateral boundary include the GFS and the Rapid Refresh (RAP) model (Benjamin et al. 2016).

Model	Vertical Levels	Initial Conditions	Lateral Boundary Conditions	Microphysics	PBL
ESRL/GSD HRRR (op)	51	RAP v2	RAP v2	Modified Thompson	MYNN
ESRL/GSD HRRR (parallel)	51	RAP v3	RAP v3	Thompson-Eidhammer	Modified MYNN
EMC HRW WRF-ARW (op)	40	RAP v3	GFS	WSM6	YSU
EMC HRW WRF-ARW (parallel)	50	RAP v3	GFS	Modified WSM6	YSU
EMC HRW WRF-NMMB (op)	40	RAP v3	GFS	Ferrier Updated	MYJ
EMC HRW WRF-NMMB (parallel)	50	RAP v3	GFS	Ferrier Updated	MYJ
EMC CONUS NAM Nest (op)	60	GFS	NAM	Ferrier-Aligo	MYJ
EMC CONUS NAM Nest (parallel)	60	GFS	NAM	Ferrier-Aligo	MYJ

Table 2.7 Daily activity schedule in local (CDT) time

<table>
<tr><td colspan="2">0800 – 0845: Evaluation of Experimental Forecasts & Guidance
Subjective rating relative to radar evolution/characteristics, warnings, and preliminary reports and objective verification using preliminary reports and MESH</td></tr>
<tr><td>Individual Hazards Desk</td><td>Total Severe Desk</td></tr>
<tr><td>• Day 1 & 2 full-period probabilistic forecasts of tornado, wind, and hail
• Day 1 4-h period forecasts and guidance for tornado, wind, and hail</td><td>• Days 1, 2, & 3 full-period probabilistic forecast of total severe
• Day 1 1-h period forecasts and guidance for total severe</td></tr>
<tr><td colspan="2">0845 – 1115: Day 1 Convective Outlook Generation
Hand analysis of 12Z upper-air maps and surface charts</td></tr>
<tr><td>• Day 1 full-period probabilistic forecasts of tornado, wind, and hail valid 16-12Z over mesoscale area of interest
• Day 1 4-h probabilistic forecasts of tornado, wind, and hail valid 18-22 and 22-02Z*</td><td>• Day 1 full-period probabilistic forecast of total severe valid 16-12Z over mesoscale area of interest
• Day 1 1-h probabilistic forecasts of total severe valid 18-00Z*</td></tr>
<tr><td colspan="2">1115 – 1130: Break
Prepare for map discussion and discuss relationship/translation from probabilities to watch</td></tr>
<tr><td colspan="2">1130 – 1200: Map Discussion
Overview and discussion of today’s forecast challenges and products
Highlight interesting findings from previous days</td></tr>
<tr><td colspan="2">1200 – 1300: Lunch
Brief EWP participants at 1245</td></tr>
<tr><td colspan="2">1300 – 1400: Day 2 Convective Outlook Generation</td></tr>
<tr><td>• Day 2 full-period probabilistic forecasts of tornado, wind, and hail valid 12-12Z over mesoscale area of interest</td><td>• Day 2 or Day 3 full-period probabilistic forecasts of total severe valid 12-12Z over mesoscale area of interest</td></tr>
<tr><td colspan="2">1400 – 1500: Scientific Evaluations</td></tr>
<tr><td>• Convection-allowing ensemble comparison (reflectivity and hourly maximum fields): SSEO, AFWA, NSSL, SSEF, SSEF EnKF, NCAR EnKF.
• EMC parallel CAM comparison (reflectivity): NAM Nest, HiResW, HRRR</td><td>• Met Office CAMs: vertical resolution
• SSEF 3DVar vs. EnKF Comparison: impact on first few hours of control forecast
• Model forecasts of explicit hail size: HAILCAST, Thompson
• MPAS</td></tr>
<tr><td colspan="2">1500 – 1600: Short-term Outlook</td></tr>
<tr><td>• Update 4-h probabilistic forecasts of tornado, wind, and hail valid 22-02Z*
• Generate 1-h probabilistic forecasts of tornado valid 22-02Z</td><td>• Update and generate 1-h probabilistic forecasts of total severe valid 21-02Z*</td></tr>
<tr><td colspan="2">* Denotes forecasts also made by participants using the PHI tool on laptops.</td></tr>
</table>

Figures

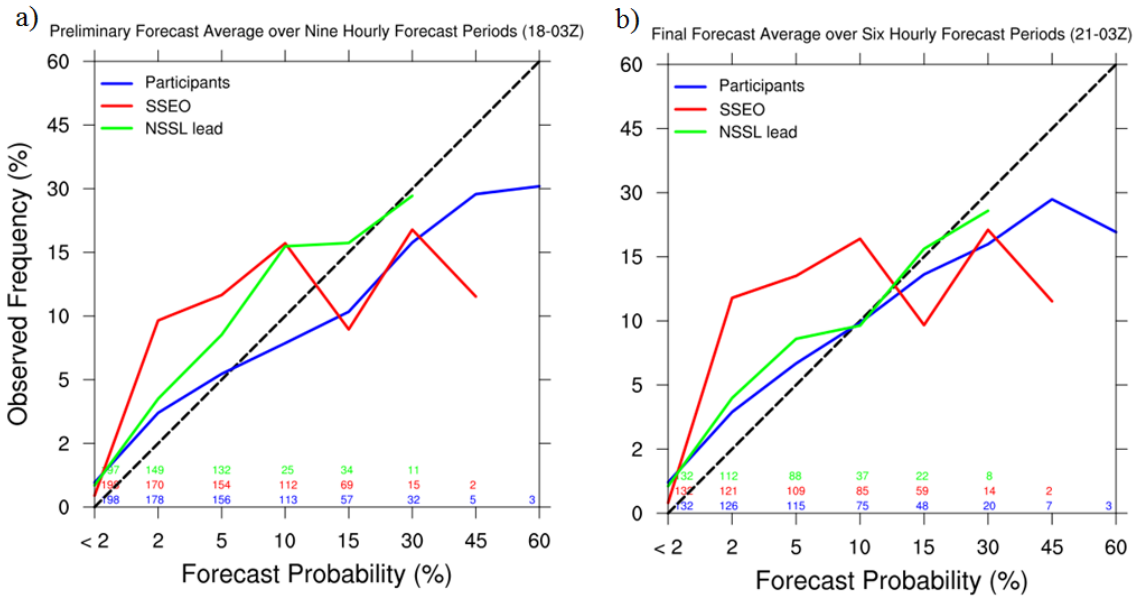

Figure 2.1 Reliability diagrams generated for SFE 2014 hourly probabilistic forecasts for (a) the nine initial hourly forecasts and (b) the six afternoon updates. The black dashed line indicates perfect reliability, and the colored numbers over the x-axis correspond to the number of forecasts with at least one forecast of that probability magnitude. From Coniglio et al. (2014).

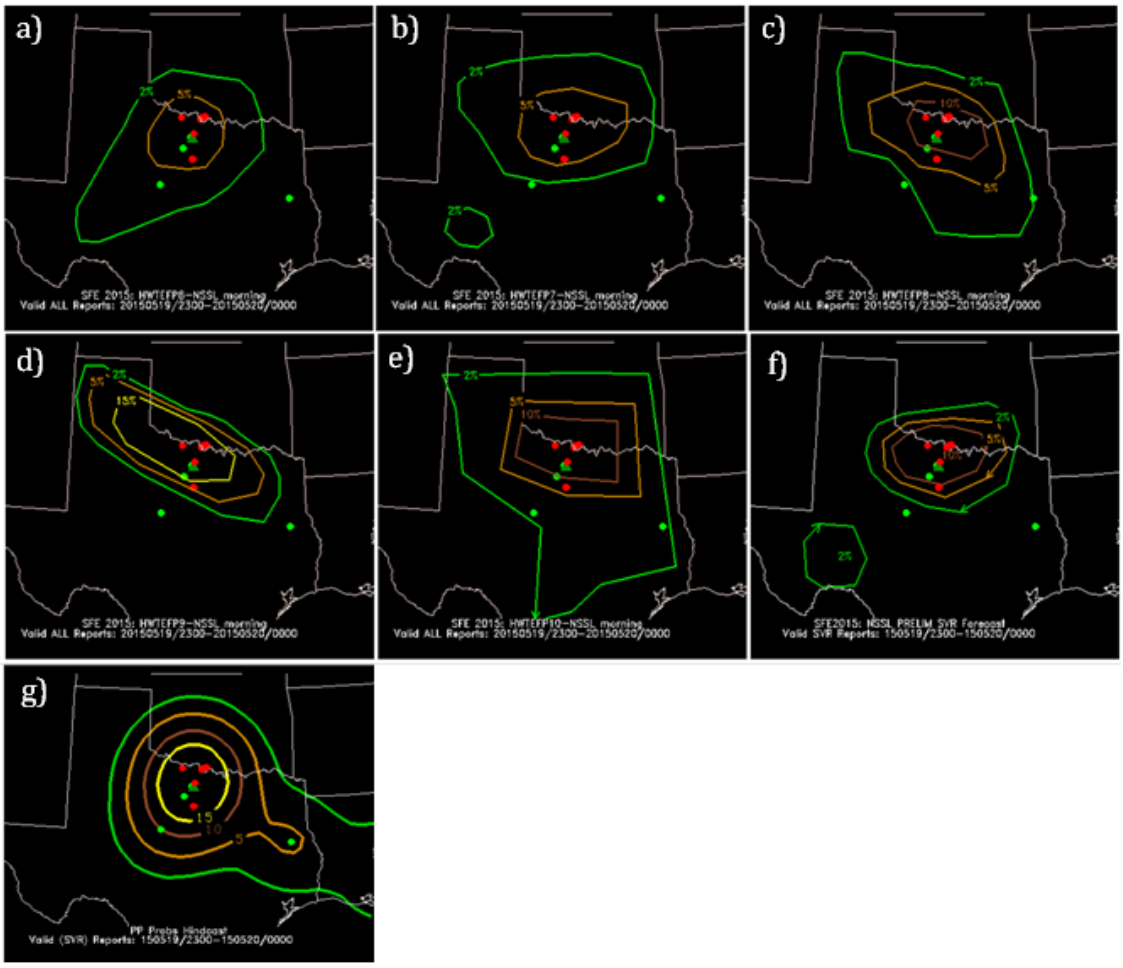

Figure 2.2 (a-e) Five participant forecasts, (f) one SPC forecaster forecast, and (g) the practically perfect forecast valid 2300 UTC 19 May 2015 – 0000 UTC 20 May 2015. Probabilistic contours indicate the likelihood of any type of severe weather (tornado, wind, or hail) during the forecast period. Overlaid red dots are tornado LSRs, green dots are hail LSRs, and dark green triangles are significant hail (hail diameter ≥ 2 inches) LSRs.

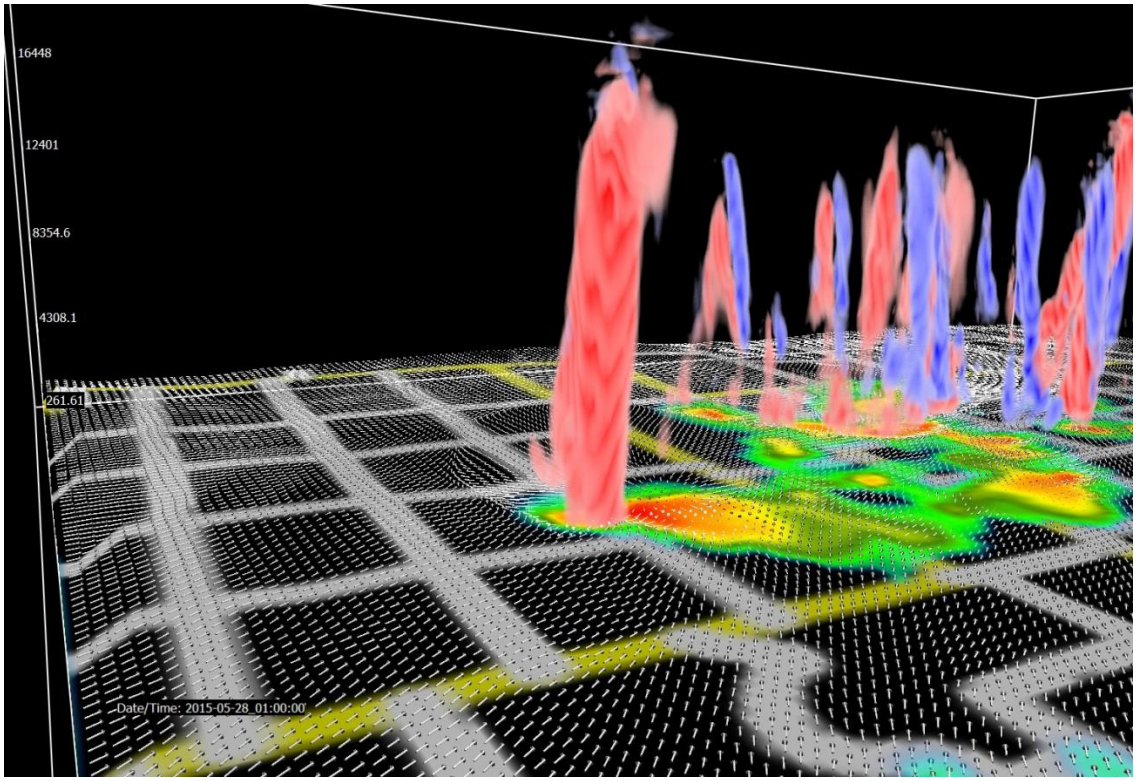

Figure 2.3 3D visualization of forecasted storms valid 0100 UTC on 28 May 2015, looking to the northwest from western Oklahoma and showing near-surface wind vectors (white), near surface radar reflectivity (2D color shaded field), and UH (red positive, blue negative). County boundaries are in white and state boundaries are in yellow.

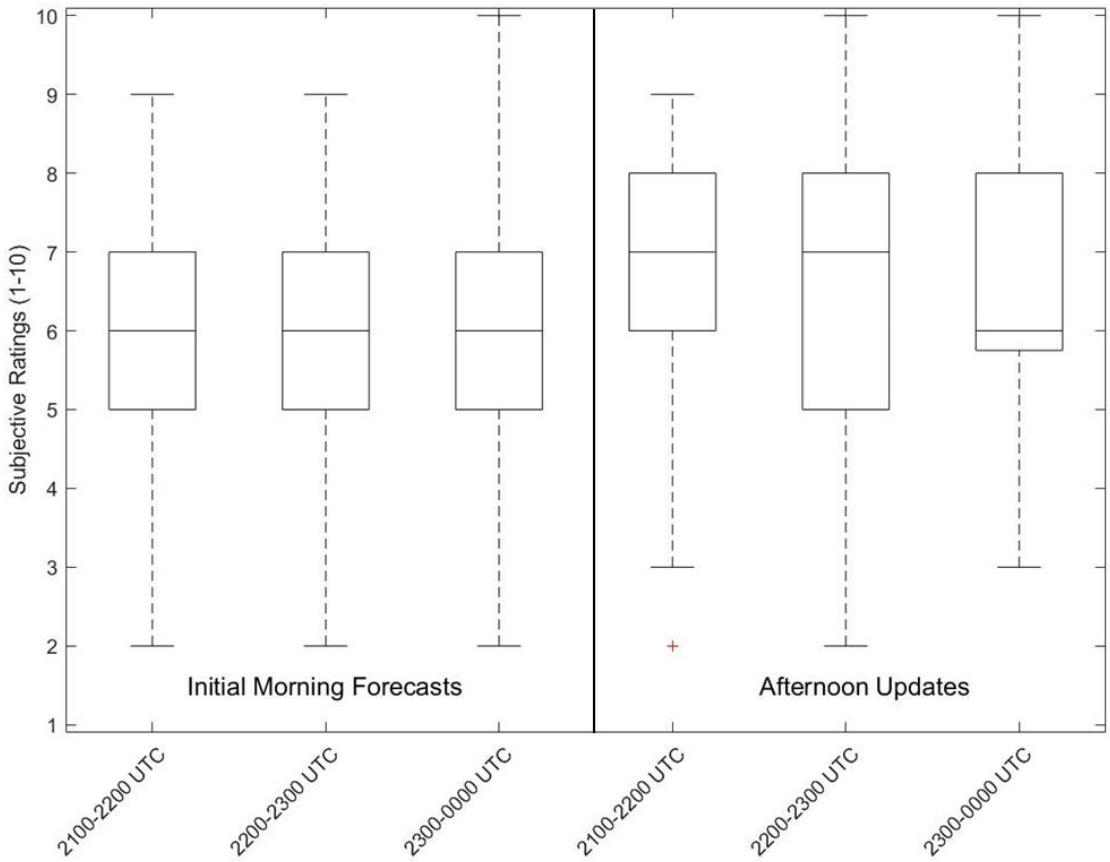

Figure 2.4 Distribution of subjective ratings (1 to 10) for the preliminary hourly experimental forecasts (left; 2100-0000 UTC) issued at 1600 UTC compared to the final experimental forecasts (right; valid 2100-0000 UTC) issued at 2100 UTC. The boxes comprise the interquartile range of the distributions and the whiskers extend to the 10th and 90th percentiles. Outliers are indicated by red plus symbols.

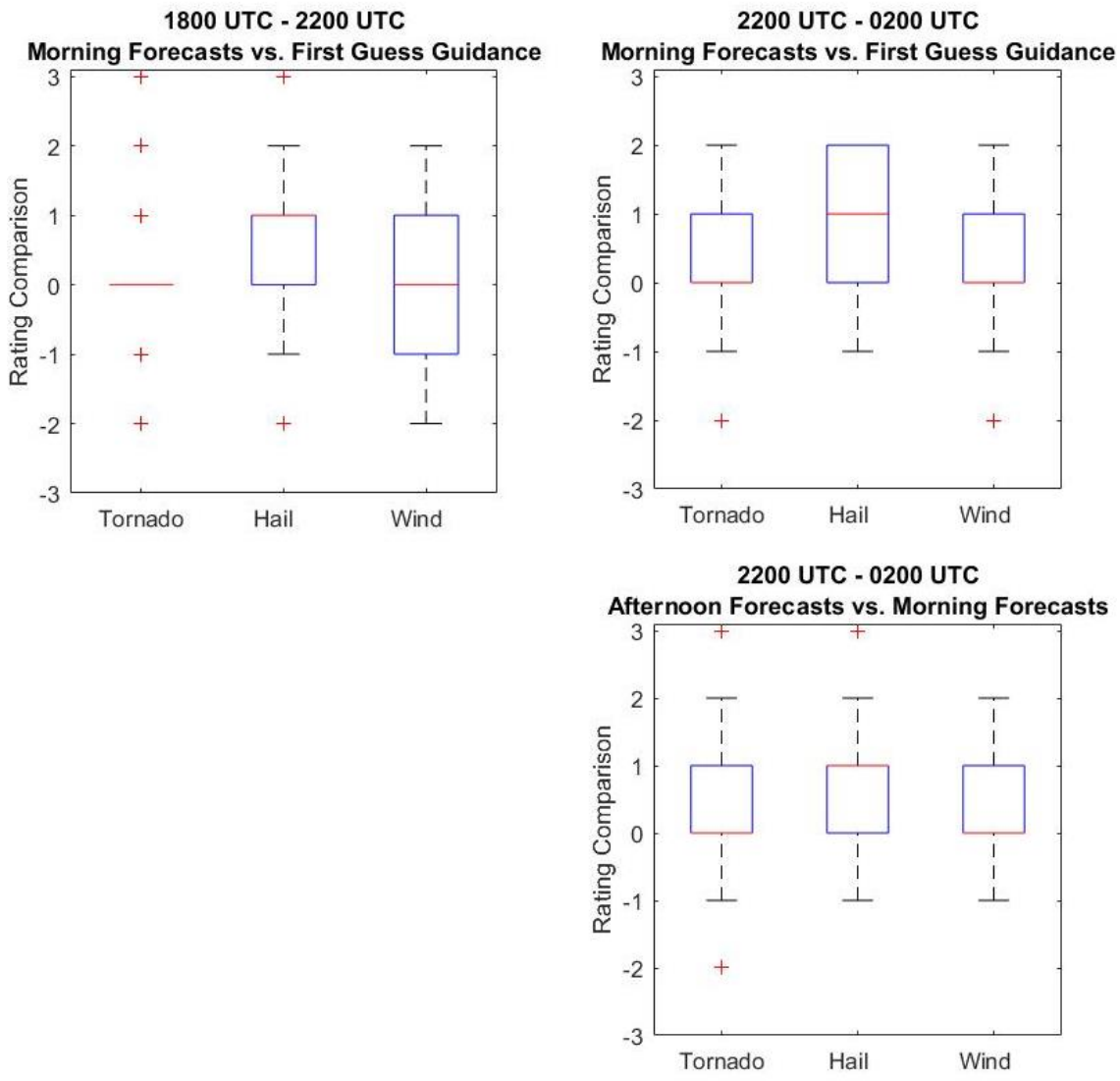

Figure 2.5 As in Fig. 2.4, except for the distribution of subjective ratings (-3 to +3) of the experimental forecasts compared to the first-guess guidance for tornado, hail, and wind during the 1800-2200 UTC (left) and 2200-0200 UTC (right) periods. The top row is the initial morning forecasts, and the bottom row is the afternoon update, which only took place for the 2200 UTC – 0200 UTC period.

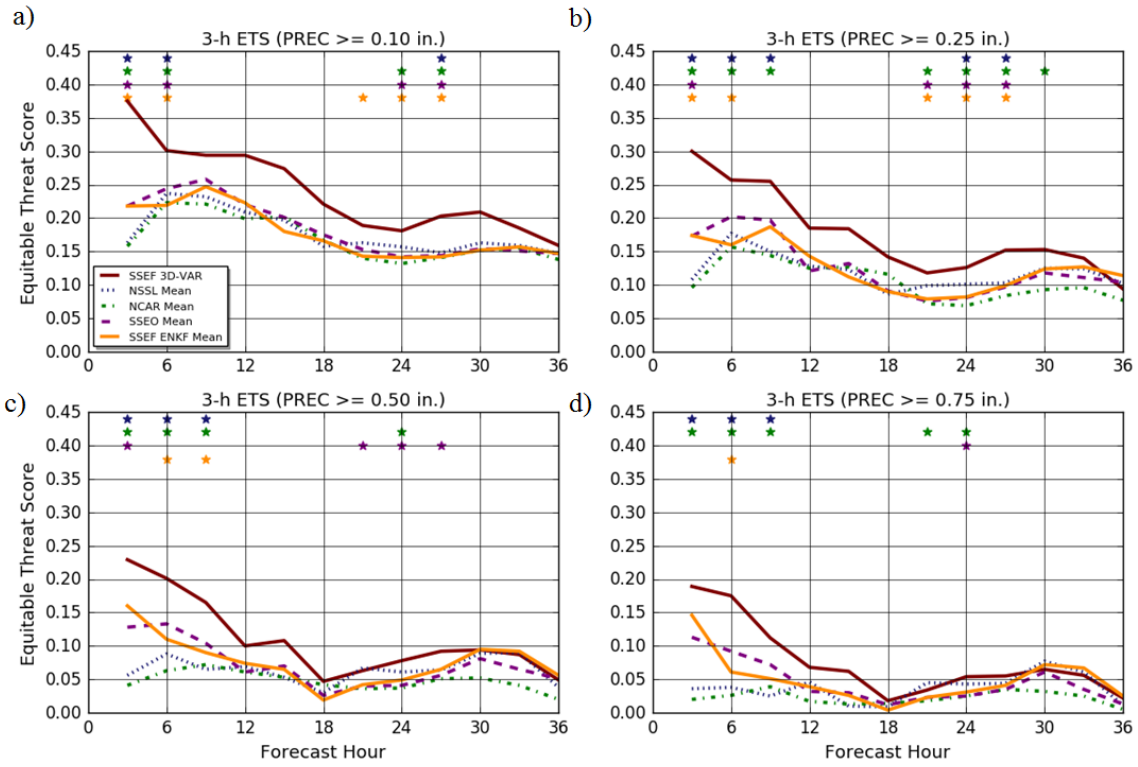

Figure 2.6 ETS scores for 3 h ensemble probability-matched mean fields at four QPF exceedance thresholds: (a) 0.10 in; (b) 0.25 in; (c) 0.50 in; and (d) 0.75 in. Different colored lines represent the different models, and colored stars indicate a significant difference between the SSEF 3DVAR ensemble and the ensemble corresponding to that color.

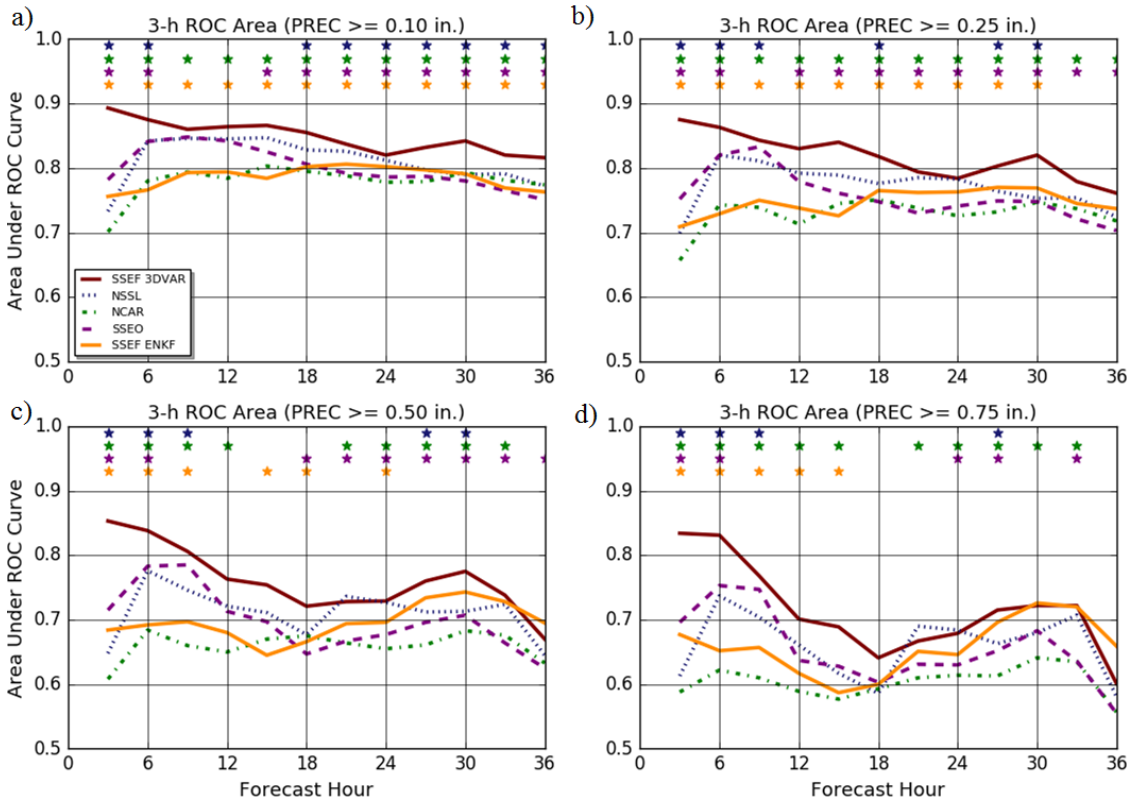

Figure 2.7 ROC area scores for 3 h ensemble probability-matched mean fields at four QPF exceedance thresholds: (a) 0.10 in; (b) 0.25 in; (c) 0.50 in; and (d) 0.75 in. Different colored lines represent the different models, and colored stars indicate a significant difference between the SSEF 3DVAR ensemble and the ensemble corresponding to that color.

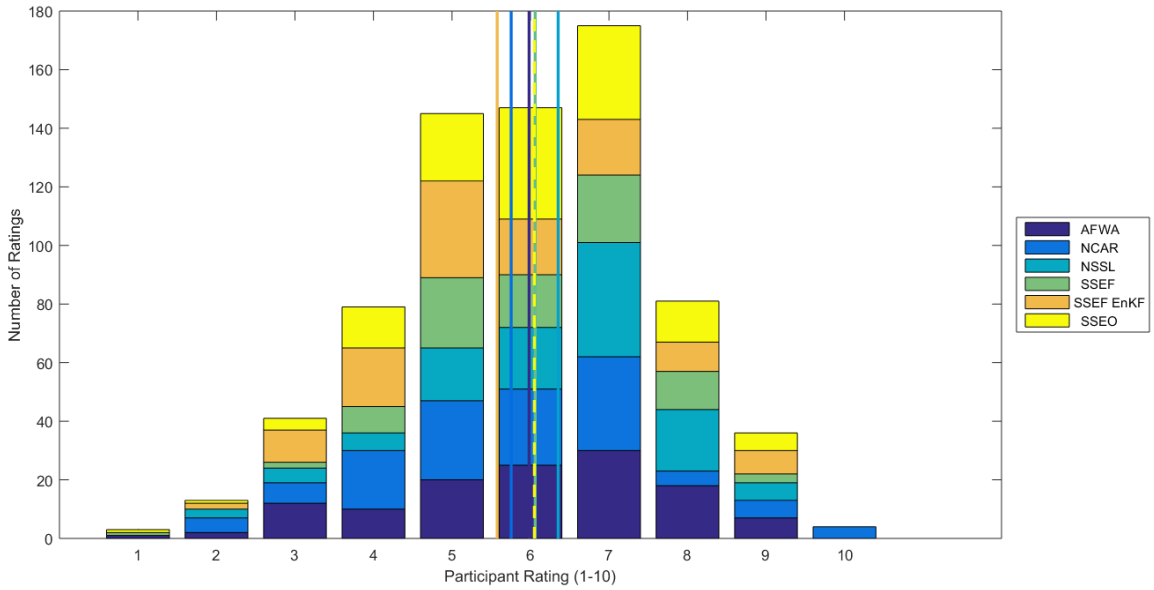

Figure 2.8 Distribution of subjective ratings (1 to 10) for the ensemble hourly maximum field forecasts compared to local storm reports for each ensemble. Mean subjective ratings are indicated by a vertical line. The dashed line indicates the mean of both the SSEF (3DVAR) and the SSEO subjective ratings.

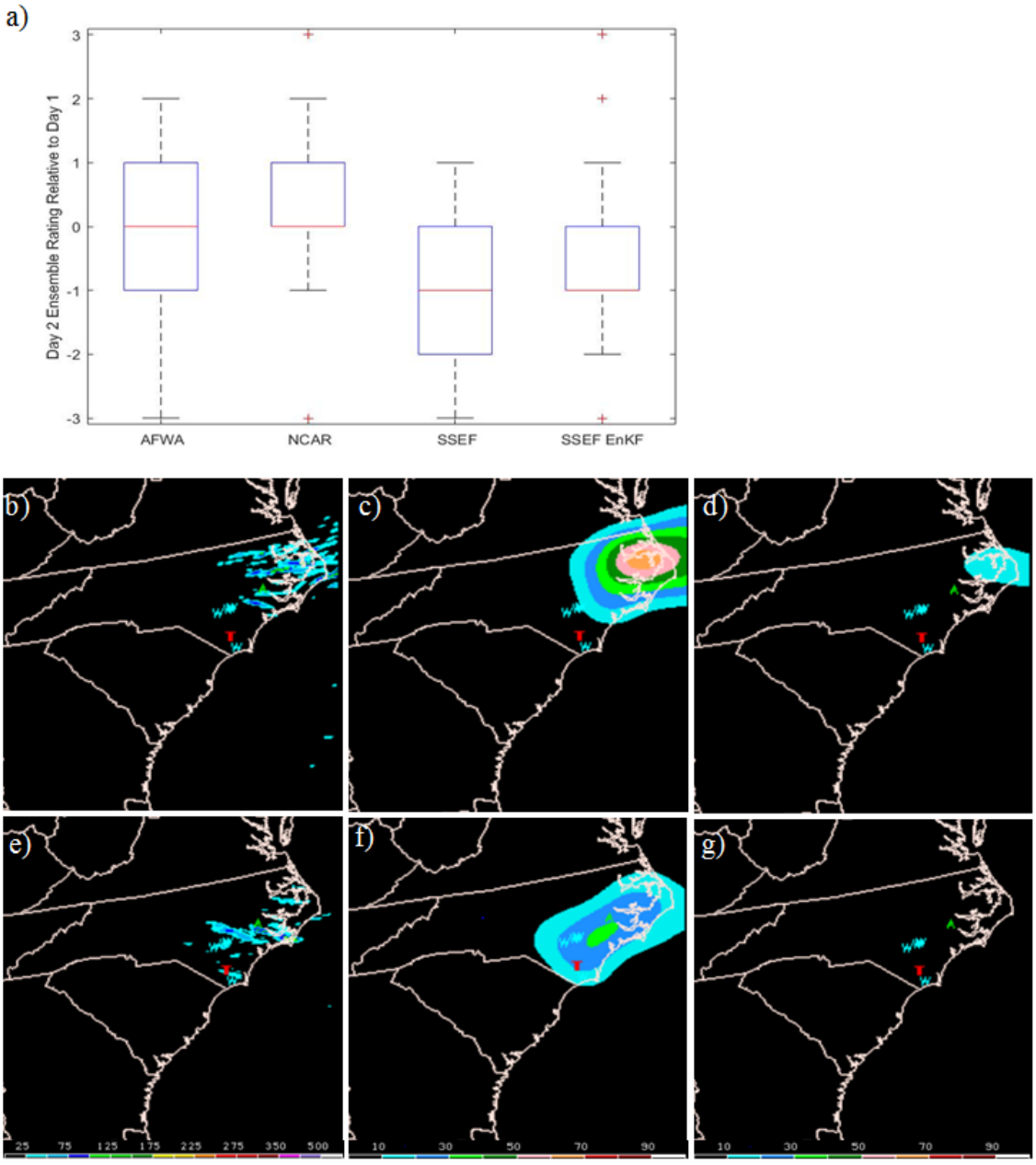

Figure 2.9 (a) As in Fig. 2.4, except for the distribution of subjective ratings (-3 to +3) for the Day 2 ensemble forecasts compared to the Day 1 forecasts, valid for the same time period. As an example, the AFWA Day 1 (b)-(d) and Day 2 (e)-(g) forecasts of 4-h ensemble maximum UH (b, e), ensemble neighborhood probability of UH ≥25 m^2s^{-2} (c, f), and ensemble neighborhood probability of UH ≥100 m^2s^{-2} (d, g) valid 1800-2200 UTC on 21 May 2015. The severe reports during this 4-h period are plotted as letters in each panel (T for tornado, W for wind, and A for hail).

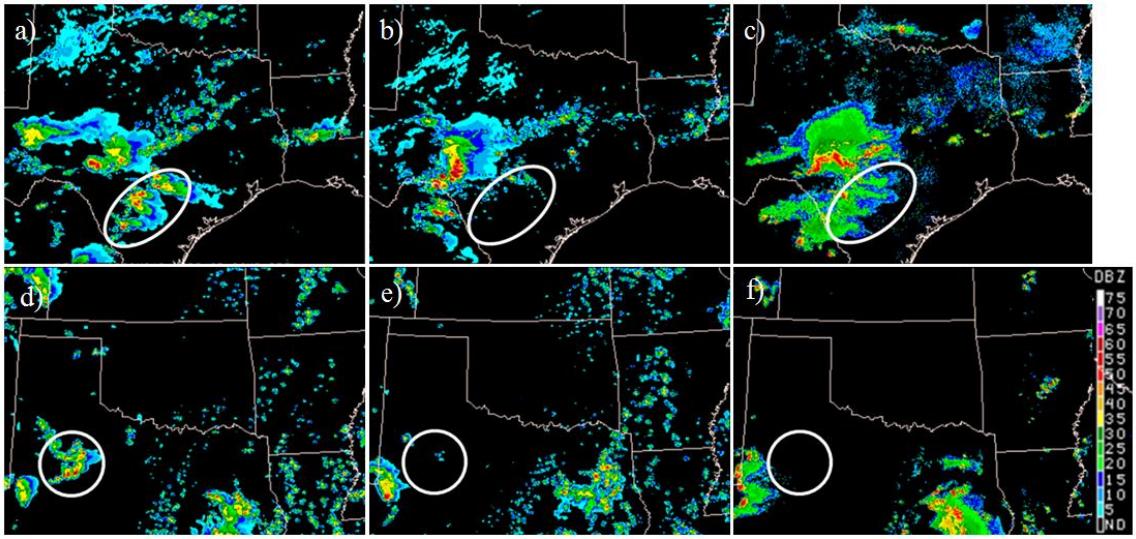

Figure 2.10 Simulated reflectivity forecasts valid at 0300 UTC on 21 May 2015 from the (a) 1500 UTC operational HRRR, (b) 1500 UTC parallel HRRR, and (c) observed reflectivity. Simulated reflectivity forecasts valid at 2200 UTC on 14 May 2015 from the (d) 1500 UTC operational HRRR, (e) parallel HRRR, and (f) observed reflectivity.

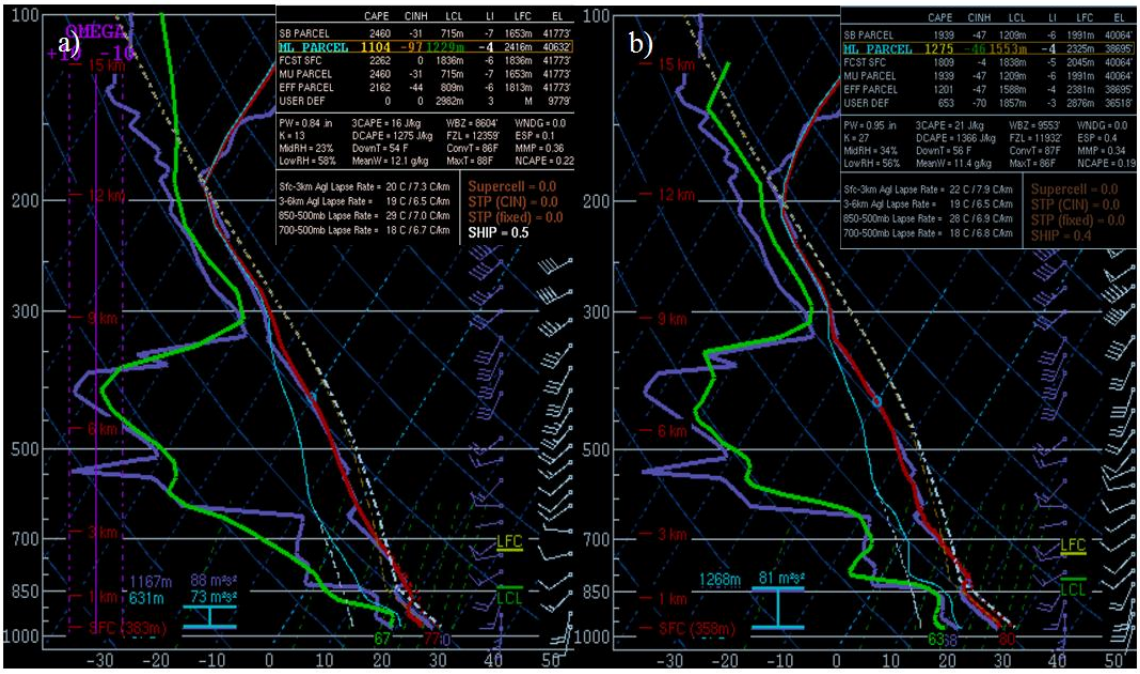

Figure 2.11 24 h forecast soundings valid 15 May 2015 for the OUN station from (a) the NSSL-WRF control member and (b) the UKMET 2.2-km model. The observed sounding is plotted in purple in each panel.

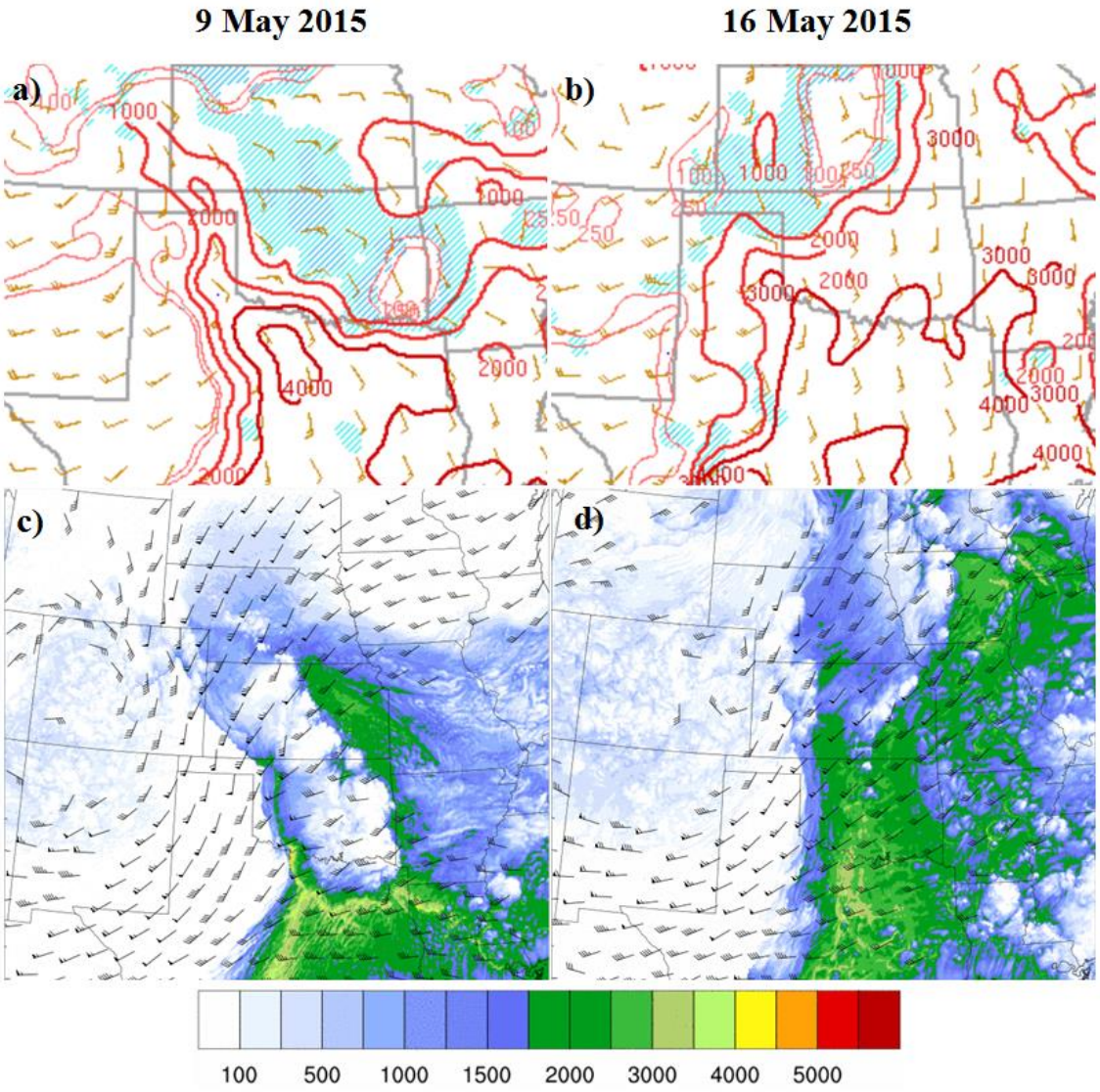

Figure 2.12 CAPE and CIN from SPC's mesoanalysis valid at (a) 2100 UTC 9 May 2015 and (b) 2100 UTC 16 May 2015. CAPE contour levels (red) are 100 J/kg, 250 J/kg, 500 J/kg, 1000 J/kg and then are spaced every 1000 J/kg. Light blue CIN indicates CIN less than -25 J/kg, and dark blue shading indicates CIN less than -100 J/kg. 69 h MPAS forecasts of CAPE and 0-6 km shear vectors beginning at 30 kts, valid (c) 2100 UTC 9 May 2015 and (d) 2100 UTC 16 May 2015.

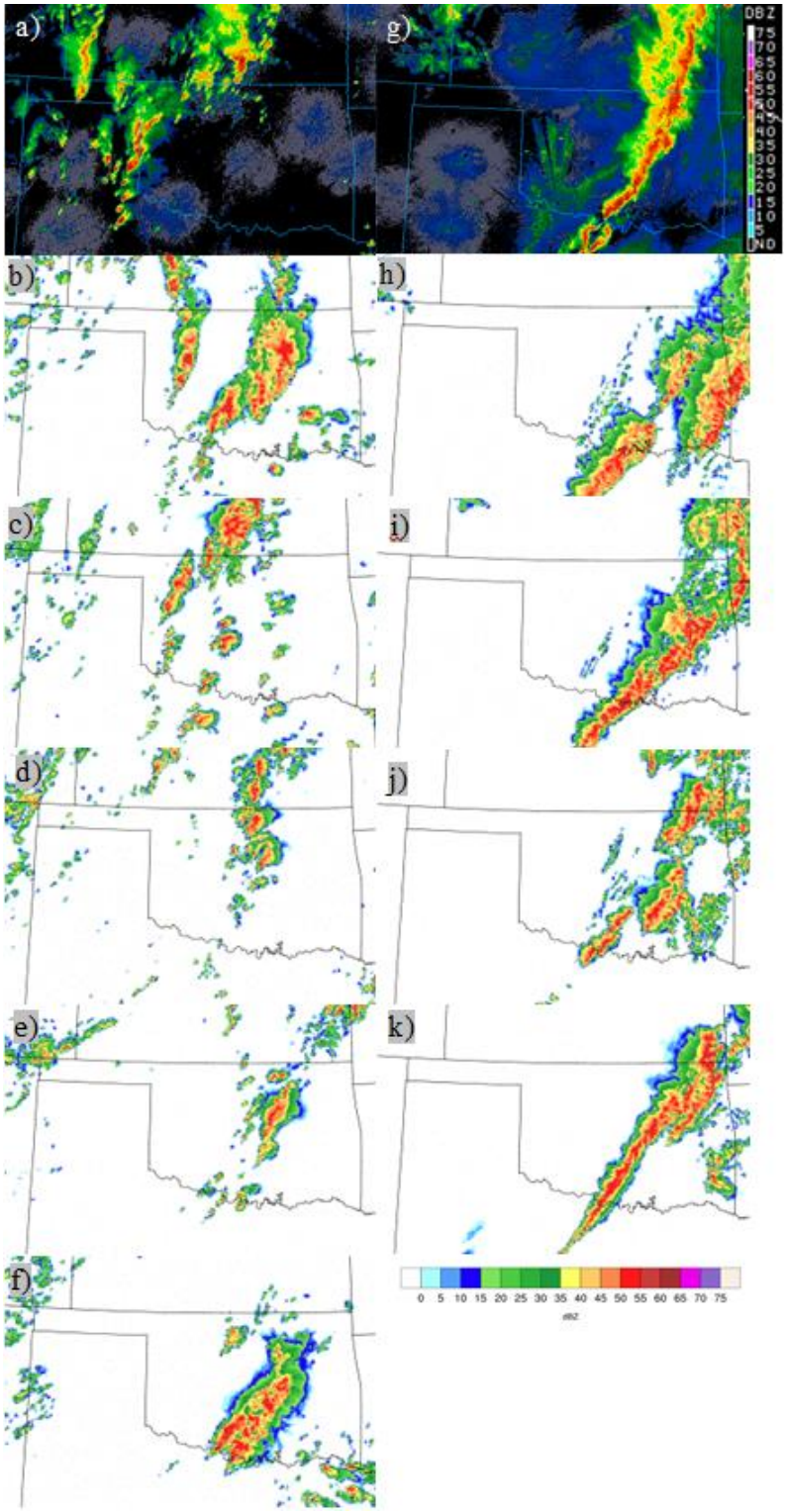

Figure 2.13 Composite reflectivity observations from (a) 2100 UTC on 16 May 2015 and (g) 0400 UTC on 17 May 2015. MPAS (b) 21 h, (c) 45 h, (d) 69 h, (e) 93 h, and (f) 117 h composite reflectivity forecasts valid on 16 May 2015 at 2300 UTC and (h) 28 h, (i) 52 h, (j) 76 h, and (k) 100 h composite reflectivity forecasts valid on 17 May 2015 at 0500 UTC.

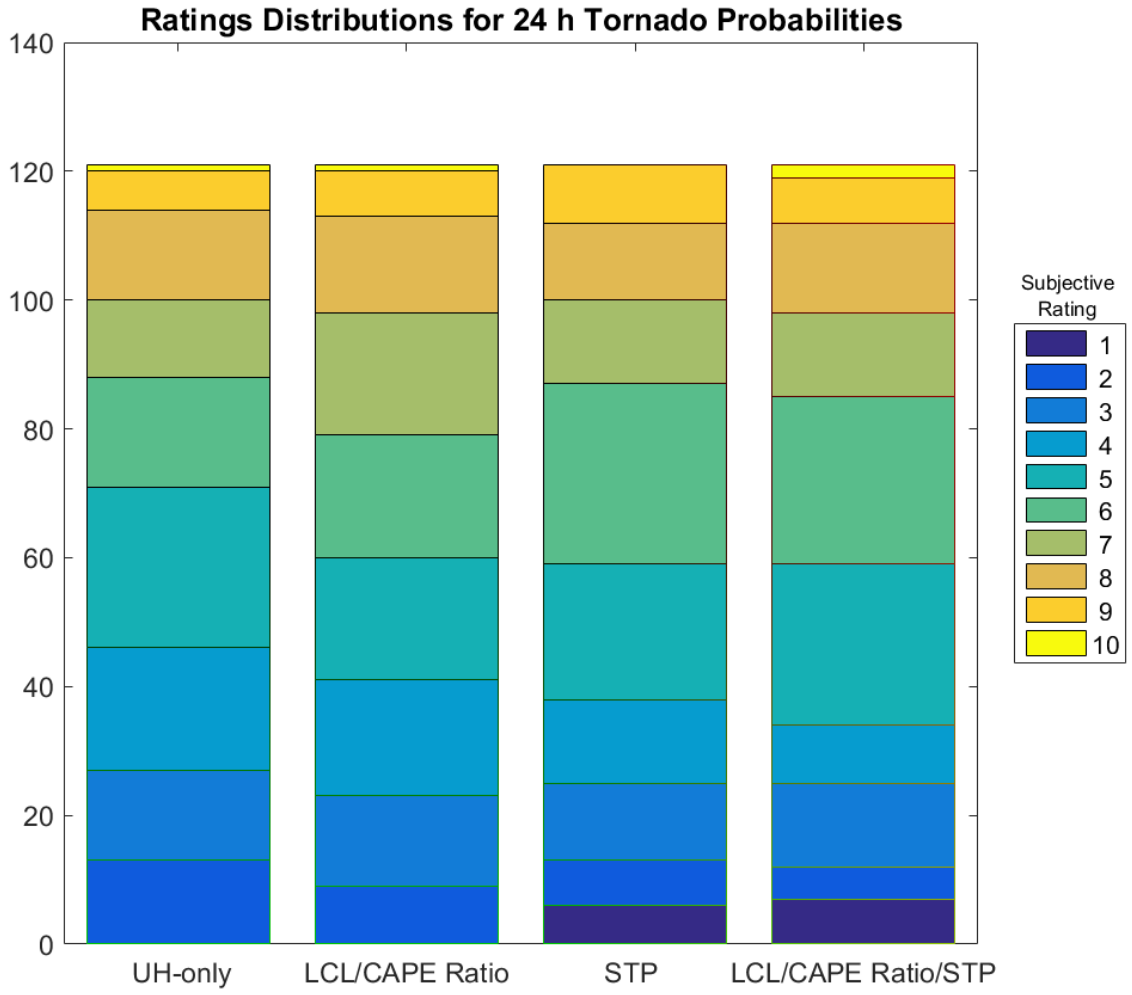

Figure 2.14 Subjective ratings of 24 h tornado probabilities generated from the NSSL-WRF ensemble requiring four different environmental criteria, along with UH $\geq$ 75m^2s^{-2}. Each set of probabilities received 121 ratings total. Adapted from Gallo et al. (2016).

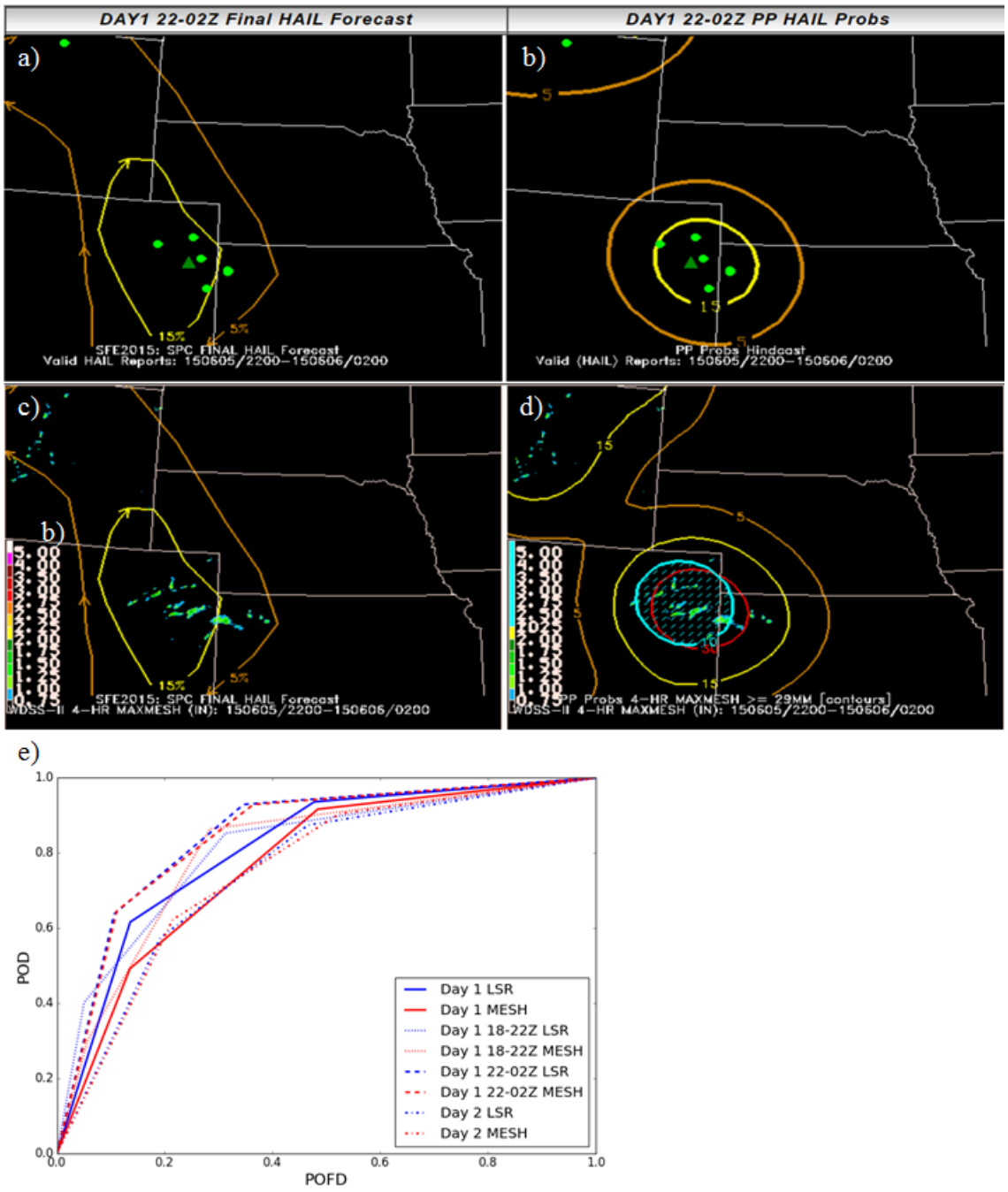

Figure 2.15 Individual hazard desk SPC forecaster's hail forecasts for 2200 UTC on 5 May 2015 to 0200 UTC on 6 May 2015 (a, c) verified against practically perfect forecasts generated using (b) hail LSRs (green dots) and significant hail LSRs (dark green triangles) and (d) MESH tracks. Full periods encompass 1600 UTC – 1200 UTC the following day. The blue hatched area is indicative of severe hail (≥2"). (e) ROC curves showing the accumulated verification results for all of SFE 2015 using LSRs and MESH.

Chapter 3: Forecasting Tornadoes using Convection-Permitting Ensembles

A paper published in *Weather and Forecasting*

Burkely T. Gallo[1], Adam J. Clark[2], and Scott R. Dembek[2, 3]

[1]School of Meteorology, University of Oklahoma, Norman, Oklahoma

[2]NOAA/OAR National Severe Storms Laboratory, Norman, Oklahoma

[3]Cooperative Institute for Mesoscale Meteorological Studies, University of Oklahoma, Norman, Oklahoma

Abstract

Hourly maximum fields of simulated storm diagnostics from experimental versions of convection-permitting models (CPMs) provide valuable information regarding severe weather potential. While past studies have focused on predicting any type of severe weather, this study uses a CPM-based Weather Research and Forecasting (WRF) ensemble initialized daily at the National Severe Storms Laboratory (NSSL) to derive tornado probabilities using a combination of simulated storm diagnostics and environmental parameters. Daily probabilistic tornado forecasts are developed from the NSSL-WRF ensemble using updraft helicity (UH) as a tornado proxy. The UH fields are combined with simulated environmental fields such as lifted condensation level (LCL) height, most-unstable and surface-based CAPE (MUCAPE and SBCAPE, respectively), and multi-field severe weather parameters such as the significant tornado parameter (STP). Varying thresholds of 2–5 km updraft helicity were tested with

differing values of σ in the Gaussian smoother that was used to derive forecast probabilities, as well as different environmental information, with the aim of maximizing both forecast skill and reliability. Addition of environmental information improved reliability and the critical success index (CSI) while slightly degrading the area under the receiver operating characteristic (ROC) curve across all UH thresholds and σ values. Probabilities accurately reflected the location of tornado reports, and three case studies demonstrate value to forecasters.

Based on initial tests, four sets of tornado probabilities were chosen for evaluation by participants in the 2015 National Oceanic and Atmospheric Administration/Hazardous Weather Testbed from 4 May – 5 June 2015. Participants found the probabilities useful and noted an overforecasting tendency.

3.1 Introduction

High-resolution convective-permitting models (CPMs) are increasingly part of an operational forecaster's severe weather toolbox (Fowle and Roebber 2003; Weiss et al. 2006; Coniglio et al. 2010; Sobash et al. 2011; Clark et al. 2012a; Schwartz et al. 2015a). These CPMs generally have grid spacing of 4 km or less, allowing them to represent bulk properties of convective circulations, skillfully differentiate convective modes (Fowle and Roebber 2003; Done et al. 2004; Weisman et al. 2008), and provide unique guidance using hourly maximum fields of simulated storm diagnostics (Kain et al. 2010). Spring Forecasting Experiments (SFEs) taking place in the National Oceanic and Atmospheric Administration (NOAA)'s Hazardous Weather Testbed (HWT) examine how well experimental CPMs can provide guidance to forecasters (Clark et al.

2012a). At the SFEs, researchers and forecasters discuss forecaster needs and current capabilities of CPMs, fostering greater understanding between research and operational communities. Input from forecasters to the research community allows for subjective information about perceived guidance value, rather than relying solely on objective measures of verification.

Murphy (1993) discusses three types of "goodness" that a forecast can possess: (1) the agreement between the forecast and the forecaster's conceptual model ("consistency"); (2) the correspondence between the forecast and observations ("quality"); and (3) the usefulness of the forecast to the end user ("value"). While objective measures asses the *quality* of the probabilities, feedback from SFE participants helps to improve the *consistency* of the probabilities, as well as the *value* of the probabilities to the forecaster. As tools for the forecaster, guidance should be consistent and valuable; working with SFE participants allows for modifying the probabilities to achieve these objectives while maintaining forecast quality.

Forecasters already use ensembles of coarser-resolution models, such as the Short-Range Ensemble Forecast (SREF; Du et al. 2014), to assess forecast uncertainty. Computing capabilities continue to improve, to the point where NOAA's Environmental Modeling Center plans to implement an operational, CPM-based ensemble in the near future (University Corporation for Atmospheric Research 2015). Compared to convection-parameterizing ensembles, CPM-based ensembles have been shown to provide better guidance in terms of precipitation forecast skill. Clark et al. (2009) found that the skill gained by upgrading ensembles to convection-permitting resolutions more than made up for the skill lost by decreasing the number of ensemble

members. However, exploring the effectiveness of CPMs at forecasting severe hazards is a relatively new endeavor. Updraft helicity (UH), a product of vertical vorticity and updraft speed, is described by:

$$UH = \int_{z_0}^{z_1} \zeta * w \, dz \,, \quad (3.1)$$

where z_0 and z_1 are the user-defined layer of the atmosphere, w is updraft speed, and ζ is the vertical vorticity (Kain et al. 2010). UH has been used to create probabilistic hazard guidance for any type of severe weather and skillfully distinguished severe weather events from non-severe weather events (Sobash et al. 2011). This skill is likely due to the detection of persistent midlevel mesocyclones – a characteristic of supercells, which cause a large percentage of severe weather reports (Duda and Gallus 2010). Indeed, hourly maximum UH correlates well with observations of mesocyclones (Kain et al. 2010).

While UH is a good predictor for severe hazards, it is not necessarily a good proxy for tornadoes when used alone. Like in reality, simulated mesocyclones often form in environments unfavorable to tornadogenesis (Clark et al. 2012b). Therefore, if generating tornado probabilities from UH alone, large areas of false alarm will occur in areas with unfavorable environments. However, adding environmental criteria for probability generation could reduce the false alarm area, increasing the precision of the tornado probabilities by combining the existence of simulated mesocyclones with environmental information conducive to tornadogenesis. This study focuses on combining model-generated rotation in the form of UH with environmental parameters

conducive to tornadogenesis as identified by numerous previous studies (Rasmussen and Blanchard 1998; Thompson et al. 2004a; Grünwald and Brooks 2011; Grams et al. 2012) to generate probabilistic forecasts of tornadoes.

Previously, high-resolution UH has been combined with coarser-resolution environmental information to separate the tornado threat from the hail and wind threat. Jirak et al. (2014) used the Storm-Scale Ensemble of Opportunity (SSEO; Jirak et al. 2012a), a CPM ensemble produced by the Storm Prediction Center (SPC), for UH fields and the 40 km SREF for environmental parameters, to extract individual hazard probabilities. The combination of large-scale environmental information with the small-scale UH diagnostic is shown to provide skillful tornado guidance, with some overprediction of hail and wind threats (Jirak et al. 2014). This study aims to investigate the benefits of combining UH with environmental parameters taken from the same model in generating probabilistic tornado forecasts. Probabilistic forecasts reflect both uncertainty in the exact location of the storms as well as whether or not an individual storm will produce a tornado. Several objective verification metrics assess the quality of the forecast probabilities, as well as examination subjective comments provided by participants in the 2015 SFE.

Section 3.2.1 of this paper will describe the ensemble system and the parameters used to generate the tornado probabilities. Section 3.2.2 will elaborate upon the probability generation methodology, and section 3.2.3 will explain both the objective and subjective verification methods. Section 3.3.1 evaluates the quality of the tornado probabilities through objective verification metrics. Differences in probability generation methods will be highlighted by three case studies in section 3.3.2. Section

3.3.3 will describe the subjective evaluation that took place, including common themes noted by the SFE 2015 participants. Finally, a summary and discussion of the results along with conclusions and suggestions for further research are provided in section 3.4.

3.2 Data and Methodology

3.2.1 The NSSL-WRF ensemble configuration

Since fall 2006, SPC forecasters have used output from an experimental, 4 km version of the Weather Research and Forecasting model (WRF; Skamarock et al. 2008) generated by the National Severe Storms Laboratory (NSSL) using the Advanced Research core WRF (WRF-ARW), known hereafter as the NSSL-WRF (Kain et al. 2010). This model runs twice daily, at both 0000 UTC and 1200 UTC. Nine additional 4 km WRF-ARW members with varying initial conditions are run at 0000 UTC, composing an ensemble of ten members known as the NSSL-WRF ensemble. Eight of the members are initialized at 0000 UTC using 3h SREF forecasts initialized at 2100 UTC for initial conditions and corresponding SREF member forecasts as lateral boundary conditions. The remaining member uses the 0000 UTC National Center for Environmental Prediction (NCEP) Global Forecast System (GFS) analysis for initial conditions and the corresponding NCEP GFS forecast as lateral boundary conditions. Physics parameterizations amongst all members are identical, using the Mellor-Yamada-Janjić (MYJ; Mellor and Yamada 1982; Janjić 2002) planetary boundary layer scheme, WRF single-moment six-class (WSM-6; Hong and Lim 2006) microphysics, the Noah (Chen and Dudhia 2001) land-surface model, the Rapid Radiative Transfer Model (RRTM; Mlawer et al. 1997) longwave radiation and Dudhia (Dudhia 1989)

shortwave radiation scheme (Table 3.1). The NSSL-WRF ensemble began running in February 2014. Each ensemble run includes 35 vertical levels and was integrated 36 h over the CONUS starting at 0000 UTC. For this study, the period from 1200 UTC to 1200 UTC the following day is considered (forecast hours 12 to 36).

Two spring seasons are examined herein: 1 April – 30 June 2014 and 1 April – 30 June 2015. Ensemble membership changed slightly in that time period, with two members initialized from EM SREF members switched for two members initialized from NMB SREF members. This change occurred because SPC forecasters noticed that the EM SREF members were much less dispersive than the other sets of SREF model cores, resulting in a clustering of members and a subsequent decrease in the ensemble variability. Thus, a switch to more NMB SREF members was hoped to increase spread and improve reliability. The change in ensemble membership was tested by comparing reliability diagrams for each year using the consistent members and reliability diagrams for each year using all of the members. Reliability diagrams plot the forecast probability versus the observed relative frequency, and only small differences occurred through the addition of the varying members to the constant members. This change did not have significant effects on the composition of the generated probabilities. Thus, the change in two members does not significantly affect the overall forecast probabilities, and the years are combined throughout the following verification.

3.2.2 Probability generation

Probabilities based on the NSSL-WRF ensemble were generated using the 2–5 km hourly maximum UH (Kain et al. 2010), defined by integrating the vertical vorticity times the updraft velocity for the 2–5 km above ground layer (e.g., Kain et al. 2008).

These hourly maximum variables contain the maximum value of UH at a given point for each hour, providing insight on trends in storm intensity and movement hour-by-hour. Hereafter, UH will refer to the hourly maximum quantity. Probabilities were generated following Hamill and Colucci (1998). For each case, the daily maximum value of UH is found at each gridpoint for each member. Next, for each gridpoint a distribution of UH values is created using the value of maximum UH within a 40 km radius for each member. Probabilities are found by determining where the chosen threshold of UH (e.g. 25 m^2s^{-2}) is within this distribution. If the threshold is greater than all members forming the distribution, the Gumbel distribution (Wilks 2011) is used. The resulting probabilities are smoothed using a Gaussian kernel density weighting function, whose weights are calculated by:

$$f(x,y) = \frac{1}{2\pi\left(\frac{\sigma}{\Delta x}\right)^2} * e^{\frac{-(x^2+y^2)}{2\left(\frac{\sigma}{\Delta x}\right)^2}} \qquad (3.2)$$

where σ is the user-defined standard deviation in units of km and Δx is the grid-spacing. Varying σ results in different levels of smoothness in the resultant probability fields – the higher the σ, the smoother the probability fields.

The first aim of this study is to determine the optimal σ for the Gaussian kernel and the optimal UH value. Previous studies have found that UH greater than or equal to 40 m^2s^{-2} generated reliable probabilities of any severe report (Sobash et al. 2011). To focus on the tornado problem rather than on the any severe problem, five thresholds of hourly maximum UH were examined beginning at 25 m^2s^{-2} and increasing to 125 m^2s^{-2} at 25 m^2s^{-2} intervals. While Sobash et al. (2011) found a relatively large smoothing

radius of 200 km to best discriminate severe events from nonevents, it is expected that in the current study usage of an ensemble framework allows for a smaller optimum σ because the ensemble members will account for much of the spatial uncertainty. This differs from the Sobash et al. (2011) study, in which the Gaussian kernel accounted for *all* spatial uncertainty.

The first set of verification statistics for probabilities with varying UH and σ without environmental information provided a baseline against which probabilities incorporating environmental information were compared. While the UH is an hourly maximum variable, the environmental variables were instantaneous and assumed to be representative of the environment into which the storm was moving. To assign values of environmental parameters to values of maximum UH at each gridpoint for each member, the hour of the maximum UH during the period of interest was determined. Then, the environmental information for the previous hour was used for that point. If the environmental information were below certain thresholds, the UH was not included in the probability generation (i.e., UH was set to zero). The environmental variables from the previous hour of the maximum UH were used in three different combinations. One combination, designed to eliminate elevated storms [where the inflow is drawn from an above-surface unstable layer; Colman (1990)] as well as high-based storms, required the ratio of surface-based convective available potential energy (SBCAPE) to most-unstable convective available potential energy (MUCAPE) to be at least .75, and the lifted condensation level (LCL) height to be below 1500m AGL. These requirements helped ensure that the storm inflow originated in the near-surface layer and that cloud bases would be relatively low. The values of .75 and 1500m were

chosen based on Clark et al. (2012b), where these values were found to successfully identify UH in environments supportive of elevated and high-based storms. In another combination, the fixed-layer significant tornado parameter (STP; Thompson et al. 2003) was required to be greater than one. Thompson et al. (2003) designed the STP to discriminate significant from non-significant or non-tornadic environments (Thompson et al. 2003), utilizing the surface-based convective available potential energy (SBCAPE), 0–6 km bulk shear (SHR6), 0–1 km storm relative helicity (SRH1), and the surface-based lifting condensation level (SBLCL):

$$STP = (SBCAPE\ /1500 Jkg^{-1})(SHR6/20 ms^{-1})(SRH1/150 m^2 s^{-2})[(2000m - MLLCL)/1000m]. \quad (3.3)$$

Since a value of 1 or greater indicates an environment supportive of significant tornadoes, it was selected as the threshold for this study. Because this study is verifying all tornadoes, both significant and non-significant, requiring STP to be at or greater than one may seem too stringent. However, based on the results (shown later), it still slightly over-predicts tornado occurrence. The final combination of environmental parameters used both prior combinations of environmental parameters: SBCAPE to MUCAPE ratio greater than .75, LCL heights below 1500 m, and STP greater than one. Each UH threshold and smoothing radius were tested for these three sets of environmental parameters.

3.2.3 Verification

Objective verification of the forecasts was conducted using reliability diagrams (Wilks 2011), receiver operating characteristic (ROC) curves, the area beneath the ROC

curves and the Critical Success Index (CSI). The area under the ROC curve measures the ability of a forecast to discern the outcome of a binary event, and is computed by plotting the probability of detection (POD), defined as:

$$POD = \frac{hits}{hits + misses} \quad (3.4)$$

against the probability of false detection (POFD), defined as:

$$POFD = \frac{false\ alarms}{false\ alarms + correct\ negatives} \quad (3.5)$$

at specified levels of probability: .5%, 1%, 2%, 3%, 4%, 5%, 10%, 15%, 20%, 25%, 30%, 35%, 40%, 45%, 50%, 55%, 60%, 65%, 70%, 75%, 80%, 85%, 90%, and 95%. Computing these statistics for smaller increments at low probabilities than high probabilities follows SPC tornado probability forecasts and accounts for large differences in area between low probability thresholds. The area under this curve is computed using the trapezoidal method (Wandishin et al. 2001), and ranges from 0 to 1. A value of 1 is a perfect forecast, a value above .5 is considered to have positive skill, and a ROC area of .7 is considered the lower limit of a useful forecast (Buizza et al. 1999). To test the statistical significance of the difference between ROC areas from two forecasts, resampling was done following Hamill (1999). Cases were randomly assigned to one of the two forecast methods (i.e., UH only vs. UH and STP) 1000 times to create a distribution of ROC area differences. If the ROC area differences calculated using the

two forecasting techniques lies outside the 95% confidence interval, they were deemed significant.

While ROC curves determine the discriminating ability between events and nonevents, the shape of the ROC curves is unaffected by probability magnitude and therefore not impacted by biased probability forecasts. To visualize the bias in the forecasts, reliability diagrams were generated by plotting the forecast probability against the observed relative frequency. A diagonal line represents a forecast probability equal to the observed relative frequency (i.e., perfect reliability). Values above (below) the diagonal represent underforecasting (overforecasting), where the observed relative frequency is higher (lower) than the forecast probability.

The final metric considered, CSI, is the number of correct "yes" forecasts divided by the total number of hits, misses, and false alarms:

$$CSI = \frac{hits}{hits + misses + false\ alarms} \quad (3.6)$$

It is a score often used in rare events (Wilks 2011), and is therefore an appropriate score to consider in tornado forecasting. Scores range from 0 to 1, with 1 being a perfect score. Visualization of CSI is through performance diagrams (Roebber 2009). Performance diagrams plot the POD versus the Success Ratio, which is defined as:

$$Success\ Ratio = 1 - \frac{false\ alarms}{hits + false\ alarms} \quad (3.7).$$

False alarms divided by hits plus false alarms is also known as the false alarm ratio, or FAR. Lines of constant reliability are plotted as dashed lines, and lines of constant CSI are plotted as solid, curved lines.

These measures were applied across the eastern two-thirds of the CONUS (Fig. 3.1). Verification was based on the Local Storm Reports (LSR) database for each day, as generated by the SPC. Reports filtered by the SPC were used to attempt to remove duplicate reports of the same tornado. While the tornado report database is flawed (Verbout et al. 2006; Doswell et al. 2009), underreporting has been reduced in current decades (Brooks and Doswell 2002) and utilizing the location of reported tornadoes for verification emphasizes the utility of CPM ensembles in highlighting spatial areas of concern. Only the starting points of tornado paths are used to assign locations of the reports, and tornado path length is not considered. Verification was performed on the 4 km grid of the NSSL-WRF and observed reports were mapped to the 4 km grid and treated as yes/no binary events, where a yes occurred if a tornado report was within a 40 km radius.

Subjective verification of the forecasts took place at the Experimental Forecast Program of the Spring Forecasting Experiment at the Hazardous Weather Testbed from 4 May – 5 June 2015. During this experiment, participants were presented with forecast probabilities and overlaid LSR tornadoes from the period of interest. The forecasters were then asked to assign ratings to the forecasts on a scale from 1 (Very Poor) to 10 (Very Good), and to provide specific comments about the forecasts and the methods of incorporating environmental parameters into the probabilities. They could also explain why they assigned the ratings they chose for each forecast.

3.3 Results

3.3.1 Objective Verification

Objective verification of the probabilities utilized the ROC curve, the area under the ROC curve, and reliability diagrams. ROC areas were first computed for the probabilities that solely incorporated UH (Fig. 3.2). The impact of changing the UH threshold and the σ value of the Gaussian kernel were tested. As the threshold of UH was increased, the ROC areas decreased at all σ levels, likely due to the probability of detection (POD) decreasing more quickly than the probability of false detection (POFD) as the UH threshold increases. However, the ROC areas remained above 0.7 for all thresholds and σ values. Decreases in ROC area were greater above the UH threshold of 50 m^2s^{-2}, decreasing by ~0.01 or less from 25 m^2s^{-2} to 50 m^2s^{-2} and from 0.01–0.03 for each 25 m^2s^{-2} of UH added to the threshold past 50 m^2s^{-2}. Differences in ROC areas between thresholds separated by 25 m^2s^{-2} were not statistically significant, but differences in ROC areas between thresholds separated by 50 m^2s^{-2} were significant. Increases in the σ value had smaller effects on the ROC area than increases in the UH thresholds. In general, as the σ value increased, (more smoothing; example given in Fig. 3.3) the ROC area increased slightly. However, at low UH thresholds, increasing σ past 50 km decreased the ROC area. The same effect at high UH thresholds was seen at σ past 100 km, suggesting a less skillful forecast. ROC area changes caused by σ variation were one to two orders of magnitude smaller than the changes caused by adjusting the UH threshold. In fact, differences in the ROC area between σ of 20 km and σ of 200 km show no statistically significant difference at any UH threshold. Thus, the variation in

the UH threshold has a larger influence on the ROC area than the smoothing level. This ROC area behavior is similar to the results of Sobash et al. (2011), who also found that ROC area decreased with increased UH threshold and generally increased with increasing σ.

The same pattern occurs when the probabilities incorporate environmental information. ROC areas for varying levels of σ and UH (Fig. 3.4) show that environmental filtering decreases the ROC area in most instances, but the ROC area remains above 0.8 for all cases except for UH ≥ 125 m^2s^{-2}, the highest UH threshold tested. The ROC area decrease depends on the filtering method, UH threshold, and σ value. The LCL/CAPE ratio method shows the smallest difference from the UH-only probabilities, with an average difference across all σ values and all UH thresholds of -0.005. Indeed, in two cases (UH ≥ 25 m^2s^{-2}/σ = 50 km and UH ≥ 50 m^2s^{-2}/σ = 100 km) the environmental information increases the ROC area compared to UH-only. However, neither of these differences were statistically significant, nor were other differences between the LCL/CAPE ratio method and the UH-only method across the σ and UH producing the largest average differences. The differences become larger and statistically significant for the STP method when compared to the UH-only method, with an average difference across all σ values and UH thresholds of -0.035. The difference from the UH-only method widens for the LCL/CAPE ratio/STP method at -0.039, while the difference between the LCL/CAPE ratio/STP method and the STP method is quite small, reflecting the large dependence on STP. The differences are larger across all methods for larger UH thresholds, often at the 0.01 order of magnitude. Therefore, the environmental information incorporated generally has as much of an

impact on the ROC area metric as the selection of UH threshold, and more of an impact than the selection of σ.

Figure 3.5 visualizes example ROC curves for five varying UH thresholds using four methods of probability generation. The σ of the Gaussian kernel is fixed at 50 km for Fig. 3.5, as 50–100 km is the range above which most ROC areas began to decrease for a given threshold. Generally, POD and POFD decrease as the UH threshold increases, because more events are being missed at higher UH thresholds. While it may seem counterintuitive that environmental information causes lower ROC areas, the curves show that most of the information loss occurs at low probabilities (i.e., less than 0.5%). Since events are rare, missing one event causes a large decrease in POD at very low thresholds. At operational probability thresholds (i.e., 2%+), the environmental information causes slight improvement in the POFD which is then offset by the decrease in POD at low probability levels.

While the ROC areas are highest for low thresholds of UH, they are heavily influenced by correct negatives, which compose a large portion of the data for tornadoes on a high-resolution grid. Thus, CSI was examined to provide a metric that excludes correct negatives. Performance diagrams (Fig. 3.6) show that the addition of environmental information increases the CSI at ranges used by the SPC operationally: 2%, 5%, 10%, 15%, 30%, and 45%. CSI also improves as UH thresholds increase. While all values considered are far from a perfect forecast of 1, they are similar to the results of Sobash et al. (2011), and roughly what is expected from high-resolution verification of very rare events such as tornadoes. Finally, CSI shows improvement with

additional environmental information, with the LCL/CAPE ratio/STP method often having the highest CSI at a given probability level.

The effect of changing σ is pronounced when considering the reliability diagrams for the UH-only probabilities (Fig. 3.7). For small values of σ, all forecast probabilities are much larger than the observed relative frequency, indicating overforecasting. This overforecasting persists as σ is increased, but the degree of overforecasting lessens with increased σ. For larger σ (Fig. 3.7e-f) the overforecasting is minimized for most levels, but sample size begins to limit the number of higher forecast probabilities, starting around a UH threshold of 75 m^2s^{-2}. Since the ROC areas of each σ level were statistically indistinguishable and a limited sample size occurred at high probability thresholds, the UH and σ combination used in SFE 2015 was selected as the more computationally efficient σ of 50 km and a UH threshold of 75 m^2s^{-2} to maintain reliable high probabilities. Though these high probabilities are larger than what is currently operationally forecasted, the reliability of these probabilities combined with the relatively high ROC areas suggest skillful forecasts.

This chosen threshold (σ of 50 km and UH of 75 m^2s^{-2}) is compared amongst all methods of probability generation (Fig. 3.8). Incorporation of the environmental information greatly increases the reliability, particularly at higher probability values. As it is harder for the probabilities to meet all environmental criteria (recall that the fixed-layer STP consists of four separate parameters), fewer ensemble members will meet the criteria in a given neighborhood. This dampens the magnitude of the probabilities and leads to a reduction in overforecasting. The environmental criteria also reduce the spatial area encompassed by the probabilities. The reduction in spatial area will be more

fully illustrated in the case study examples given in Section 3.2.2b. As can be seen in Fig. 3.8b, incorporating LCL height and the CAPE ratio increases the reliability at high forecast probabilities. However, only a slight increase occurs at the lower magnitudes. When the STP is considered, as in Fig. 3.8c and Fig. 3.8d, the reliability increases for all magnitudes of probability, and the overforecasting is more uniform than in both Fig. 3.8a and Fig. 3.8b.

When these results are compared to probabilities generated without UH, instead requiring that STP ≥ 1, vast overforecasting occurs at all levels, and large swaths of very high probabilities occur (Fig. 3.9). These results emphasize the need for multiple methods of evaluating the probabilities, as the ROC area from both spring seasons is 0.90, similar to that found with solely using UH. However, the large swaths of high probability seen on individual days (Fig. 3.9a) demonstrate how difficult it would be to use these probabilities as a first guess forecast, as extremely high probabilities encompass much of Texas and Oklahoma. There is also a very sharp gradient in probabilities, reflecting the larger overforecasting problem illustrated by the reliability diagram (Fig. 3.9b).

Forecasters develop intuition about various models and products; these statistics may help calibrate forecasters. The high probabilities in all cases involving environmental parameters demonstrate high observed relative frequency, occasionally even underforecasting high probability events. While the sample size at high probabilities is fairly small, when high probabilities occur, a tornado is relatively likely and forecasters can proceed with heightened awareness. The high values of ROC area

found across all probabilities also indicate that these forecasts can successfully distinguish areas of tornado occurrence from areas without tornado occurrence.

3.3.2 Example cases

Three example cases are discussed in this section. The first case was a typical synoptic setup for spring in the southern Plains, with ample CAPE, strong shear, and relatively little convective inhibition, spawning multiple tornadoes across southern Oklahoma and northern Texas. These tornados were well-depicted by the probabilities. The second case is a late spring case, taking place in the northern Plains with a secondary area of focus across the mid-Atlantic. This case had more tornadoes than the first case, and demonstrates the performance of the probabilities in less climatologically favored regions for tornadoes. The final case demonstrates a day where the probabilities had difficulty pinpointing the area of highest tornado risk, instead portraying a broad area of false alarm, with the tornado reports occurring away from the highest magnitude of probabilities. While it is unwise to judge the quality of probabilistic forecasts based on individual days, these probabilities are meant to be tools for forecasters. As such, the potentially operational end products are presented here. These case studies further emphasize the operational potential of these forecasts.

a. 19 MAY 2015

On 19 May 2015 at 1200 UTC, a 500 hPa shortwave trough progressed across the Great Basin area, with a 500hPa speed maximum of 55–60 kt located over Arizona and New Mexico (Fig. 3.10). At 850 hPa (not shown), moist air was advected northwestward from the Gulf of Mexico, and dewpoints across Oklahoma and northern Texas reached 10°C–14°C. While this setup is often associated with outbreaks of severe

weather across the southern Plains (Corfidi et al. 2010; Mercer et al. 2012), this case was complicated by the presence of ongoing convection across the Texas Panhandle. On this day, a slight risk was issued by the SPC despite the high values of shear and potential for large CAPE, largely due to the morning convection and subsequent cloud cover, and a lack of an elevated mixed layer as discussed in the 1630 UTC Day 1 convective outlook. This case took place during SFE 2015, and both experiment leaders and participants agreed that the convective mode, evolution, and timing were particularly difficult to forecast due to the ongoing storms and mixed numerical guidance regarding convective mode. Many models showed multiple mesoscale convective systems (MCSs) moving across the region of interest during the day, but some suggested that supercellular storms would form in the warm sector ahead of the ongoing convection and south of an east-west oriented surface stationary front.

This front progressed slowly northward throughout the day, and tornadic supercells formed after the passage of the weak MCS generated by the morning convection. These supercells grew upscale into a second MCS that stretched across Oklahoma into northern Texas. Behind these supercells, surface heating was able to initiate a third MCS over the Texas Panhandle late in the day, which eventually caught up to and merged with the second MCS into an east-west oriented MCS located along the stationary front. A few supercells also initiated off of the Davis Mountains in southern Texas, far from the morning convection. At the end of the day, 29 tornadoes were reported across Oklahoma and Texas.

Clearly, this was a difficult day to forecast specific hazards. The mixed-mode signal suggested that wind, hail, and tornado threats were possible. The tornado

probabilities provided an excellent first guess for the locations of the tornado reports (Fig. 3.11). UH-only probabilities (Fig. 3.11a) broadly highlight northern Texas and southern Oklahoma, as well as a secondary area of concern associated with the Davis Mountains. The highest probabilities were centered along the Red River, which forms the border between southern Oklahoma and northern Texas, and the highest magnitudes were ~45%. This bullseye was where the highest concentration of tornado reports occurred. The LCL and SBCAPE/MUCAPE ratio method (Fig. 3.11b) maintained the high magnitude of probabilities around the Red River, but correctly diminished the high probabilities of the UH-only method across the Texas Panhandle. The low-end probabilities generally encompassed the same area as the UH-only probabilities, but magnitudes decreased (Fig. 3.11b). The STP method (Fig. 3.11c) greatly reduces the magnitude of probabilities far from the bullseye while maintaining high probabilities in the bullseye, although the magnitude of the probability reduction is less than with the SBCAPE/MUCAPE ratio method. The area of false alarm initially present across the Texas Panhandle (Fig. 3.11a,b) is also greatly reduced by the STP method. Finally, the method with LCL height, CAPE ratio, and STP decreases the probabilities the most, and provides the greatest correspondence of the probabilities with the location of the tornado reports (Fig. 3.11d). The secondary bullseye of higher probability across the Texas Panhandle is greatly diminished, while the area of higher probability remains present across the Davis Mountains.

The high magnitude of the probabilities along the Red River is maintained in all methods of tornado forecast generation, showing that incorporating environmental information maintains the high risk of tornadoes across this area. This contrasts with the

area of relatively high probabilities across the Texas Panhandle, which was greatly reduced by using the environmental information. While the broad area encompassed by the probabilities remained consistent, the highest risk was shifted toward the observations through the addition of the environmental information, and highlighted the area of highest tornado risk despite mixed signals regarding the convective mode, evolution, and timing of the day's storms.

b. 27 JUNE 2015

On 27 June 2015, a 500 hPa trough at 1200 UTC was across the Mississippi valley (Fig. 3.12). 250 hPa wind speeds (not shown) were high considering the location and time of year, reaching over 100 kt ahead of the main trough axis. Two separate areas of tornadic storms formed: one across the eastern Dakotas and one across the Mid-Atlantic and the Carolinas. The SPC issued an enhanced risk across both areas, and encompassed most tornado reports within either the enhanced or the slight risk area. The SPC noted in their 1630 UTC convective outlook that the warm front in the east provided backed wind profiles capable of supporting rotating storms, as well as steep lapse rates and strong upper level winds associated with the shortwave trough evolving from Canada into the Dakotas. However, the weak anticipated low-level wind shear caused uncertainty with regards to the tornado risk due. By the end of the event, 35 tornadoes were reported, with a majority of the tornadoes occurring across North Dakota into Minnesota.

On this day, the probabilities highlighted the northern system (Fig. 3.13). The probabilities emphasized tornadic risk across the Dakotas, while maintaining low risk across the Mid-Atlantic. The orientation of the probabilities in both cases also closely

matched the orientation of the reports, suggesting that the synoptic setup was accurately portrayed. Comparing Fig. 3.13a and Fig. 3.13d demonstrates the reduction in both magnitude and areal coverage of probabilities provided by adding environmental information. The difference plots shown in Fig. 3.13b and Fig. 3.13c show that the STP method in this case caused a much larger reduction than the LCL/CAPE ratio method. The STP method also eliminates the area of false alarm in Alabama. While all of the northern tornado reports remain within the envelope of probabilities with all environmental criteria (Fig. 3.13d), the focus of the tornado probabilities in the Mid-Atlantic is much more northerly than the reports. Though the mid-Atlantic probabilities encompassed two of the tornado reports, on this day many of the North Carolina tornadoes were missed.

This case is discussed to demonstrate that the probabilities are useful across the United States; wherever the environmental conditions are favorable for tornadogenesis and UH is present within the ensemble, probabilities will ocur.

c. 28 MAY 2015

On 28 May 2015, a shortwave trough was located across the Rocky Mountains, with several smaller shortwave impulses along the larger trough axis. One such shortwave impulse ejected from northern Oklahoma, with another impulse set to eject northeastward over Texas throughout the day (Fig. 3.14). Upper-level wind speeds at the trough's base were approximately 55–65 kt at 250 hPa (not shown), and low level moisture was abundant. Prior convection left remnant outflow boundaries across Kansas, Oklahoma, and Texas, and the Storm Prediction Center's 1300 UTC and 1630 UTC convective outlooks noted their potential as foci for convective initiation later in

the day. Despite morning convection, storms initiated along the outflow boundaries and produced one tornado before they quickly grew upscale into a large MCS that spanned Texas, while farther northward isolated supercells across western Kansas also grew upscale into clusters. The supercells in Kansas produced a string of tornado reports, as did supercells near the Oklahoma/Colorado border.

Probabilities on this day suggested a widespread region of risk from southern Nebraska south to the Texas-Mexico border (Fig. 3.15a). These probabilities exceeded 30% across most of Texas. While one report did occur in this area, the majority of reports took place away from the area of highest probabilities. In addition, false alarm was present across most of Oklahoma and Texas. Again, usage of the environmental information decreased the probabilities (Figs. 3.15b,c). The decrease in probabilities using the LCL/CAPE ratio method (Fig. 3.15b) was fairly uniform across Texas and Oklahoma, but lowered the probabilities where most of the tornado reports occurred. The STP method (Fig. 3.15c) reduced the probabilities much more than the LCL/CAPE ratio method did, but again the highest-magnitude reductions were near the actual string of reports. The large area of false alarm remained over Oklahoma and eastern Texas, and was not reduced much by the inclusion of environmental information. When all of the environmental information is included in the probabilities (Fig. 3.15d), a large area of false alarm persists, particularly in south-central Texas, far from the majority of the reports. In addition, one of the tornado reports included in the UH-only method (Fig. 3.15a) now is outside the envelope of probabilities.

This case highlights the difficulties of calculating probabilities in MCS situations. While the mode is often easily discernable when looking at simulated

reflectivity, the presence of UH within the squall lines and the presence of ingredients conducive to tornadogenesis in systems presents a difficult problem. Further, MCSs occasionally do produce tornadoes, and ideally probabilities would reflect this potential. It is beyond the scope of this work to lower the probabilities when the expected mode is linear in nature, while maintaining probabilities that reflect the MCS tornado threat.

3.3.3 Subjective Verification

Subjective verification of the tornado probabilities took place during SFE 2015, from 5 May – 4 June 2015. Each participant was asked on a daily basis to rate the four probabilities from the previous day generated using a UH threshold of 75 m^2s^{-2} and a σ of 50 km. In the case of Monday, the most active day from the previous weekend was considered. These ratings ranged from 1 (Very Poor) to 10 (Very Good), in response to the question:

> "Subjectively rate the NSSL-WRF 24 h tornado probabilities using a rating scale of Very Poor (1) to Very Good (10). We are testing the use of updraft helicity as forecast by the NSSL-WRF ensemble to derive tornado probabilities at time and space scale consistent with SPC outlooks. UH $\geq$ 75 is used as a proxy for tornadoes and various methods are tested to only consider UH in environments typically supportive of tornadoes."

Incorporation of environmental information produced higher mean subjective ratings (Fig. 3.16) over the UH-only method for the 24 h probabilities. Of the 22 days of evaluation, the LCL/CAPE ratio method had or was tied for the highest average rating on 9 days, the STP method and the LCL/CAPE ratio/STP method had or were tied for the highest average rating on 8 days, and the UH only method had or was tied for the highest average rating on 6 days. UH only and LCL/CAPE ratio were rated the same on four days, and the STP and LCL/CAPE ratio/STP method were rated the same on seven

days. Thus, many participants saw a strong similarity between the STP and the LCL/CAPE ratio/STP method, although the STP method peaks at a higher rating than the LCL/CAPE ratio/STP method.

Overall, the participants' comments described some common themes. Most of the participants found the guidance to be useful, and noted that the incorporation of environmental information focused the area of interest and reduced false alarm as per the aim of this study, with multiple comments such as:

> "All products capture the area, axis, and grouping of the tornado reports very well. The naive UH probabilities show too much false alarm area in SW Oklahoma, but the additional filters correct that area very well."

These comments suggested that forecasters would like to have the probabilities available when they are forecasting, and that they would glean information at-a-glance, rather than mentally integrating all of the ensemble data upon which these probabilities are based.

The participants' main concerns were the high magnitude of probabilities on multiple days and displacement of the "bullseye" of high probabilities from eventual tornado reports on multiple days. The high magnitude of the probabilities correspond to relatively high risk categories as assigned by the SPC, resulting in comments such as:

> "Several reports occurred outside of the bullseye of tornado probs, and there were only a few tornadoes in the area in Oklahoma that had probs over 30%, even with the most discriminating filters. 30% is high risk, and the reports did not seem like a high risk day to me."

However, from the objective verification discussed previously, high magnitudes are only slightly overforecast according to the reliability diagrams.

Adding environmental information did occasionally have a downside, as was noted in the 27 June 2015 case study; the STP-inclusive probabilities were occasionally

noted by the participants as too limiting, and excluding tornado reports that the less restrictive methods maintained within low probabilities:

> "Large false alarm areas. However the two tornado reports were near the high probability areas. After filtering, the Wyoming tornado was missed although the false alarm area was greatly reduced."

Forecasters have different opinions about whether it is more important to not miss events or to reduce false alarm, and through the SFE the probabilities were rated by forecasters with a mix of these views.

Finally, the participants noted the difference between days with few tornadoes and days with more tornadoes; namely, that marginal days posed more difficulties due to the weaker environmental parameters naturally present on those days:

> "Some displacement from the area where reports occurred. Max probability value ~ 4X greater than the density of storm reports - so the parameter is running quite hot. Missed event, which probably is more a miss of the underlying forecast than any aspect of the parameter space shown. The filters that included STP reduced the max values, which for this event moved closer to observed report density."

Since the probabilities are mostly ingredients-based, it is to be expected that the days with less favorable environments would produce fewer tornadoes, and that the probabilities would have difficulty pinpointing exactly where these tornadoes would occur.

Overall, the participant comments were positive and reinforced the results produced through objective analysis while providing insight into how a forecaster might utilize these probabilities operationally. They also highlighted areas of potential improvement and concerns, which will be taken into account in future work.

3.4 Summary and discussion

High-resolution models are a very useful resource for forecasters, but the amount of information available from these models continues to grow while the amount of time a forecaster has often is fixed. This work attempts to provide a "first guess" forecast of tornadoes from the high-resolution NSSL-WRF ensemble. Information output by the ensemble, such as UH, STP, LCL height, and SBCAPE/MUCAPE ratio are synthesized into probabilities. The first question addressed by this study asks which UH threshold and σ value maximized both reliability and skill in forecasting tornadoes. Utilizing the area under the ROC curve, CSI, and reliability diagrams, this study suggests a UH threshold of 75 m^2s^{-2} maximizes reliability, while producing graphics of similar smoothness to those already issued operationally and maintaining a high ROC area. Lower thresholds of UH were also considered, but produced large areas of overforecasting. However, all thresholds of UH produced less overforecasting than what was found when considering environmental information, such as STP, without considering UH. Small smoothing radii greatly overforecasted and produced noisy graphics; using a larger σ ensures that the probabilities are not tied to specific UH tracks within the model.

When our results are compared to the calibrated tornado forecasts of Jirak et al. (2014), they demonstrate higher CSI at UH thresholds above 25 m^2s^{-2}. As Jirak et al. (2014) used calibrated probabilities based on historical relative frequencies, these probabilities have the advantage of higher CSIs while not requiring historical report information. Reliabilities between the two studies were comparable, and both performed more poorly than the SPC Day 1 Outlooks reported by Jirak et al. (2014). However, the

addition of higher-resolution ensemble data appears to improve the CSI of these uncalibrated probabilities beyond the calibrated probabilities using coarser-resolution environmental information, suggesting that the higher-resolution environmental information benefits the probabilities.

The second question of this study asked whether the incorporation of environmental information to UH information would improve the probabilities. While ROC areas decreased slightly with the addition of environmental information across all UH and σ thresholds, CSI increased. ROC area reduction is thought to be due to lower skill at very low probability thresholds and the large influence of correct negatives, as supported by the CSI. However, the inclusion of environmental information reduced the area of false alarm in many individual cases, STP generally more so than LCL height and CAPE ratio. The inclusion of environmental information also led to an improvement in reliability across all cases.

Subjectively, this finding was supported by participants during SFE 2015, in their comments and their ratings, which favored the probabilities incorporating environmental information over the UH-only probabilities. Both verifications suggest that high-resolution environmental information helps distinguish tornadoes from other severe convective hazards. Subjective evaluation also suggests that these probabilities are useful to forecasters, particularly from SFE 2015 participant comments. The integration of environmental parameters with UH values into one map of probabilities saves forecasters time and effort. To that end, three case studies are presented in which the probabilities could give forecasters an idea of tornado threat. An overwhelmingly mixed-mode day and a day with the potential for tornadoes in a climatologically less-

favored area for tornadoes than the central Great Plains show the ability of the probabilities to handle a multiple tornadic scenarios. A third case demonstrates weaknesses of the probabilities, and provides focus of the future work.

Future work includes ongoing collaboration with SPC forecasters on using UH and STP to generate empirically calibrated probabilities. Preliminary results suggest that these probabilities could provide very different guidance from the method described in this study. Future work will also focus on exploring the relationship between model-generated STP and STP obtained from the ROC re-analysis of tornado events, as well as the relationship between model-generated UH and the radar-observed rotational velocity of storms. Future probabilities will be tested in upcoming SFEs and objectively analyzed, to provide the best possible "first guess" tool for forecasters in their pursuit of an accurate tornado forecast.

Acknowledgments

The authors would like to thank Chris Melick and Robert Hepper of the SPC for providing regridded SPC forecasts, as well as Andrew Dean of the SPC for obtaining the environmental and radar data used in the climatological frequency calculation. This material is based upon work supported by the NSF Graduate Research Fellowship under Grant No. DGE-1102691, Project #A00-4125. AJC and SRD were provided support by NOAA/Office of Oceanic and Atmospheric Research under NOAA-University of Oklahoma Cooperative Agreement #NA11OAR4320072, US Department of Commerce. AJC also received support from a Presidential Early Career Award for Scientists and Engineers.

Tables

Table 3.1 A summary of the NSSL-WRF ensemble configurations with differing lateral boundary conditions and initial conditions. All members use WSM6 microphysics, Dudhia shortwave radiation, RRTM longwave radiation, the Noah land surface model, and the MYJ boundary layer. Members with years in parentheses by the ensemble member were only part of the ensemble for that year. Aside from the control NSSL-WRF member and _GFS member, members are initialized using 3 h SREF member forecasts initialized at 2100Z for the initial conditions and lateral boundary conditions.

Ensemble Member	ICs/LBCs	Microphysics	PBL	Radiation	Land-surface
1	00Z NAM	WSM6	MYJ	RRTM/Dudhia	Noah
2	00Z GFS	WSM6	MYJ	RRTM/Dudhia	Noah
3	21Z em_ctl	WSM6	MYJ	RRTM/Dudhia	Noah
4	21Z nmb_ctl	WSM6	MYJ	RRTM/Dudhia	Noah
5	21Z nmb_p1	WSM6	MYJ	RRTM/Dudhia	Noah
6	21Z nmm_ctl	WSM6	MYJ	RRTM/Dudhia	Noah
7	21Z nmm_n1	WSM6	MYJ	RRTM/Dudhia	Noah
8	21Z nmm_p1	WSM6	MYJ	RRTM/Dudhia	Noah
9 (2015)	21Z nmb_n1	WSM6	MYJ	RRTM/Dudhia	Noah
10 (2015)	21Z nmb_p2	WSM6	MYJ	RRTM/Dudhia	Noah
11 (2014)	21Z em_n1	WSM6	MYJ	RRTM/Dudhia	Noah
12 (2014)	21Z em_p1	WSM6	MYJ	RRTM/Dudhia	Noah

Figures

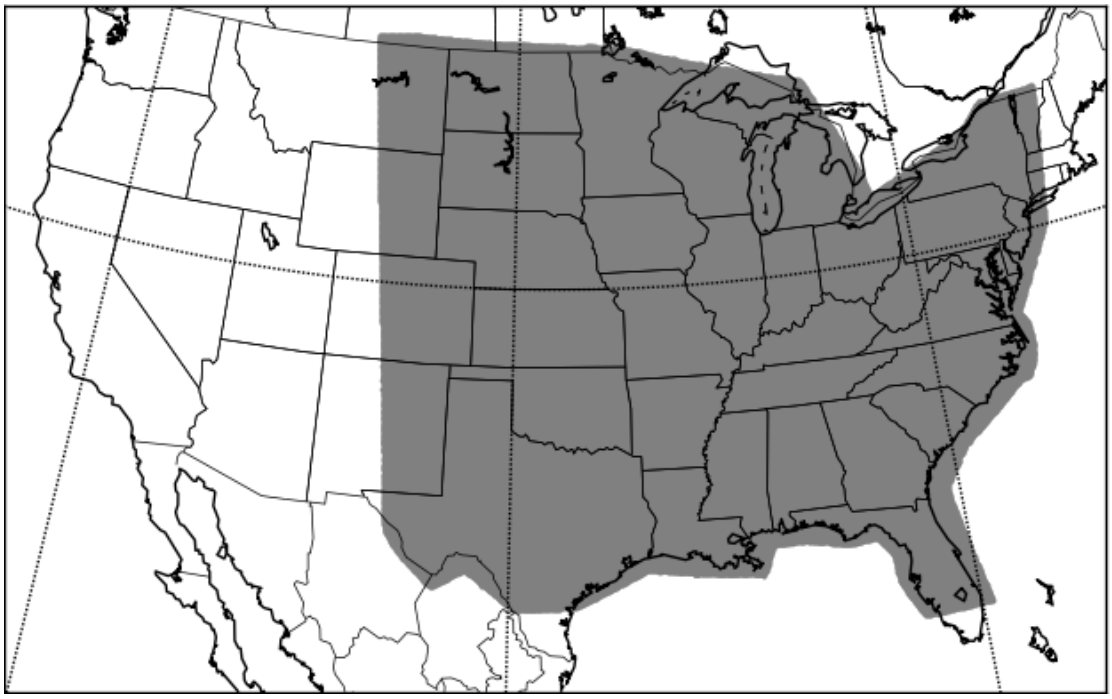

Figure 3.1 The model domain for the NSSL-WRF ensemble. The shaded region shows where objective verification measures were computed.

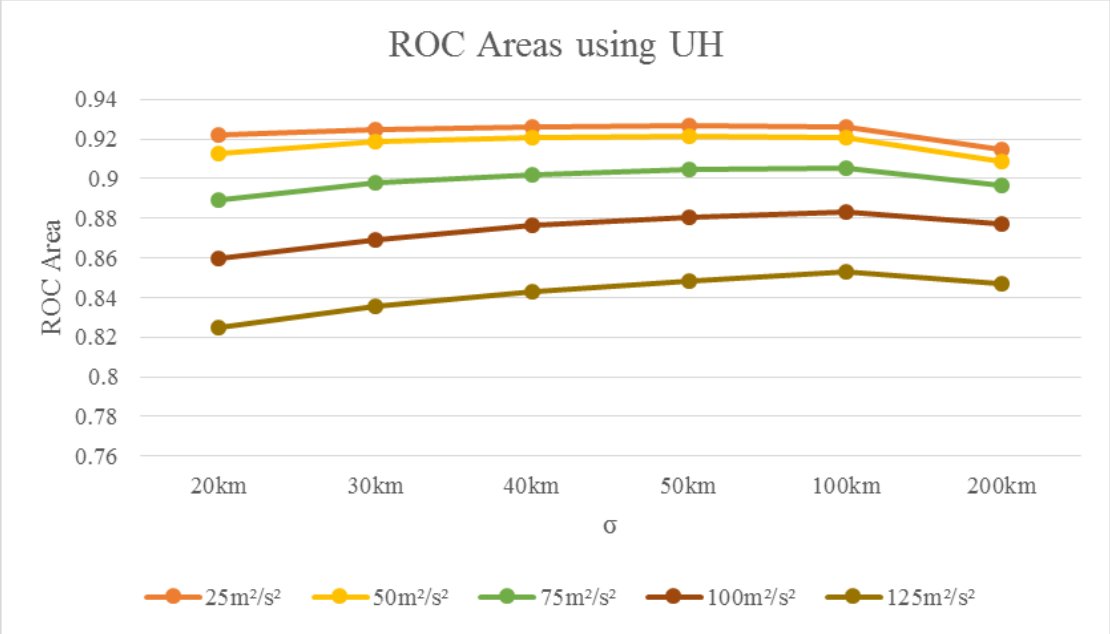

Figure 3.2 ROC areas for tornado probabilities formed using differing σ values and UH thresholds. Different UH thresholds are shown in different colors. All ROC areas are for probabilities formed without incorporation of environmental information.

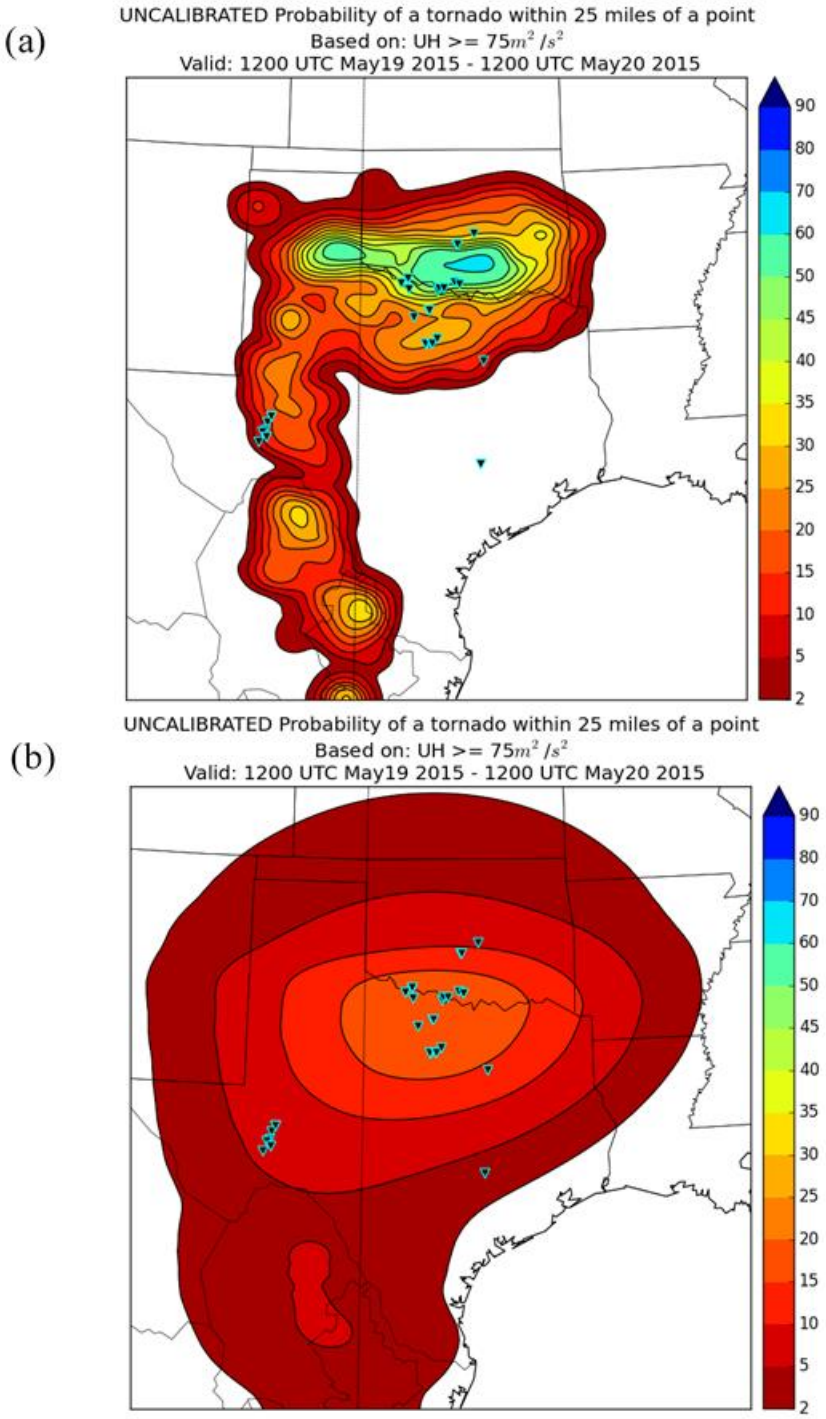

Figure 3.3 Tornado probability maps valid from 1200 UTC 19 May 2015 – 1200 UTC 20 May 2015 for a UH threshold of 75 m^2s^{-2} and a Gaussian kernel of (a) σ = 20km and (b) σ = 200km. Probabilities are shaded contours, and tornado reports are overlaid black inverted triangles with cyan borders.

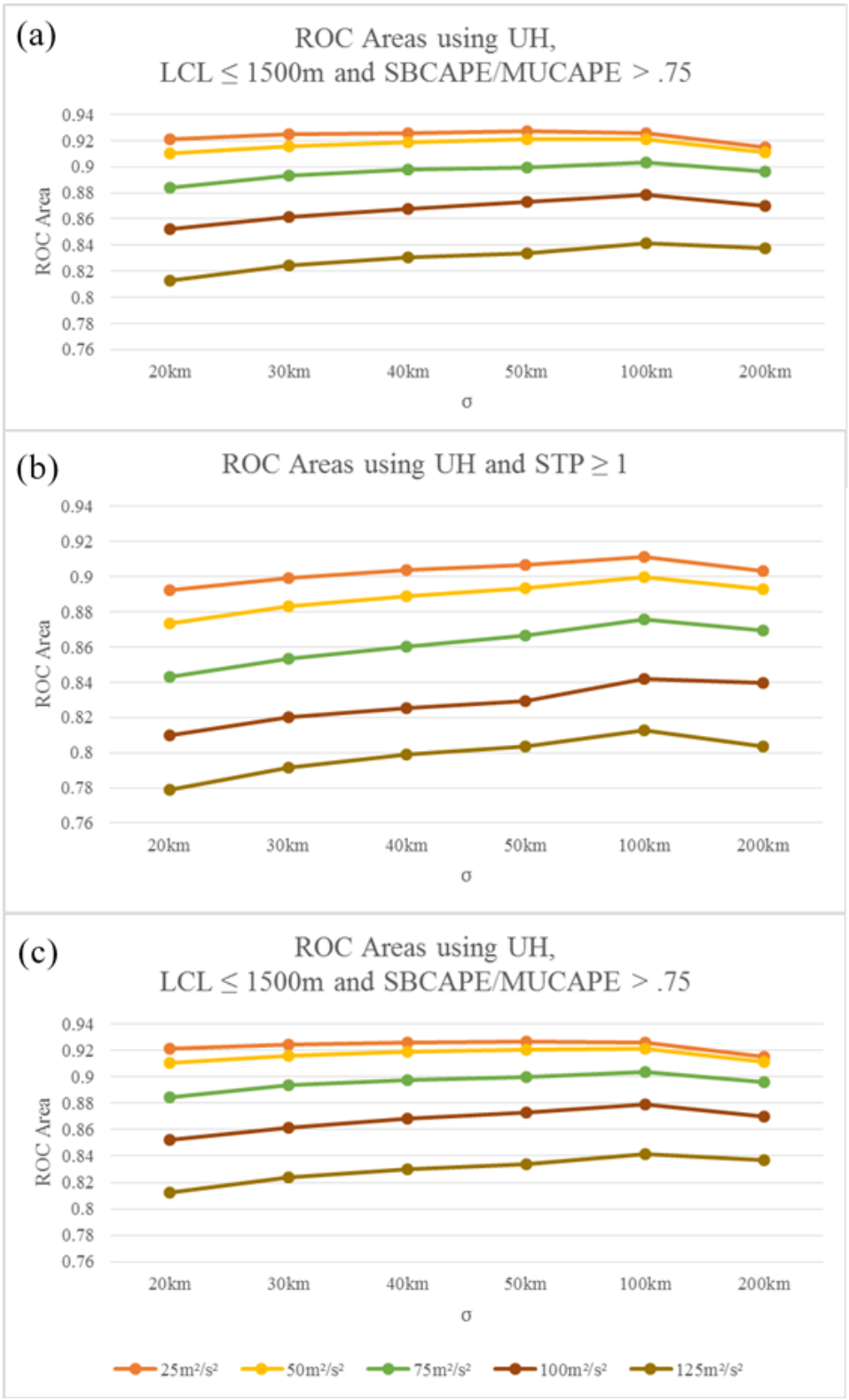

Figure 3.4 ROC areas for tornado probabilities formed using differing σ values and UH thresholds. Different colors represent different UH thresholds. ROC areas are from probabilities incorporating (a) LCL ≤ 1500 m and SBCAPE/MUCAPE > .75, (b) STP ≥ 1, and (c) LCL ≤ 1500 m, SBCAPE/MUCAPE > .75, and STP ≥ 1.

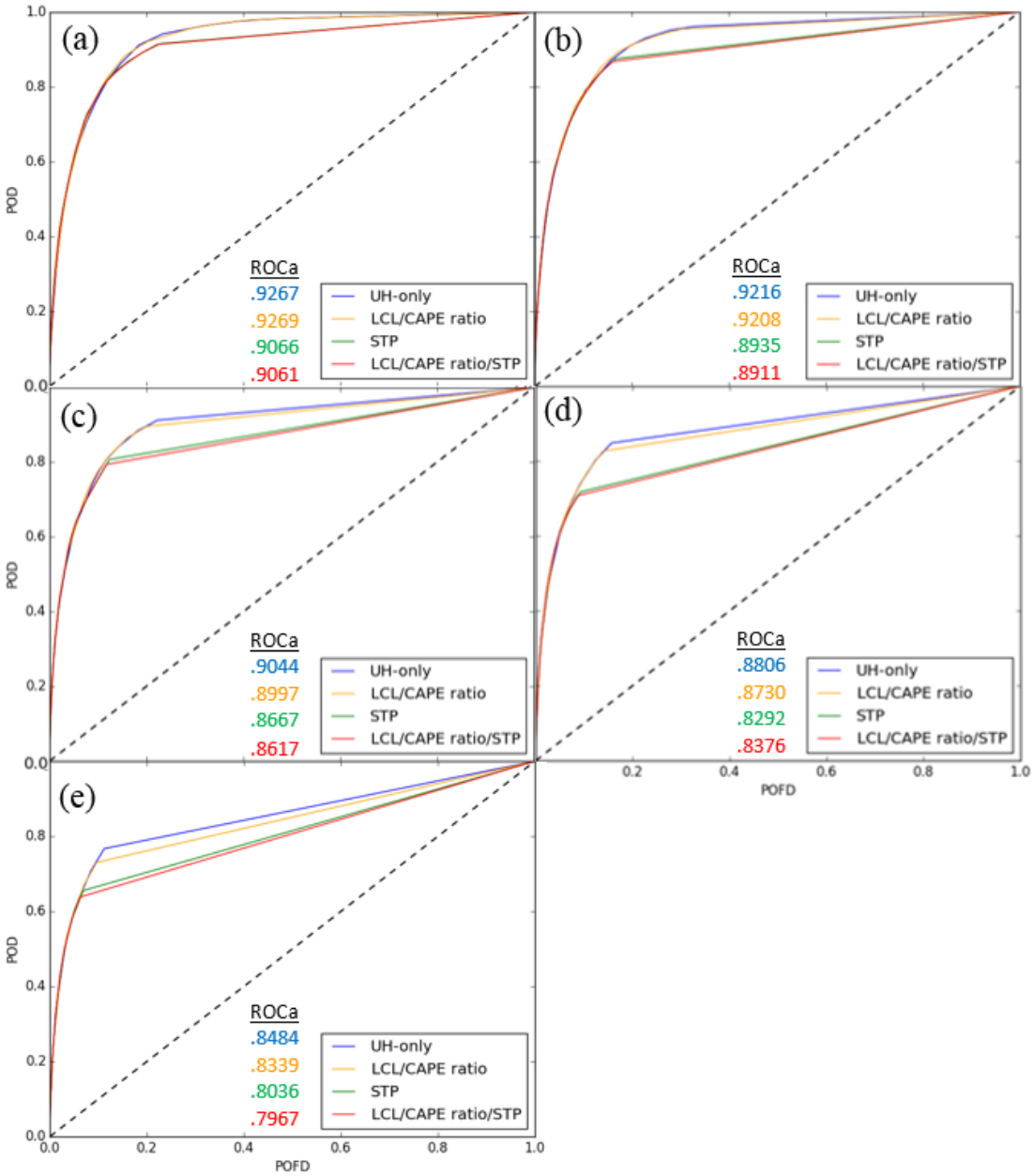

Figure 3.5 ROC curves for σ = 50, four different methods of probability generation, and five different UH thresholds: (a) 25 m^2s^{-2}, (b) 50 m^2s^{-2}, (c) 75 m^2s^{-2}, (d) 100 m^2s^{-2}, and (e) 125 m^2s^{-2}. ROC curves show the probability of detection (POD) vs. the probability of false detection (POFD). Different colors represent methods of probability generation, and ROC areas are listed beside the legend. The dashed diagonal represents the ROC curve that a random forecast would create, and is a reference for comparison.

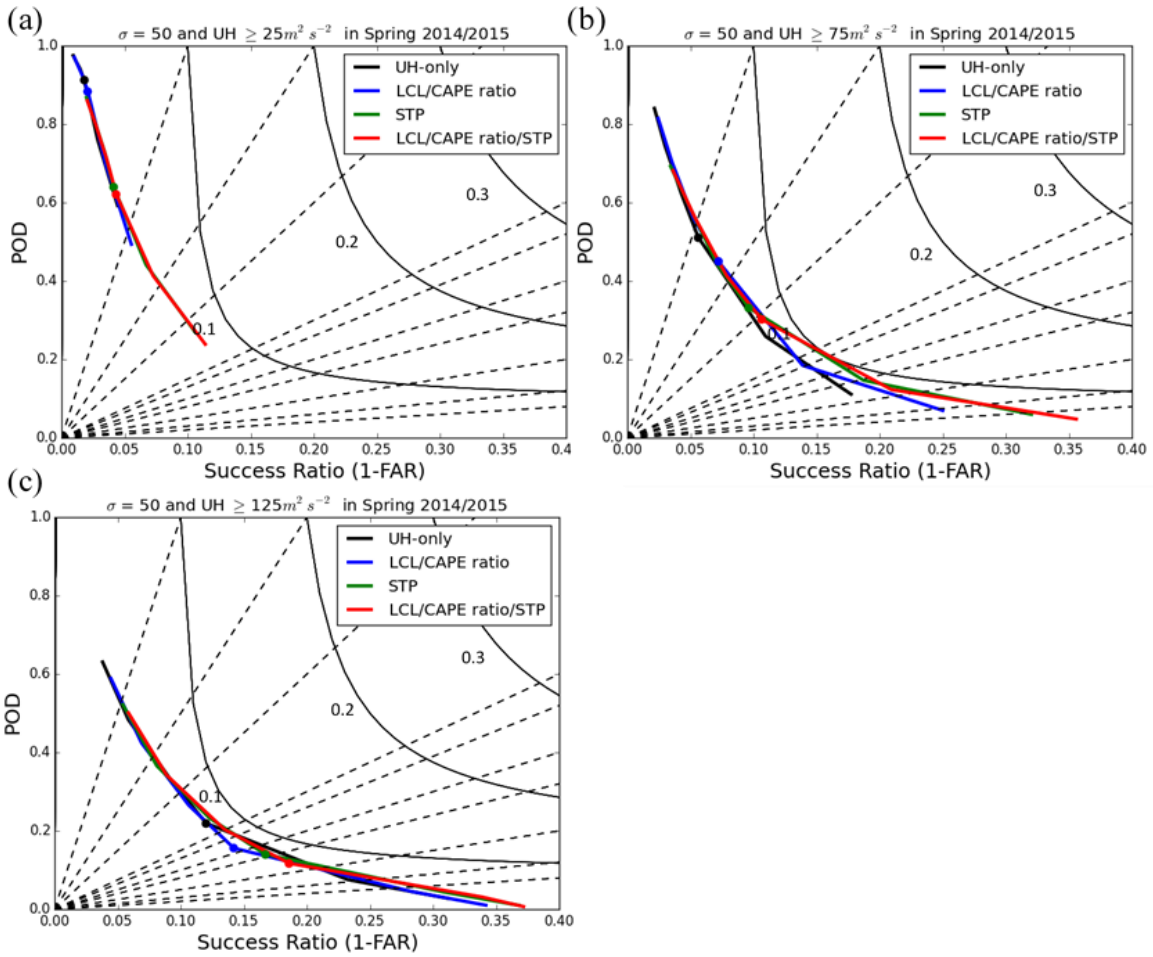

Figure 3.6 Performance diagrams with σ = 50 corresponding to differing UH thresholds: (a) 25 m^2s^{-2}, (b) 75 m^2s^{-2}, and (c) 125 m^2s^{-2}. Colored curves represent the POD plotted vs. the success ratio (1-FAR) at all probability levels forecasted, and the colored dot highlights 15% probability. Dashed lines are of constant bias, and curved lines are of constant CSI. Probability methods include: UH only (black); LCL < 1500 m and SBCAPE/MUCAPE > .75 (blue); STP ≥ 1 (green); and LCL < 1500 m, SBCAPE/MUCAPE > .75, and STP ≥ 1 (red).

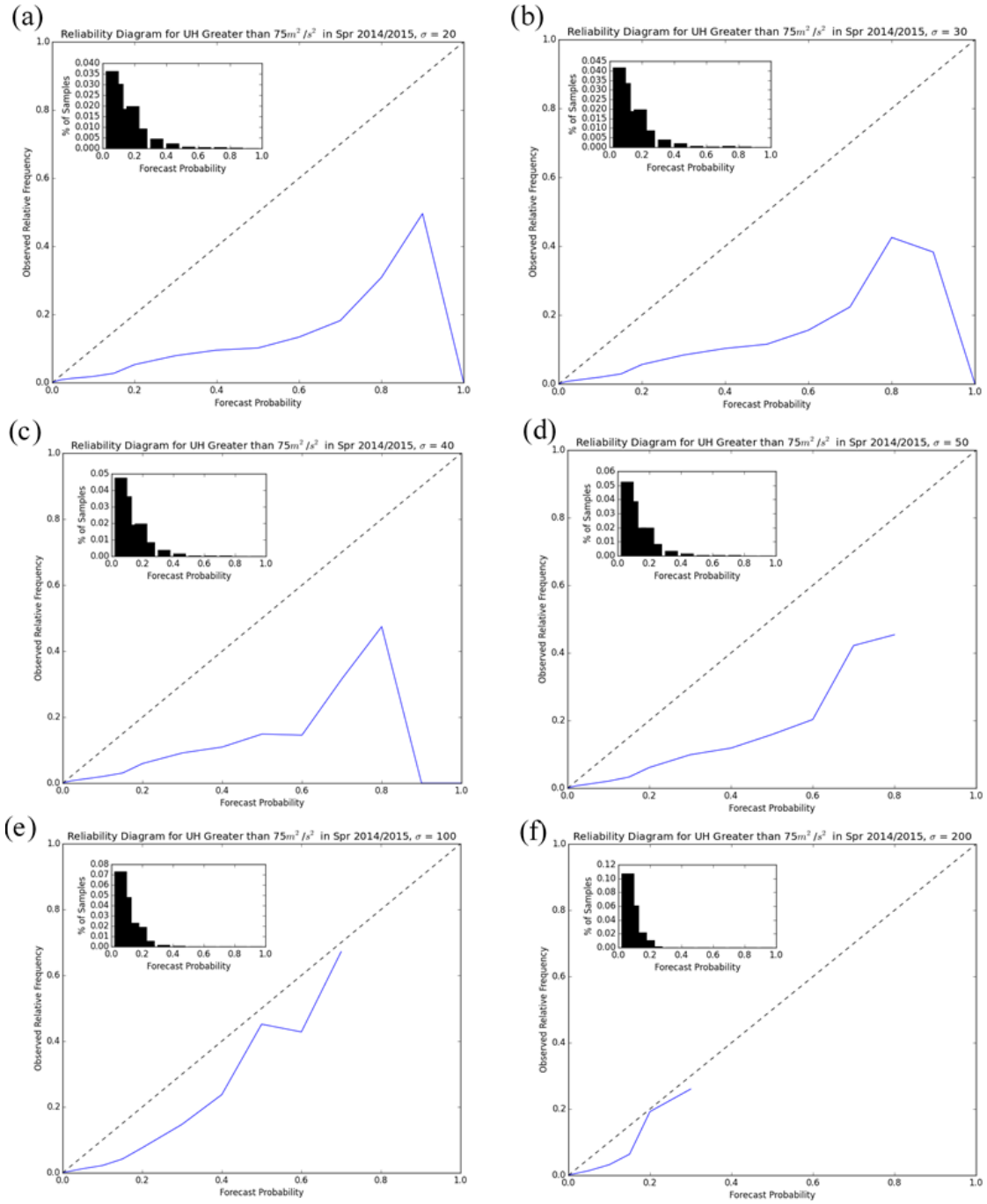

Figure 3.7 Reliability diagrams for tornado probabilities solely incorporating UH > 75m^2s^{-2} and Gaussian smoothing kernel σ values of: (a) σ = 20 km, (b) σ = 30 km, (c) σ = 40 km, (d) σ = 50 km, (e) σ = 100 km, and (f) σ = 200 km. The dashed black line indicates perfect reliability, area above the line indicates underforecasting, and area below the line indicates overforecasting. Histograms in the corner show the percentage of samples in each forecast probability bin, with the 0% bin excluded for clarity due to its overwhelming majority of samples.

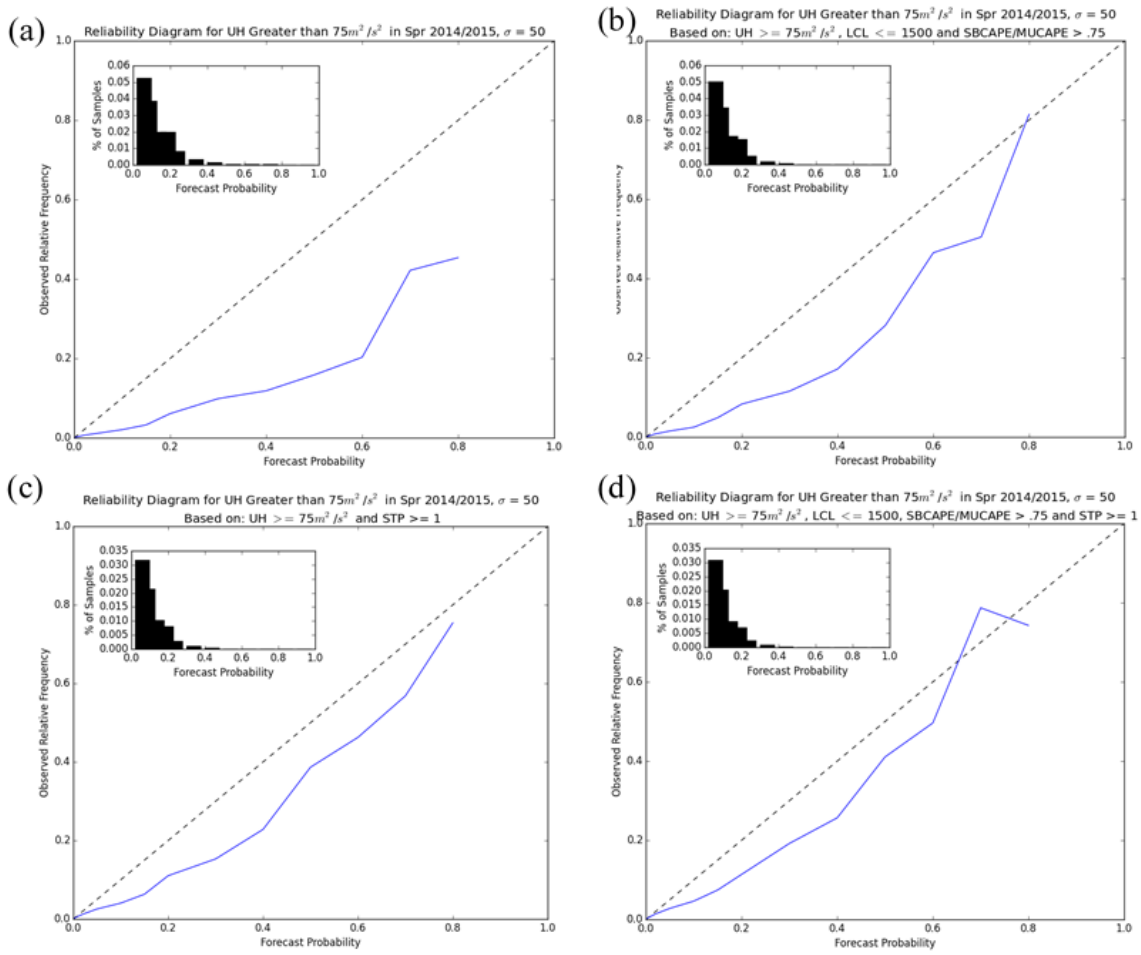

Figure 3.8 Reliability diagrams for tornado probabilities with a UH threshold of 75 m^2s^{-2} and Gaussian smoothing kernel σ values of σ =50 km for (a) no additional environmental information; (b) LCL < 1500 m and SBCAPE/MUCAPE > .75; (c) STP ≥ 1; and (d) LCL < 1500 m, SBCAPE/MUCAPE > .75, and STP ≥ 1. The dashed black line indicates perfect reliability.

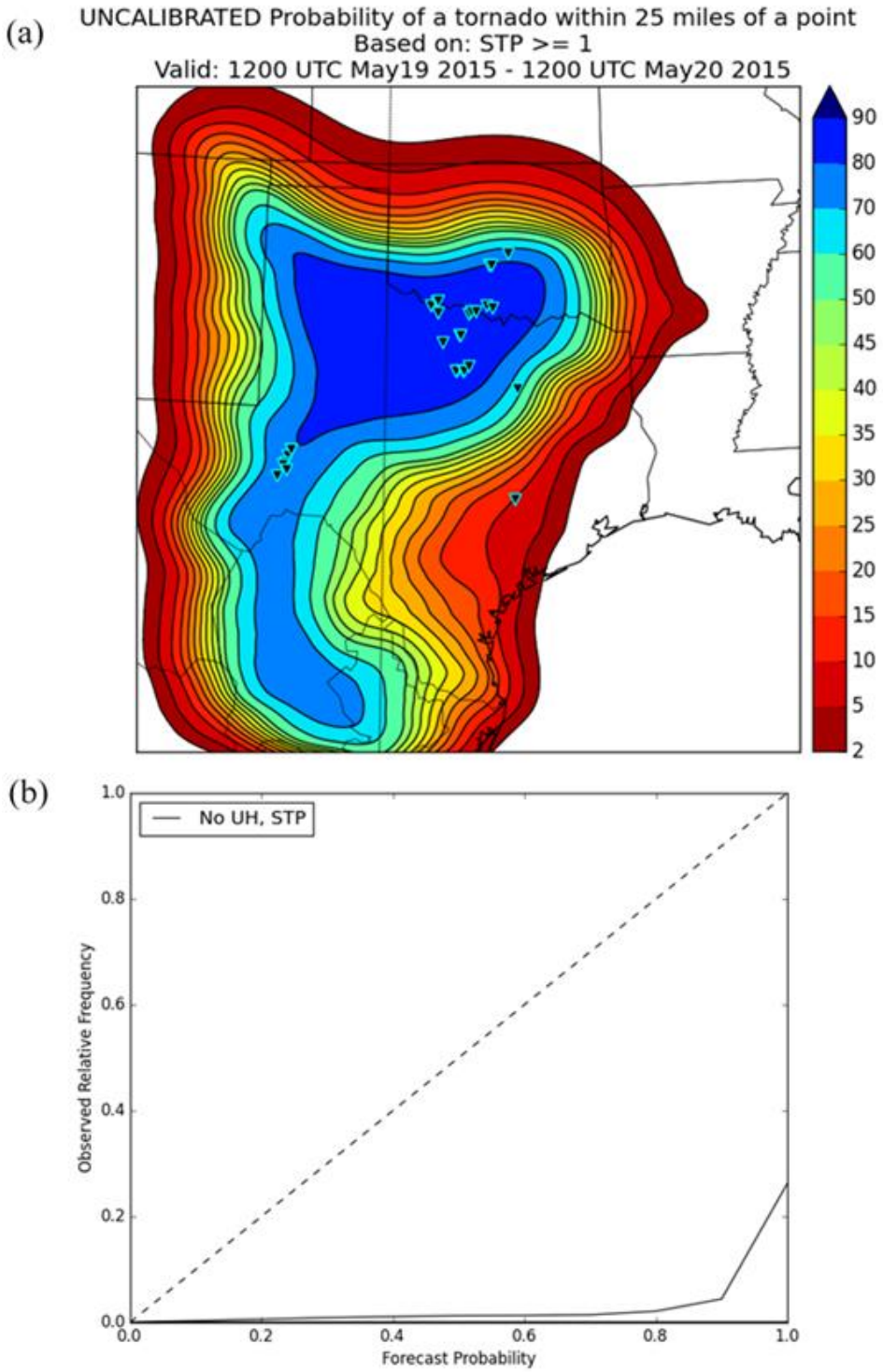

Figure 3.9 (a) Tornado probability map valid from 1200 UTC 19 May 2015 – 1200 UTC 20 May 2015 generated solely using STP ≥ 1 and σ = 50 km, with tornado reports as overlaid black inverted triangles with cyan borders and (b) the reliability diagram for Spring 2014-2015 for probabilities using solely STP ≥ 1. The dashed black line indicates perfect reliability.

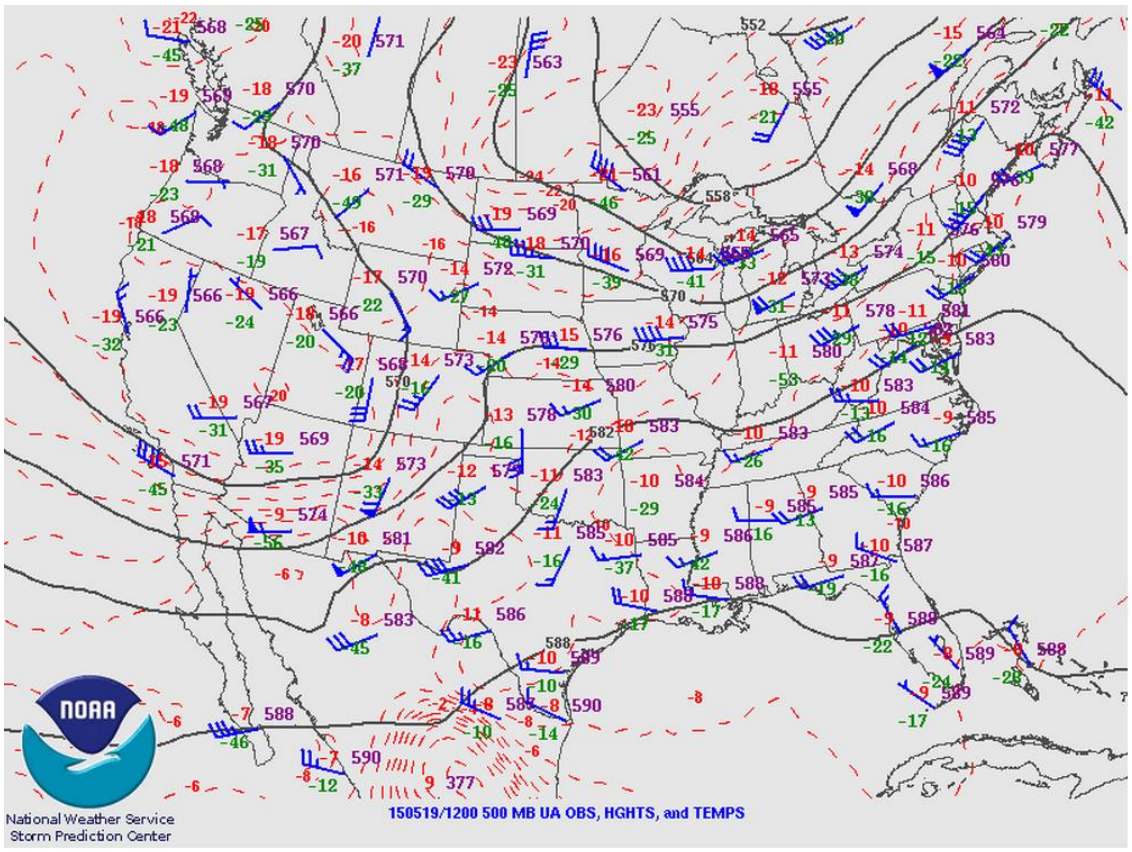

Figure 3.10 A 500 hPa map valid at 1200 UTC on 19 May 2015. Solid black lines are isobars, dashed red lines are isotherms, and blue barbs are 500 hPa wind speed and direction. Pressures (purple), temperatures (red), and dewpoints (green) at observation points are also shown. Obtained from the SPC website: www.spc.noaa.gov/exper/archive/event.php?date=20150519.

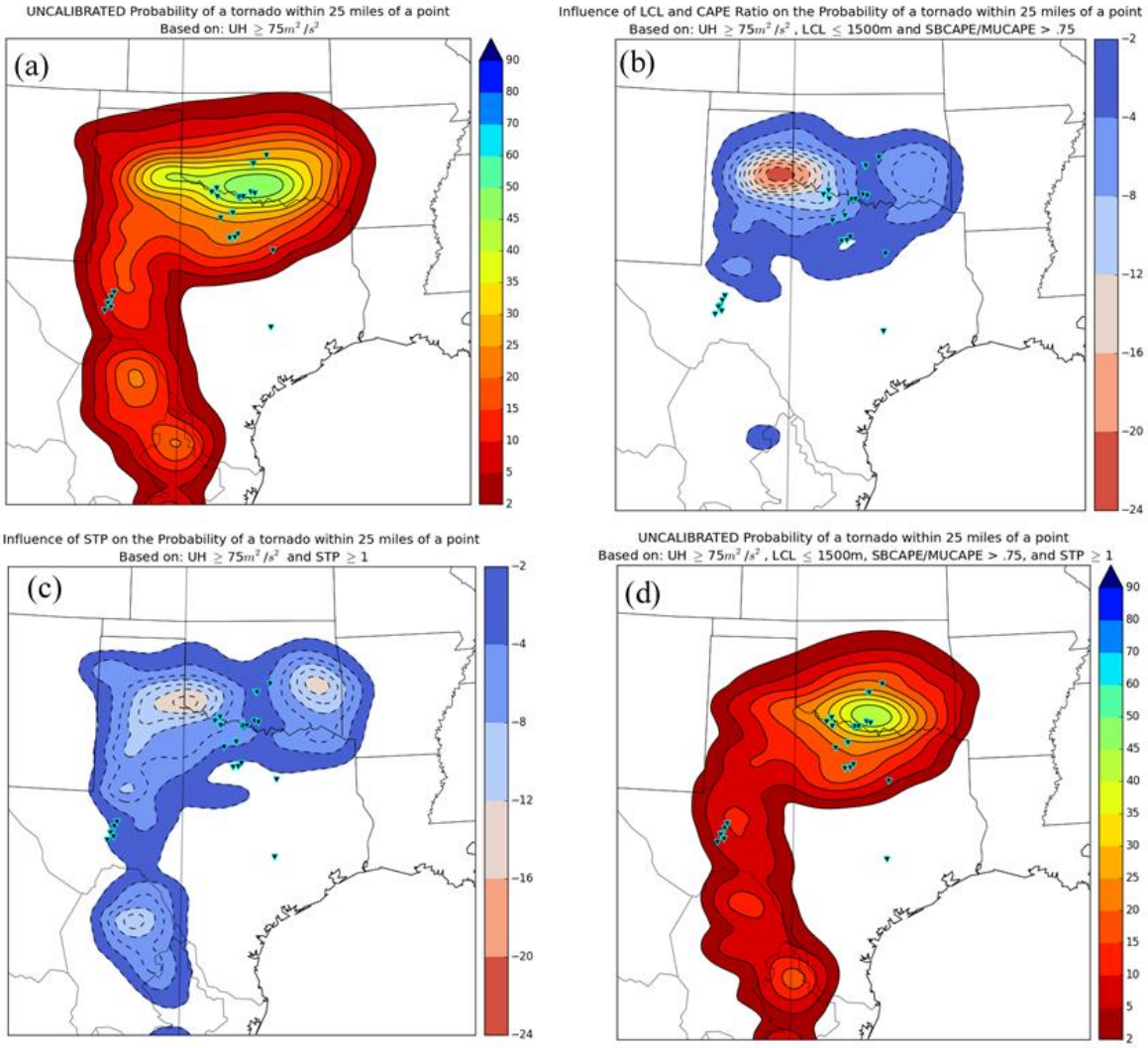

Figure 3.11 (a) Tornado probability map valid from 1200 UTC 19 May 2015 – 1200 UTC 20 May 2015 for a UH threshold of 75 m^2s^{-2} and σ = 50 km generated using solely UH and (d) including environmental information. Probabilities are shaded contours, and tornado reports are overlaid black inverted triangles with cyan borders. (b) and (c) are difference maps between probabilities generated solely using UH and (b) requiring LCL < 1500 m and SBCAPE/MUCAPE > .75; (c) requiring STP ≥ 1. Dashed contours are drawn every 2%, starting at 0%. Negative numbers indicate a reduction in probability compared to (a).

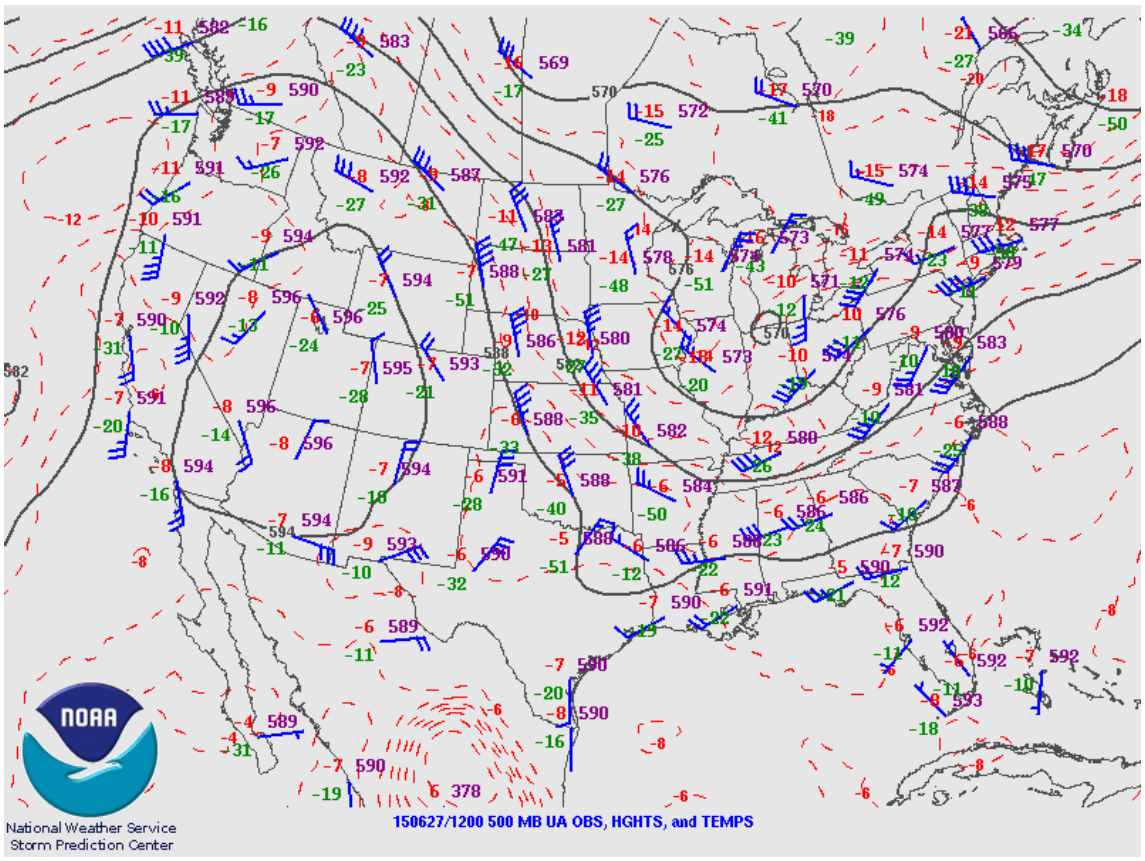

Figure 3.12 A 500 hPa map valid at 1200 UTC on 27 June 2015. Solid black lines are isobars, dashed red lines are isotherms, and blue barbs are 500 hPa wind speed and direction. Pressures (purple), temperatures (red), and dewpoints (green) at observation points are also shown. Obtained from the SPC website: www.spc.noaa.gov/exper/archive/event.php?date=20150627.

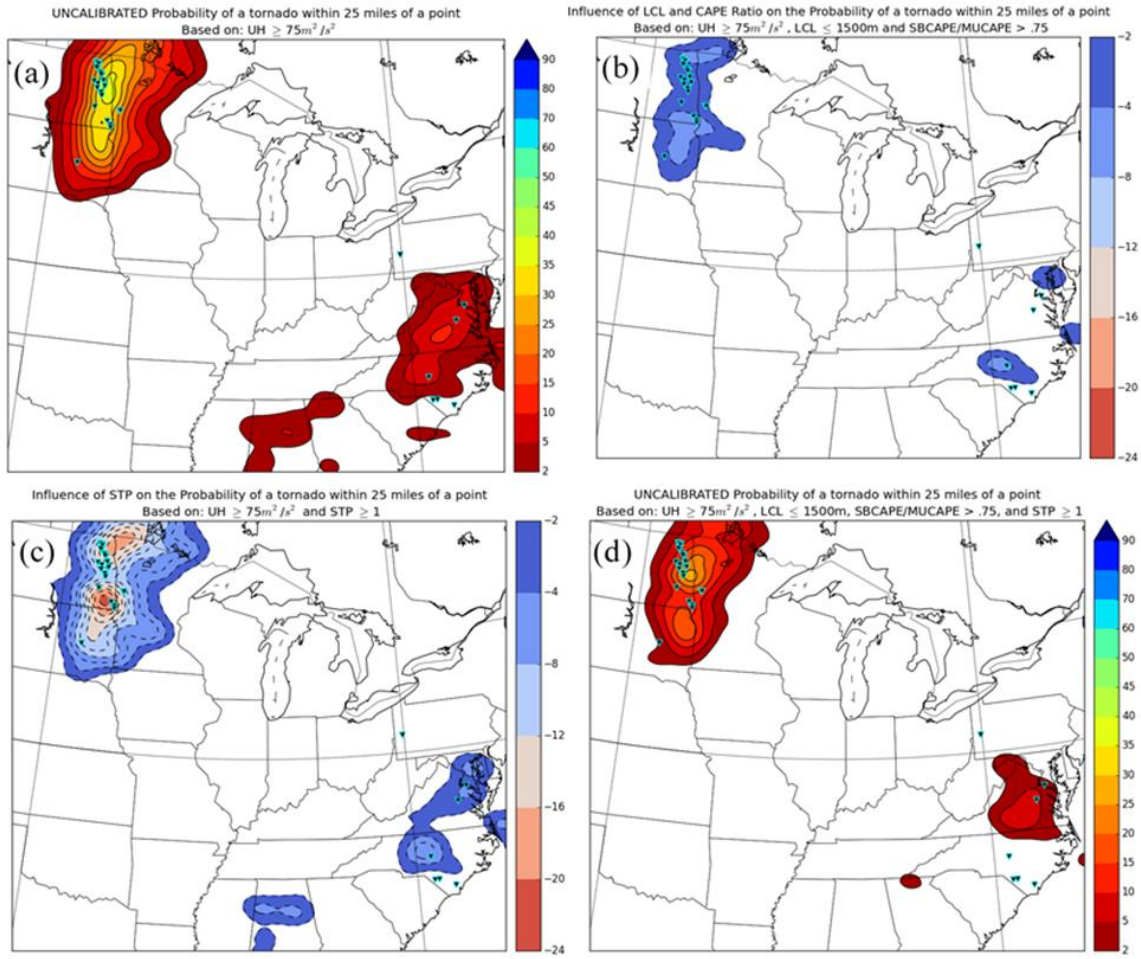

Figure 3.13 (a) Tornado probability map valid from 1200 UTC 27 June 2015 – 1200 UTC 28 June 2015 for a UH threshold of 75 m^2s^{-2} and σ = 50 km generated using solely UH and (d) including environmental information. Probabilities are shaded contours, and tornado reports are overlaid black inverted triangles with cyan borders. (b) and (c) are difference maps between probabilities generated solely using UH and (b) requiring LCL < 1500 m and SBCAPE/MUCAPE > .75; (c) requiring STP ≥ 1. Dashed contours are drawn every 2%, starting at 0%. Negative numbers indicate a reduction in probability compared to (a).

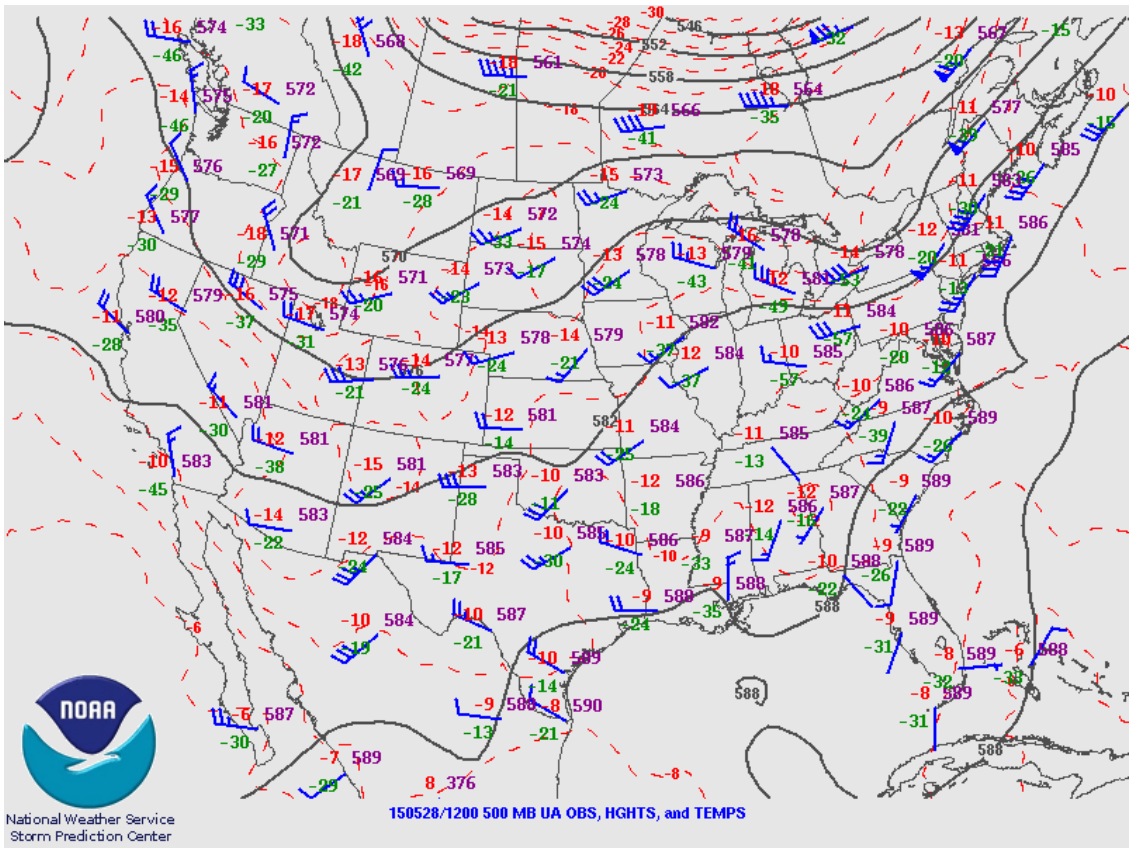

Figure 3.14 A 500 hPa map valid at 1200 UTC on 28 May 2015. Solid black lines are isobars, dashed red lines are isotherms, and blue barbs are 500 hPa wind speed and direction. Pressures (purple), temperatures (red), and dewpoints (green) at observation points are also shown. Obtained from the SPC website: www.spc.noaa.gov/exper/archive/event.php?date=20150528.

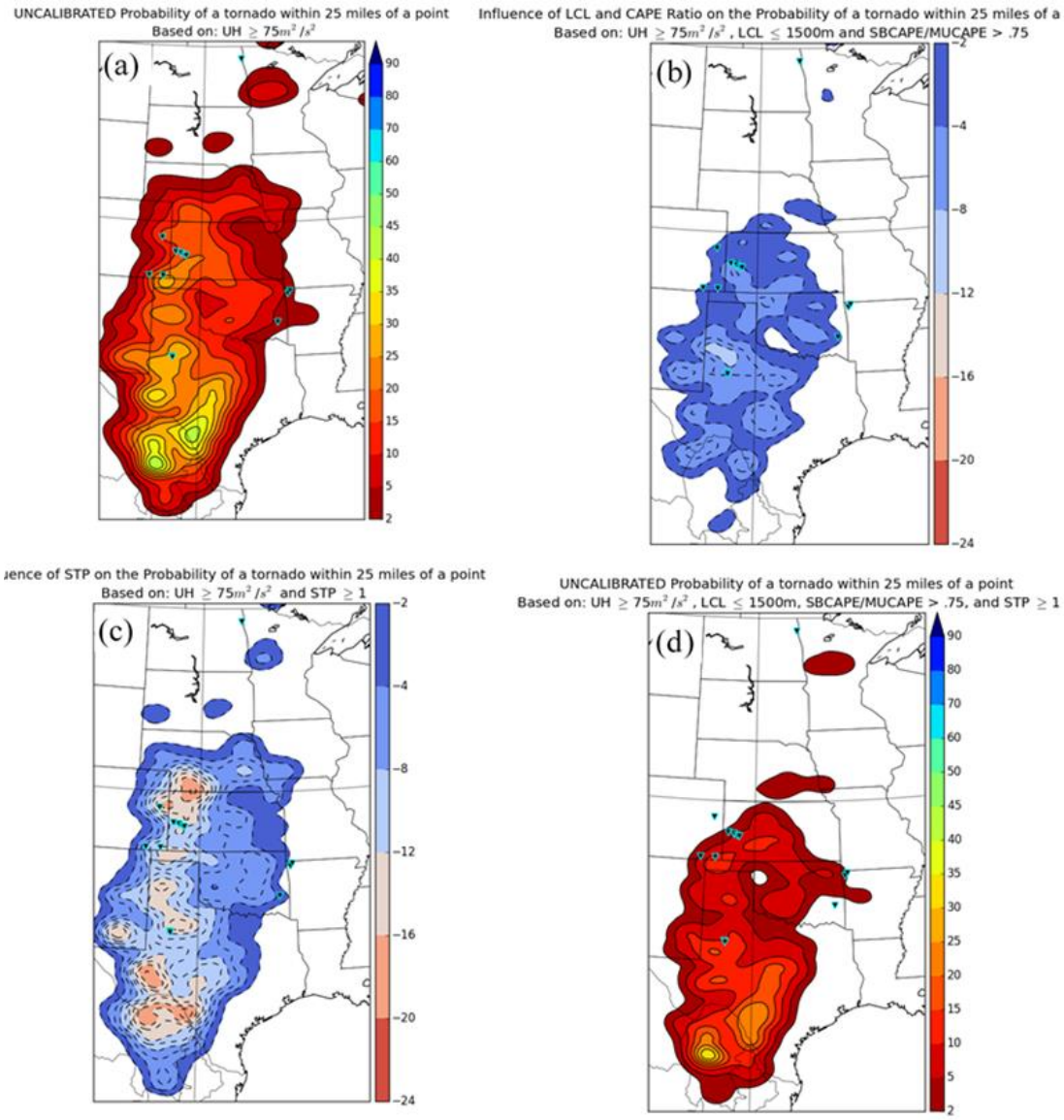

Figure 3.15 (a) Tornado probability map valid from 1200 UTC 28 May 2015 – 1200 UTC 29 May 2015 for a UH threshold of 75 m2s-2 and σ = 50 km generated using solely UH and (d) including environmental information. Probabilities are shaded contours, and tornado reports are overlaid black inverted triangles with cyan borders. (b) and (c) are difference maps between probabilities generated solely using UH and (b) requiring LCL < 1500 m and SBCAPE/MUCAPE > .75; (c) requiring STP ≥ 1. Dashed contours are drawn every 2%, starting at 0%. Negative numbers indicate a reduction in probability compared to (a).

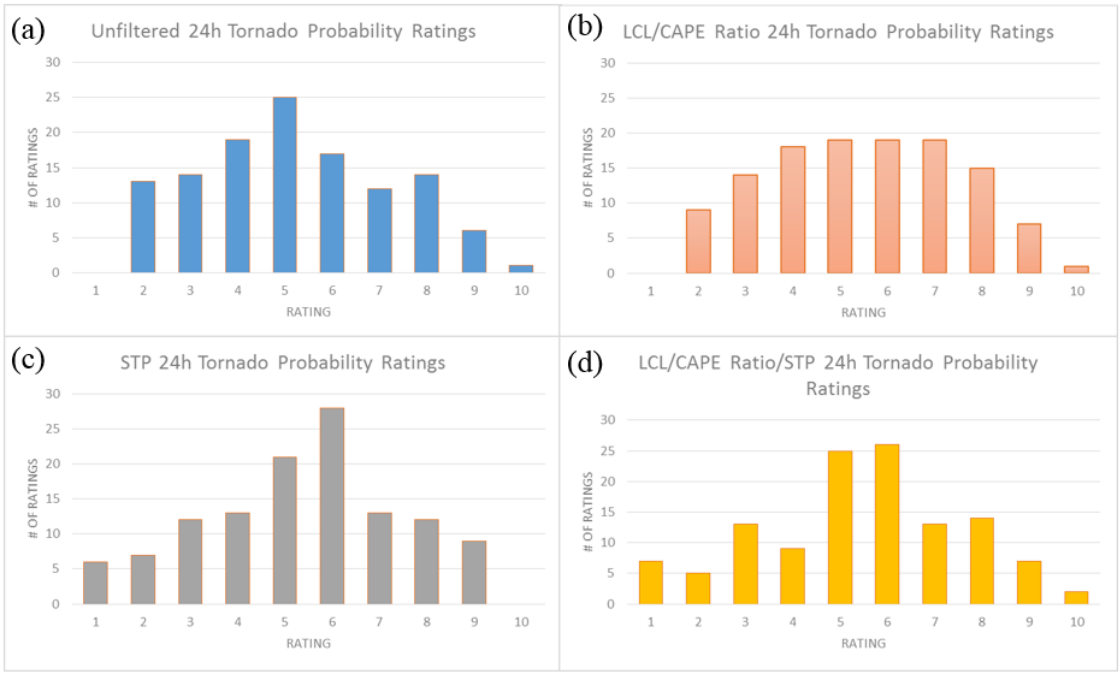

Figure 3.16 Subjective ratings of the tornado probabilities by participants in SFE 2015 for: (a) UH only; (b) requiring LCL < 1500 m and SBCAPE/MUCAPE > .75; (c) requiring STP ≥ 1; and (d) requiring LCL < 1500 m, SBCAPE/MUCAPE > .75, and STP ≥ 1. Ratings encompassed twenty-four cases.

Chapter 4: Blended Probabilistic Tornado Forecasts: Combining Climatological Frequencies with NSSL-WRF Ensemble Forecasts

A paper conditionally accepted by *Weather and Forecasting*

Burkely T.Gallo[1], Adam J. Clark[2], Bryan T. Smith[3], Richard L. Thompson[3], Israel Jirak[3], and Scott R. Dembek[2,4]

[1]School of Meteorology, University of Oklahoma, Norman, Oklahoma

[2]NOAA/OAR National Severe Storms Laboratory, Norman, Oklahoma

[3]NOAA/NWS/NCEP Storm Prediction Center, Norman, Oklahoma

[4]Cooperative Institute for Mesoscale Meteorological Studies, University of Oklahoma, Norman, Oklahoma

Abstract

Attempts at probabilistic tornado forecasting using convection-allowing models (CAMs) have thus far used CAM attribute [e.g., hourly maximum 2–5 km updraft helicity (UH)] thresholds, treating them as binary events—either a grid point exceeds a given threshold or it does not. This study approaches these attributes probabilistically, using empirical observations of storm environment attributes and the subsequent climatological tornado occurrence frequency to assign a probability that a point will be within 40 km of a tornado, given the model-derived storm environment attributes. Combining empirical frequencies and forecast attributes produces better forecasts than solely using mid- or low-level UH, even if the UH is filtered using environmental parameter thresholds.

Empirical tornado frequencies were derived using severe right-moving supercellular storms associated with a local storm report (LSR) of a tornado, severe wind, or severe hail for a given significant tornado parameter (STP) value from Storm Prediction Center (SPC) mesoanalysis grids in 2014–2015. The NSSL-WRF ensemble produced the forecast STP values and simulated right-moving supercells, which were identified using a UH exceedance threshold. Model-derived probabilities are verified using tornado segment data from just right-moving supercells and from all tornadoes, as are the SPC-issued 0600 UTC tornado probabilities from the initial Day 1 forecast valid 1200 UTC–1159 UTC the following day. The STP-based probabilistic forecasts perform comparably to SPC tornado probability forecasts in many skill metrics (e.g., reliability) and thus could be used as first-guess forecasts. Comparison with prior methodologies shows that probabilistic environmental information improves CAM-based tornado forecasts.

4.1 Introduction

Discriminating a tornado threat from an overall severe convective threat poses a unique forecast challenge. Forecasters incorporate knowledge of internal storm dynamics and environments conducive to tornadogenesis, a thorough understanding of current observations, and numerical weather prediction (NWP) to forecast tornadoes. Until very recently, NWP has been too coarse to depict specific storm modes, but recent expansion of computational resources has enabled models that explicitly depict convection and can thus provide specific information on mode, initiation, and evolution (Kain et al. 2008; Clark et al. 2012a).

Several parameters have been associated with environmental conditions supportive of supercells, which can produce tornadoes. Supercell environments require enough convective available potential energy (CAPE) to maintain convection and strong deep-layer shear to create midlevel rotation (Weisman and Klemp 1982, 1984, 1986; Weisman and Rotunno 2000). Supercells produce all types of severe convective weather (defined herein as hail ≥ 2.54 cm in diameter, thunderstorm wind gusts ≥ 25 m s^{-1}, and tornadoes). However, distinguishing which storms in an environment will become tornadic is more difficult than determining if environmental conditions could support supercells, and remains a large forecast challenge (Anderson-Frey et al. 2016). Environments conducive to supercell-based tornadogenesis typically have low lifted condensation levels (LCLs) and high 0–1 km storm-relative helicity (SRH; Rasmussen 2003; Craven and Brooks 2004; Thompson et al. 2012). Thompson et al. (2003) combined these parameters into the fixed-layer significant tornado parameter (STP), which attempts to distinguish significantly tornadic (EF2+) environments from non-tornadic environments. The formulation was then updated by Thompson et al. (2012) to incorporate convective inhibition (CIN) and effective shear terms:

$$STP = \frac{MLCAPE}{1500\ J\ kg^{-1}} * \frac{EBWD}{20ms^{-1}} * \frac{ESRH1}{150m^2s^{-2}} * \frac{(2000m-MLLCL)}{1000m} * \frac{(200+MLCIN)}{150Jkg^{-1}}, \quad (4.1)$$

where MLCAPE, MLCIN and MLLCL are the CAPE, CIN and LCL calculated using the lowest 100 hPa mean parcel, EBWD is the effective bulk wind difference, and ESRH1 is the effective storm relative helicity [calculated using the Bunker et al. (2000) storm motion estimate]. If the STP is ≥ 1.0, the environment is more supportive of significant tornadoes.

STP as a composite parameter also better discriminates between weak and significant right-moving supercellular (RM) tornadoes than individual thermodynamic or kinematic parameters (Thompson et al. 2013). Smith et al. (2015) examined tornadic storms from 2009–2013 within 101 miles of a WSR-88D, creating conditional probabilities of maximum hourly tornado intensity based on the maximum STP within 80 km of each tornadic storm. Larger STPs yielded generally stronger tornadoes in a grid point hour, further extending the application of STP as a discriminatory parameter.

While potential storm environment evolutions depicted by convection-parameterizing NWP helps forecasters understand large-scale environmental conditions, key storm characteristics depend on smaller-scale features such as boundaries (Markowski et al. 1998; Boustead et al. 2013) and storm-to-storm interactions (e.g., Klees et al. 2016). These fine-scale details, which CAMs can depict, often determine how convective mode and subsequent hazards evolve (Fowle and Roebber 2003). CAMs also supply storm-scale metrics such as hourly maximum updraft helicity (UH; Kain et al. 2010), which has been successfully used as a midlevel (Kain et al. 2008; Clark et al. 2012b) and low-level (Sobash et al. 2016b) mesocyclone-scale rotation diagnostic. Swaths of positive UH typically indicate simulated right-moving supercells (similarly, swaths of negative UH typically depict simulated left-moving supercells). Since supercells often generate severe weather reports, UH can indicate severe storm occurrence in both deterministic (Sobash et al. 2011) and ensemble frameworks (Sobash et al. 2016a).

Extending UH application from severe convective forecasting to tornado forecasting has begun in recent years. Taking a countrywide perspective, daily

accumulated UH swaths positively correlate with total tornado path length over the CONUS (Clark et al. 2013). On an individual storm level, Sobash et al. (2016b) argue that 0–3 km UH can serve as a tornado proxy by showing that simulated storms with strong low-level mesocyclone-scale rotation occur in simulated environments with STP and individual kinematic and thermodynamic parameters similar to observed proxy soundings from tornadic storm environments. Combining UH and environmental information can also help parse the tornado threat from the overall severe convective threat (Jirak et al. 2014; Gallo et al. 2016). Since simulated mesocyclones often occur in environments unfavorable to tornadogenesis (Clark et al. 2012b), environmental criteria can reduce false alarms by limiting probabilistic tornado forecasts to favorable environments (Rasmussen and Blanchard 1998; Rasmussen 2003; Thompson et al. 2003; Grünwald and Brooks 2011; Grams et al. 2012; Thompson et al. 2012; Thompson et al. 2013). Indeed, both coarse-scale (Jirak et al. 2014) and fine-scale (Gallo et al. 2016) environmental information demonstratively improves tornado guidance skill beyond forecasts generated solely using UH.

This work blends CAM environmental and storm-scale output with observed, empirical frequencies of a tornado of any intensity given environmental characteristics from right-moving supercells. Smith et al. (2015) developed initial frequencies from environmental tornado climatologies, which Thompson et al. (2017) improved upon by determining the frequency of a tornado given a right-moving supercell [as defined by Smith et al.(2012)] with a Local Storm Report (LSR) using data from 2014 and 2015. By applying these observed frequencies to the NWP output, this study creates forecasts resembling Storm Prediction Center (SPC) convective outlooks using a paradigm that

represents each point as having a probability of tornado occurrence rather than assuming a tornado if deterministic attribute thresholds are exceeded. This process was also designed to reduce the over-forecasting seen in prior probabilistic tornado forecasts (Jirak et al. 2014; Gallo et al. 2016; Sobash et al. 2016b) by constraining the magnitude of the probabilities to observed frequencies roughly based on the environmental probabilities from Thompson et al. (2017). The forecasts produced by this methodology are also compared to other methods of probability generation described in the literature, including using 2–5 km UH or 0–3 km UH as a tornado proxy sans environmental information [as in Sobash et al. (2016b)], or by requiring 2–5 km UH exist in an environment exceeding a threshold of STP [as in Gallo et al. (2016)].

Section 4.2.1 of this paper describes the modified STP used throughout this study, which is a surface-based parcel and fixed-layer shear version of the effective-layer STP (Thompson et al. 2012). Section 4.2.2 describes the empirical climatological frequency generation, while section 4.2.3 outlines the ensemble system and probabilistic forecast generation algorithm. Sections 4.2.4 and 4.2.5 specify SPC forecasts and objective verification metrics used in this study, respectively. Determination of the optimum STP percentile composes section 4.3.1, while section 4.3.2 compares four probability generation methods and the 0600 UTC SPC forecasts. Case studies in Section 4.3.3 illustrate the daily tornado probabilities on two high-end days and a more marginal day. Finally, Section 4.4 summarizes and discusses the results and future research directions.

4.2 Data and Methodology

4.2.1 STP Formulation

The STP calculation herein uses surface-based parcels and fixed layer calculations within the effective-layer STP equation (Thompson et al. 2012):

$$STP = \frac{SBCAPE}{1500\ J\ kg^{-1}} * \frac{SHR6}{20ms^{-1}} * \frac{SRH1}{150m^2s^{-2}} * \frac{(2000m-SBLCL)}{1000m} * \frac{(200+SBCIN)}{150Jkg^{-1}}, \quad (4.2)$$

where the SBCAPE, SBLCL, and SBCIN are the surface-based CAPE, LCL, and CIN. As in the fixed-layer STP, the CAPE and LCL height are calculated from surface-based parcels due to availability constraints within the NSSL-WRF ensemble, and the shear and SRH are computed from fixed layers. Similar to the effective-layer STP, the modified STP includes CIN, albeit calculated from the surface-based parcel rather than the 100 mb mixed layer parcel. Additionally, the capping terms (e.g., if SHR6 < 12.5 kts, the SHR6 term is set to zero) are taken from the effective-layer STP. This STP formulation utilizes improvements within the effective-layer STP while balancing the computational expense of running a CONUS-wide CAM ensemble (i.e., the inability to calculate the effective-layer inflow for each grid point and time on a 4-km grid efficiently).

4.2.2 Tornado Frequency Calculation

The climatological frequency of tornado occurrence was calculated following Thompson et al. (2017), but using the modified STP formulation described in Section 4.2.1. LSRs from 1 February 2014–31 December 2015 were filtered in three ways: (1) all tornado reports were filtered by maximum EF-scale per 40-km grid hour[1], (2) all hail/wind reports were required to meet effective bulk wind difference (Thompson et al.

[1] This study does not use intensity information; this step was performed such that the most intense tornado supported by each environment was used.

2007) criteria ($\geq$ 20 kt for 2014, $\geq$ 40 kt for 2015[2]), and (3) a convective mode filter ensured that only right-moving supercells and right-moving marginal supercells were included. The supercell definition required an azimuthal velocity difference of ≥ 10 m s^{-1} across less than ~7 km throughout more than one quarter of the storm's depth for at least 10–15 minutes (Smith et al. 2012). After filtering, 1202 tornadic cases and 5422 non-tornadic cases were used to generate the climatological frequencies. To ensure separation of the training and testing dataset, weekly frequencies were generated withholding the reports for that week. Each week's frequencies were then used in probability generation. This cross-validation technique (Elsner and Schmertmann 1994) has previously been applied to surrogate severe probabilities (Sobash and Kain 2017). Hourly SPC objective analyses (Bothwell et al. 2002) provided the nearest 40 km grid point modified STP assigned to each event. The weekly climatological tornado frequency in each STP bin equaled the tornadic storm count divided by the total number of storms in that bin (Fig. 4.1). Variability in the equations was largest at high STP values, which have more limited sample sizes than lower STP values.

4.2.3 Probabilistic Forecast Generation

Probabilistic tornado forecasts were generated using output from a 4-km horizontal grid-spacing ensemble based around an experimental version of the Weather Research and Forecasting model (WRF; Skamarock et al. 2008), generated by the National Severe Storms Laboratory (NSSL) using the Advanced Research core WRF

[2] The more strict effective bulk wind difference criteria for 2015 was estimated to reduce the number of potential 40-km grid hour events by ~35% for 2015 based on 2014 data, thereby reducing workload while capturing a majority of the low-level circulations within the sample. For further details, see Thompson et al. (2017).

(WRF-ARW) and known as the NSSL-WRF (Kain et al. 2010). The NSSL-WRF ensemble contains the NSSL-WRF and nine additional members with varied initial conditions and lateral boundary conditions (Gallo et al. 2016; Clark 2017; Table 4.1). Ensemble runs began in February 2014, and produce forecasts to 36-h beginning at 0000 UTC. Probabilistic tornado forecasts were generated for the spring seasons (defined as 1 April–30 June) of 2014 and 2015; seasonal statistics are aggregated over that time. The probabilistic forecasts herein are intended as automated first-guess tornado forecasts for 12–36 h lead time covering the Day 1 period defined by the SPC.

Ensemble membership shifted slightly between June 2014 and April 2015, exchanging two members initialized from Eulerian mass (EM) Short-Range Ensemble Forecast (SREF) members for two members initialized from Non-hydrostatic Multiscale Model on the B-grid (NMB) SREF members. This change occurred when SPC forecasters noticed tight clustering within the EM SREF members compared to other subsets. The ensemble membership shift has minimal impact on subsequent tornado forecasts (Gallo et al. 2016), and therefore the 2014 and 2015 spring seasons are combined.

This work compares four methods of probabilistic forecast generation. Method 1 uses 2–5 km UH $\geq$ 75 m^2s^{-2} as a coarse proxy for tornado occurrence from the daily maximum UH field of each member, as in Gallo et al. (2016) and following the Hamill and Colucci (1998) method for calculating probabilities. Each member has a distribution of UH values from the daily maximum UH within a 40 km radius of a point, and probabilities are generated by determining where 75 m^2s^{-2} occurs within the distribution. Methods 2 and 3 are similar, but use 2–5 km UH $\geq$ 75 m^2s^{-2} only at points

where the preceding hour had STP ≥ 1 or use 0–3 km UH ≥ 33 m^2s^{-2}, respectively. The 0–3 km threshold was chosen by determining the percentile of 2–5 km UH corresponding to 75 m^2s^{-2} during the study period and the subsequent value of 0–3 km UH at that percentile. These three methods are derived from those previously explored in the literature, and solely use output from CAM ensembles.

The final probabilistic tornado forecast method (i.e., Method 4) combines ensemble information and the observed climatological frequencies described in Section 2b (Fig. 4.2). First, forecast hours 12–36 of each ensemble member are checked for 2–5 km UH ≥ 25 m^2s^{-2}, indicating a right-moving supercell (Clark et al. 2013; Gallo et al. 2016; Sobash et al. 2016a). If a gridpoint exceeds the UH threshold, the STP from the prior hour is collected from every point where the threshold is exceeded within a 40-km radius, creating a STP distribution at each gridpoint and for each hour. From these STP distributions, a percentile value is extracted and assigned to the gridpoint and hour. The percentiles examined herein are the 10^{th}, 25^{th}, 50^{th}, 75^{th}, 90^{th}, and 100^{th} (maximum value). Once each gridpoint and hour has a STP value the daily maximum STP is assigned to the point, representing the most favorable environment over a 24-h period. The climatological frequency values are then used to assign a STP-based tornado probability at that gridpoint. The calculated climatological frequency values (Fig. 4.1) represent the centerpoint of their bins, and linear interpolations between the bin centers assign frequencies between centerpoints.

The final step averages the individual member probabilities and smooths the resultant field using a Gaussian kernel density weighting function with weights determined by:

$$f(x,y) = \frac{1}{2\pi\left(\frac{\sigma}{\Delta x}\right)^2} * e^{\frac{-(x^2+y^2)}{2\left(\frac{\sigma}{\Delta x}\right)^2}}, \quad (4.3)$$

where σ is the user-defined standard deviation in km and Δx is the grid spacing. Varying σ were tested (not shown), and σ = 50 km creates a field of comparable resolution to SPC tornado probabilities.

4.2.4 SPC Forecasts

All ensemble probabilities were verified in conjunction with the initial SPC Day 1 tornado probabilities issued at 0600 UTC (valid 1200 UTC–1159 UTC the following day) to compare the skill of the first-guess probabilities and initial SPC tornado forecasts. For these probabilities to become a useful first-guess forecast, the resolution and accuracy should resemble the SPC forecasts. The SPC issues 0600 UTC tornado forecasts using information from 0000 UTC, making them the most applicable comparison to the first-guess forecasts since the ensemble initializes at 0000 UTC. The outlooks herein were largely independent of the NSSL-WRF ensemble probabilities, as the ensemble fields were unavailable to forecasters producing the 0600 UTC outlooks. The SPC probabilities were regridded to the NSSL-WRF grid before verification, ensuring consistency between the ensemble and SPC forecasts.

4.2.5 Verification

Verification occurred across approximately the eastern two-thirds of the CONUS (Fig. 4.3). All probabilities (NSSL-WRF and SPC) were considered only within this domain and over the 182 days of April–June 2014 and 2015. Tornado path data were georeferenced to the 4-km grid of the NSSL-WRF ensemble and treated as

binary yes/no events. Yes events occurred if a tornado passed within 40 km of a point. Though the severe report database has documented shortcomings regarding tornado reports (Doswell and Burgess 1988; Brooks et al. 2003; Verbout et al. 2006; Doswell et al. 2009) and hail reports (Blair et al. 2017), more low-magnitude tornadoes have been reported in recent decades (Brooks and Doswell 2001).

Two subsets of the tornado database were considered for this project. The first subset included tornado path data from all modes of parent convection. The second subset solely included tornadoes produced by either right-moving supercells or marginal right-moving supercells (RM tornadoes). Since the new methodology derives probabilities from observed climatological frequencies of RM tornadoes, applying the forecasts to the second subset is truer to the underlying data than using them as forecasts of all tornadoes. Comparing the verification methods may help determine whether the probabilities are appropriate as tornado forecasts or should solely be considered a forecast of RM tornadoes. The other methods previously documented in the literature were also verified with both datasets.

Forecasts were verified using reliability diagrams (Wilks 2011), performance diagrams (Roebber 2009), and the area under the receiver operating characteristic (ROC) curve, which measures the ability of a forecast to discern an event from a non-event by plotting the probability of detection (POD) against the probability of false detection (POFD) at different thresholds. POD and POFD were generated using a standard 2 X 2 contingency table and defined as:

$$POD = \frac{hits}{hits + misses}, \quad (4.4) \text{ and}$$

$$POFD = \frac{false\ alarms}{false\ alarms + correct\ negatives}. \quad (4.5)$$

One POD and one POFD were defined for each probabilistic tornado forecast threshold that the SPC issues: 2%, 5%, 10%, 15%, 30%, 45%, and 60%. The model forecast verification occurred at these thresholds to enable comparisons. The area under the curve was then computed using the trapezoidal method (Wandishin et al. 2001). ROC areas range from 0.0 to 1.0, where 1.0 indicates a perfect forecast, and 0.5 is the skill of a random forecast. Generally, a score of 0.7 or higher is considered skillful (Buizza et al. 1999). Both seasonally aggregated and daily ROC areas were computed.

The ROC area difference between the SPC forecasts and ensemble forecasts was tested for statistical significance using resampling, following Hamill (1999). All cases were randomly assigned to one of the two forecasts, seasonally aggregated ROC areas were calculated for the two groups, and the difference was computed 1000 times to create a ROC area difference distribution. Significant ROC area differences between the SPC forecasts and the NSSL-WRF ensemble forecasts fell outside of the 95% confidence interval of this subsequent distribution.

Reliability diagrams plot the observed relative frequency against the forecast probability, providing information about bias to supplement the ROC areas, which are insensitive to bias. A perfect forecast follows the 45° diagonal: when there is a 40% probability of a tornado, a tornado observation occurs in four out of ten forecasts. The SPC's forecasts largely occur at low probabilities, and are only issued at specific thresholds: forecasters typically assume some higher probabilities exist within the contours that do not exceed the following threshold. For example, the 15% contour may contain probabilities as high as 29.99%, since 30% is the next probabilistic contour

issued. Thus, SPC forecasts by design under-forecast according to the reliability diagram, resulting in values that are above the diagonal. Conversely, over-forecasting results in values beneath the diagonal.

Performance diagrams visualize four different statistical metrics including the Critical Success Index, defined as:

$$CSI = \frac{hits}{hits + misses + false\ alarms}. \quad (4.6)$$

This is typically a rare event score (Wilks 2011), and has verified prior tornado forecasts (Gallo et al. 2016; Sobash et al. 2016b). It ranges from 0.0 to 1.0, with 1.0 being a perfect score. Performance diagrams plot POD versus Success Ratio (SR), defined as:

$$Success\ Ratio = 1 - \frac{false\ alarms}{hits + false\ alarms}, \quad (4.7)$$

with lines of constant CSI and bias to aid in interpretation. The false alarms divided by hits plus false alarms is otherwise known as the false alarm ratio, or FAR. Reliability information at each threshold can also be extracted (i.e., ideally a SR of 15% would occur at the 15% forecast threshold).

4.3 Results

4.3.1 STP Percentile Sensitivity

The seasonally aggregated SPC 0600 UTC tornado forecasts had a ROC area of 0.824 for all tornadoes and 0.865 for RM tornadoes (Table 4.2), showing that the SPC is more skillful at forecasting RM tornadoes than tornadoes from other convective modes. However, both subsets easily exceed the 0.7 criteria determining a skillful forecast.

Similarly, the ensemble-based probabilities achieved skillful ROC areas for all tested percentiles, ranging from a low score of 0.845 for probabilities using the 10^{th} percentile of STP and verified on all tornadoes to a high score of 0.921 for probabilities using the maximum STP and verified on RM tornadoes (Table 4.2). Across all percentiles, verification on RM tornadoes scored higher than verification on all tornadoes, indicating that the forecasts were more adept at discerning areas of RM tornadoes. Given the underlying climatological frequencies and the strong correlation between UH and supercells, the probabilities were expected to particularly highlight areas where RM tornadoes may occur. Higher percentiles attained significantly higher ROC areas than the SPC, likely due to their broader coverage as a harsh penalty is imposed by the ROC area when missing a tornado report (Gallo et al. 2016).

ROC curves for all STP percentiles had higher POD and POFD than the SPC forecasts, particularly at lower forecast thresholds such as 2% (Fig. 4.4a,d). The curves also showed that the increase in ROC area at higher STP percentiles comes mostly from increased POD at the 2% and the 5% threshold. Above the 5% threshold, the POD and the POFD were nearly indistinguishable from the SPC's forecasts. Thus, STP-based ensemble forecasts could provide forecasters with objectively skillful first-guess tornado probabilities, particularly for RM tornadoes, with the understanding that at low thresholds the improvement in POD is accompanied by a slightly higher POFD. The largest difference between verifying with all tornadoes and RM tornadoes stemmed from the POD difference at low forecast thresholds, with all forecasts having a higher POD for RM tornadoes than for all tornadoes.

A day-to-day comparison between the ROC areas illustrated another operationally relevant facet of the probabilities (Table 4.2). For the lower percentiles of STP, the SPC probabilities had a higher ROC area than the NSSL-WRF ensemble probabilities and vice versa slightly over one-third of the time. Remaining cases had a tied ROC area of 0.5, which occurred when no tornadoes happened or when tornadoes occurred entirely in regions below 2% forecast probabilities. The percentage of days the NSSL-WRF ensemble ROC area exceeded the SPC ROC area was highly dependent on the percentile of STP used to generate the NSSL-WRF ensemble probabilities. The NSSL-WRF ensemble most often scored higher than the SPC when ensemble probabilities were generated using the maximum percentile. Conversely, the SPC most often scored higher than the NSSL-WRF ensemble when ensemble probabilities were generated using the 10^{th} percentile. These results were consistent between both verification datasets, suggesting that some days only non-RM tornadoes occurred within the 2%+ probability. The higher the percentile of STP used to generate the ensemble probabilities, the higher the percentage of days the ensemble scored higher than the SPC, likely because increased coverage of the probabilities missed fewer tornadoes.

Since the ROC area solely distinguishes events from non-events, forecast reliability is key in determining the practical usefulness of the probabilities. Reliability diagrams showed that the ensemble-based probabilities closely resembled the SPC forecasts when they were generated using the 10^{th} percentile of STP (Fig. 4.4b,e). Higher percentiles over-forecasted all tornadoes, especially at low probabilities (Fig. 4.4b); only the 10^{th} percentile forecast was nearly reliable until the 30% forecast probability. When forecasting RM tornadoes over-forecasting increased, and the 10^{th}

percentile remained most reliable (Fig. 4.4e). The increase in over-forecasting when looking at RM tornadoes compared to all tornadoes was expected, since the RM constraint ensures fewer tornadoes in the verification dataset.

Performance diagrams allow a closer examination of individual probabilistic forecast thresholds. Since tornadoes rarely occur, the ideal forecast would contain a majority of tornadoes with limited false alarm, leading to a SR equal to the probability at each probability threshold. At nearly all percentiles and probability thresholds, the ensemble forecasts had a higher POD and a lower SR than the SPC probabilities (Fig. 4.4c,f). An exception occurred with the probabilities generated using the 10^{th} percentile of STP for the 10% or 15% threshold, when the ensemble forecasts had higher PODs and higher SRs than the SPC forecasts. SPC forecasts of 10% and 15% are reserved for high-impact days, and so these thresholds warrant special attention.

Performance diagram results were consistent between all tornadoes (Fig. 4.4c) and RM tornadoes (Fig. 4.4f), but the RM tornadoes generally had a lower CSI despite having an increased ROC area. Since RM tornadoes are a subset of all tornadoes, when verifying solely on RM tornadoes the false alarm and correct negatives will increase, the misses will decrease, and at best the number of hits will remain the same (if the probabilities are encompassing all RM tornadoes) or decrease. In a rare event scenario, false alarms are often the largest term in the CSI (compared to hits and misses), and the increased false alarm of verifying on RM tornadoes decreases the CSI. False alarms affect CSI more than the ROC area because the CSI does not incorporate correct negatives. False alarms are incorporated in the ROC area through the POFD, which is

overwhelmingly dominated by correct negatives in the rare-event scenario. The ROC area is instead sensitive to the POD, and increases because of the decreased misses.

4.3.2 Probability Generation Method Comparison

The probabilities generated using the 10th percentile of STP were the most reliable while maintaining high skill, so those forecasts were compared with other methodologies of probability generation (Gallo et al. 2016; Sobash et al. 2016b). From this point, the STP-based probabilities denote the probabilities computed using the 10th percentile of STP. Seasonally aggregated ROC areas between the 0–3 km UH-only, 2–5 km UH-only, and STP-based probabilities were similar, while the filtered 2–5 km UH had a much lower ROC area. However, neither the filtered 2–5 km nor the STP-based method were statistically significantly different from the SPC forecasts for either verification dataset (Table 4.3). Across both verifications, ROC curves of the UH-only methods had higher POD and POFD at low probability thresholds (Fig. 4.5). The filtered 2–5 km UH method had lower POFDs than the other methods accompanied by a much lower POD than the other methods and the SPC forecasts. The STP-based probabilities had a slightly lower POD than the UH-only methods, but also had a lower POFD that more closely resembles the SPC forecasts. The most obvious difference between the RM tornado verification and the all tornado verification was that the RM tornadoes produced higher ROC areas than the all tornado dataset across all methods, mostly due to an increase in POD at low thresholds. Otherwise, the results were consistent between verifications.

The percentage of days each method achieved a higher ROC area than the 0600 UTC SPC probabilities varied greatly, from a low of 28.6% for the filtered 2–5 km UH

verified against RM tornadoes to a high of 50.0% for the 0–3 km probabilities verified against all tornadoes (Table 4.3). The STP-based probabilities more often outscored the SPC than the filtered 2–5 km UH, but less often than the two UH-only methods. Overall, most daily forecasts were skillful for both verifications, although some days had large spread between the methods showing that method choice had a large impact on forecast skill (Fig. 4.6). Marginal days, in which one or two tornadoes occurred on the edge of the forecast area, were often the most impacted by method choice. In those cases, increased coverage (occurring with more widespread UH and less environmental criteria) achieved a higher ROC area by covering more "tornado event" points. Additionally, STP-based probabilities typically scored higher than filtered 2–5 km UH probabilities, suggesting that incorporating STP probabilistically generated a better forecast than using STP as an additional binary criterion. The shift to higher scores for RM tornadoes (i.e., more points in the upper right corner of the graph) occurred due to an overall improvement in ROC areas for both the SPC and the NSSL-WRF ensemble.

Methods differed immensely in their reliability (Fig. 4.7). High SPC forecast probabilities are rare, and unnecessarily high first-guess ensemble probabilities can mislead forecasters trying to anticipate the severity of a day (Gallo et al. 2016). Vast over-forecasting occurred in the methods solely using UH despite their high ROC areas, and verification using only the RM tornado dataset exacerbated this signal. Filtering the 2–5 km UH probabilities by requiring STP ≥ 1 improved reliability, but still over-forecasted. The STP-based probabilities, however, were remarkably reliable, particularly when forecasting RM tornadoes. The SPC was also extremely reliable for both verification methods. Indeed, the SPC forecasts achieved nearly perfect reliability

up to 15% when forecasting RM tornadoes, while the STP-based probabilities over-forecasted at 10% and below. Clearly, using empirical observations as a basis for the probabilistic tornado forecasts improved reliability over the other methods, which solely rely on an ensemble and Gaussian smoother to moderate the probabilities.

A performance diagram illustrates verification statistics at SPC forecast thresholds (Fig. 4.8). At the 2% level the UH-only and STP-based methods have similar SRs, although the STP-based method had higher CSI and lower POD than the UH-only methods. However, beginning at the 5% level, all methods except the STP-based probabilities have much higher POD and lower SR than the SPC forecasts. At the 10% and 15% threshold, the STP-based probabilities have higher CSI, POD, and SR than the SPC forecasts for all tornadoes and for RM tornadoes, although the increase in SR was larger for all tornadoes than for RM tornadoes. As the probability threshold increases, so do the discrepancies between the methods, with the UH-based methods having much higher POD and much lower SRs than the SPC and the STP-based method and corresponding to their high bias.

4.3.3 Case Studies

To demonstrate how the probabilities appear to a forecaster, three case studies are now presented. The first illustrates a high-impact day, with high probabilities and multiple tornadoes. The second highlights an area where forecast upscale growth contained embedded supercells, emphasizing that these probabilities are intended as a tool for forecasting supercellular tornadoes. The final case occurred on a more marginal day, and had a relatively large false alarm area.

a. 28 APRIL 2014

Late April 2014 saw a multi-day outbreak spanning from the Great Plains to the east coast, with the most tornadoes occurring on 28 April. In fact, this day had the largest number of tornadoes (121) of any day in our dataset. Four of these tornadoes caused fifteen deaths across Mississippi, Alabama, and Tennessee. On the 28th, a 500 mb closed low was located over Nebraska and a negatively tilted shortwave trough stretched from the central Great Plains into eastern Oklahoma and Louisiana. At the base of this trough, a 500 hPa jet streak with wind speeds exceeding 80 kt existed over Arkansas and moved eastward throughout the day. Thermodynamic parameters were also favorable, with MLCAPE exceeding 2000 Jkg^{-1} where tornadoes would later occur. Objectively analyzed STP ranged from 3.0–6.0 in the area of interest (not shown).

The SPC forecasted this event well in advance, issuing a Day 3 moderate risk. The SPC's 0600 UTC tornado probabilities (Fig. 4.9a) had a broad area of 15% probability, corresponding to a "moderate" categorical risk. The 2000 UTC update to this forecast increased the tornado probabilities to 30% (not shown), leading to a categorical upgrade to high risk. The 0600 UTC SPC-issued probabilities successfully captured the largely RM tornado reports for that day, and most of the tornadoes occurred in the upper-tier probabilities. The NSSL-WRF ensemble also highlighted the Southeast, with high ensemble STP and abundant UH, creating high probabilities for all methods (Fig. 4.9b-e).

This case demonstrates the value of restricting the maximum probability using observed frequencies. Initially, using midlevel rotation (Fig. 4.9b) or low-level rotation (Fig. 4.9d) alone created extremely high probabilities both within and well outside the region with numerous tornadoes. The over-forecasting of the 2–5 km UH probabilities

(Fig. 4.9b) was not tempered much by requiring STP ≥ 1 (Fig. 4.9c), since high STP was abundant. However, the STP-based probabilities (Fig. 4.9e) had a maximum magnitude equivalent to the SPC's updated forecast: 30%, which is categorically equivalent to a high risk, although they had lower probabilities than the other methods within the region containing numerous tornadoes. All forecasts on this day had ROC areas above 0.95 (Fig. 4.9f).

b. 3 JUNE 2014

The second case contained mixed modes, where clusters of supercells produced most of the tornadoes. A vigorous short-wave trough was initially located across the north-central plains, with strong 250 hPa wind speeds (not shown). According to the 0600 UTC convective outlook, severe convection was expected to occur near a warm front. The forecast environment had ample shear and sufficient CAPE to support rotating storms. Isolated, high-based storms were anticipated initially, but much NWP guidance showed fast upscale growth into one or more mesoscale convective systems (MCSs). As a result, a 10% tornado threat was highlighted by the 0600 UTC SPC convective outlook (Fig. 4.10a), along with a 45% damaging wind threat (not shown). Although upscale growth occurred, many of the storms retained supercellular characteristics early in their convective life cycle. Six RM tornadoes and one non-supercellular tornado resulted.

As in the previous case, the 2–5 km UH (Fig. 4.10b) and the 0–3 km UH (Fig. 4.10d) had vast swaths of probability exceeding 60% (the highest possible tornado probability contour issued by the SPC), including in areas outside of the region with several tornadoes. However, the probabilities captured the tornado in western Kansas,

which was missed by the 0600 UTC outlook (the 1630 UTC outlook extended the 2% probabilities into western Kansas). Capturing that tornado report increased ROC area for those probabilities at the significant expense of increasing the forecast probabilities across the region (Fig. 4.10f). Forecasters might have excessive difficulty determining the appropriate magnitude of the probabilities given this over-forecasting, as was seen in Gallo et al. (2016). Incorporating environmental information by requiring an exceedance of STP reduced the probabilities somewhat (Fig. 4.10c), but the peak magnitude remained above 60% and the Kansas tornado was now outside the 2% contour, decreasing the ROC areas. The STP-based probabilities (Fig. 4.10e), however, handled the magnitude of the event best of any automated probabilities, although the highest probabilities occurred east of the area with the most tornadoes. The highest probability contour was only one category higher than the official SPC forecast on this day, making them the most useful first-guess of any ensemble probabilities as the forecaster would not have to mentally calibrate the probabilities to typical operational values. Verifying solely on RM supercells doesn't have much of an effect on this case, although a slight decrease in the SPC's ROC area was caused by no longer counting the non-supercellular tornado in southwestern Illinois. This case also demonstrates the struggle the probabilities have with mode, in that UH swaths associated with MCSs can produce areas of false alarm, as seen across Illinois in all ensemble-generated methods.

c. 5 MAY 2015

The third case examined herein demonstrates how these probabilities are best used for forecasting RM tornadoes, and shows the difficulties they may have on more weakly forced days. According to the SPC 0600 UTC convective outlook, a shortwave

trough was forecast to evolve across the CONUS throughout the period of interest. Ongoing thunderstorms were expected to limit the instability across the central High Plains. A sharpening dryline and remnant boundaries from the morning convection were anticipated as the focus of the subsequent severe convection. Such mesoscale detail poses a forecasting challenge to humans and NWP alike, making this a difficult day to forecast. Effective bulk shear was noted by the SPC as sufficient for supercells with a tornado threat east of the dryline, leading to an area of 5% tornado probability across the Texas Panhandle and a broader area of 2% stretching southward, where shear was weaker (Fig. 4.11a). Subsequent outlooks reduced the area of 5% and eventually shifted it southward (not shown).

While the UH-only methods had lower probabilities than in the prior two cases, they still showed areas of 10% (2–5 km UH-only; Fig. 4.11b) and 15% (0–3 km UH-only; Fig. 4.11d), which are typically used by the SPC on high-end days. These probabilities encompassed all of the tornadoes that occurred on 5 May, with the exception of the non-RM tornado in Oklahoma. Filtering the UH by requiring STP ≥ 1 decreased the area of false alarm in Oklahoma, but just excluded the tornadoes that occurred in central Texas and maintained the high-magnitude false alarm in southern Texas (Fig. 4.11c). Using the STP-based probabilities decreased the false alarm overall, and the maximum probability magnitude matched that of the SPC: 5%. Probabilities across southern Texas were especially reduced. However, the area highlighted by the 5% was in southwestern Oklahoma, which had no tornadoes, and some of the southern tornadoes were excluded.

This case shows how different daily statistics can be when verifying RM tornadoes vs. all tornadoes. The SPC's ROC area increased greatly, from 0.84 to 0.96 (Fig. 4.11f), as the only tornado not in the SPC's forecast area was non-RM. Such increases emphasize the importance of capturing all of the reports to ROC areas in a rare-event scenario such as tornado forecasting, which is also demonstrated by comparing the increase in ROC area among the forecast methods. All methods showed some increase in ROC area when verifying on RM tornadoes as compared to all tornadoes, but the increase for the 0–3 km UH-only probabilities was much greater than the increase in the filtered 2–5 km filtered UH probabilities (Fig. 4.11f). The only tornadoes not captured by the 0–3 km UH-only probabilities were non-RM, so excluding them from verification greatly increased the ROC area despite substantially over-forecasting. However, the exclusion of the non-RM tornadoes in the filtered 2–5 km UH probabilities led to fewer misses and more correct negatives, which would respectively increase the POD and decrease the POFD. Nevertheless, since the forecast still excludes most of the RM tornadoes (some of the tornadoes likely occurred within 40 km of the edge of the 2% probabilities), the ROC area did not increase by much.

4.4 Summary and Discussion

Forecast probabilities generated using combined ensemble output and observed climatological tornado frequencies performed comparably to the SPC 0600 UTC forecasts for all tornadoes and solely RM tornadoes. These model forecasts are designed for quick forecaster interpretation by summarizing relevant environmental and convective ensemble parameters into one graphic. Additionally, the ensemble forecasts

currently become available for the 1300 UTC forecast updates, allowing forecasters to adjust the magnitude and location of the 0600 UTC tornado probabilities if they think the ensemble forecast probabilities add value. Incorporating this method into other ensembles would even allow the probabilities to be available in time for the initial Day 1 forecast at 0600 UTC, and is the subject of ongoing work.

These probabilities are the first to incorporate observed climatological frequencies given environmental parameters, unlike other ensemble-based tornado forecast techniques to date. The climatological frequencies calibrate the tornado probability given model-based storm environments and attributes, improving upon the idea of using thresholds of simulated environmental values, as is seen in Gallo et al. (2016). Calibrating on the STP magnitude presumes that tornado occurrence in a high-STP environment when a supercell is present is more probable, all else being equal. By calculating the probability using the value of environmental STP, the probabilities provide more information than a simple threshold exceedance paradigm. To construct the probabilities and ensure that the environmental STP remains free of storm influences, each point and time has a unique STP distribution. The probabilities are calculated by taking different percentiles of this distribution, finding the maximum resultant STP throughout the day, and assigning the probability based on the climatological frequency to that point and ensemble member. Once all ensemble members have a probability field, a Gaussian-smoothed member average yields the final values.

Of the different percentiles of STP used for probability generation, the 10^{th} percentile had the highest reliability while maintaining high ROC areas and was

compared to other probabilistic forecast generation methods. The methods tested herein produced vastly different statistics. Using solely 2–5 km UH or 0–3 km UH as proxies for tornado occurrence produced large ROC areas as seen in previous studies (Jirak et al. 2014; Gallo et al. 2016; Sobash et al. 2016b), capturing many tornado events but over-forecasting. While the exact probability calculation method using the 0–3 km UH differed from Sobash et al. (2016b), using a UH threshold that produced the most reliable forecasts also misses many tornado events as evidenced by the relatively low ROC areas in Sobash et al. (2016b). Since these probabilities are to be operational forecasting tools, the 0–3 km UH threshold selected herein minimized missed events at the expense of perfect reliability.

Statistically, the STP-based probabilities resembled the 0600 UTC tornado forecasts issued by the SPC more than any other method, when verified by all tornadoes or solely by RM tornadoes. While the UH-only methods captured more tornado events than the STP-based probabilities (i.e., higher ROC areas), both low-level and midlevel UH over-forecasted threat areas and magnitude. Incorporating environmental information by requiring STP ≥ 1 increased reliability compared to solely using UH, but excluded some tornadoes, lowering the ROC area and still over-forecasting. The STP-based probabilities scored high ROC areas by increasing the POD with a slight increase in the POFD at the low forecast thresholds that compose most of the SPC's forecasts. They also drastically reduced over-forecasting, with relatively reliable forecasts at most probabilistic thresholds, especially when considering all tornadoes. Until NWP models can directly resolve tornado-like vortices with finer grid-spacing, environmental information still adds value to tornado forecasts at ~3–4 km grid spacing.

On a day-to-day basis, the STP-based probabilities often performed comparably to the SPC forecasts, while the opposite was true for probabilities determined using a threshold of STP. The STP-based probabilities achieved these higher ROC areas while issuing lower probabilities, as shown in the case studies. Since these forecasts are designed to be available and can be considered a first-guess for operational forecasters (with caveats of the ensemble correctly forecasting the convective mode and environment), magnitudes that are more accurate save forecasters from trying to mentally calibrate unrealistically high probabilities. For example, forecasters on 3 June 2014 could have seen the potential for supercellular tornadoes, despite the forecasted upscale growth into linear convective modes. With this guidance, it may have been easier to determine that embedded supercells were a threat within the large storm clusters, although the UH generated by the linear MCSs would lend caution to the veracity of the underlying tornado probabilities. Indeed, only one non-RM tornado occurred after the line grew upscale.

The case studies also demonstrate limitations of using environmental parameter thresholds. On 28 April 2014, STP was abundant throughout the domain of concern, so limiting the probabilities by requiring that STP exceed one still created widespread high probabilities. On 3 June 2014, high STP occurred even after the storms grew upscale, leading to high probabilities east of where most tornadoes occurred. However, using the STP-based method, the probabilities were lowered and somewhat constrained. This method also decreased the magnitudes of the probabilities in less severe cases such as 5 May 2015 and focused the probabilities on the RM tornadoes, although weakly forced cases remain challenging.

The probabilistic paradigm discussed herein generates a probabilistic forecast from each ensemble member before averaging those forecasts. Therefore, this methodology is applicable to deterministic forecasts and ensembles of multiple sizes and implementation in such ensembles is the subject of future work. Future work will also extend these forecasts to differing modes and tornado intensities, perhaps developing similar probabilities for tornadoes with quasi-linear convective systems or forecasting the probability of a significant tornado. Further work also remains in isolating mode: a great improvement to these probabilities would eliminate the false alarm produced by UH from MCSs, which are far less likely to produce significant tornadoes than supercellular modes. Additionally, the data examined herein covered only spring seasons; in order for these probabilities to be increasingly validated by forecasters, applicability across seasons must be tested. While these probabilities are running daily online (at www.nssl.noaa.gov/wrf/newsite) and anecdotally appear to be useful outside of the peak convective season, formal operational evaluation has yet to occur.

Acknowledgements

The authors would like to thank Chris Melick and Robert Hepper of the SPC for providing regridded SPC forecasts, as well as Andrew Dean of the SPC for obtaining the environmental and radar data used in the climatological frequency calculation. Thanks also go to Ryan Lagerquist for insight to the cross-validation technique performed herein. This material is based upon work supported by the National Science Foundation Graduate Research Fellowship under Grant No. DGE-1102691, Project #A00-4125. BTG and SRD were provided support by NOAA/Office of Oceanic and

Atmospheric Research under NOAA-University of Oklahoma Cooperative Agreement #NA11OAR4320072, US Department of Commerce. AJC also received support from a Presidential Early Career Award for Scientists and Engineers. We would also like to thank three anonymous reviewers for their comments, which improved the content and clarity of the manuscript.

Tables

Table 4.4.1 Specifications for the NSSL-WRF ensemble. All members use WSM6 microphysics, Dudhia shortwave radiation, RRTM longwave radiation, the Noah land surface model, and the MYJ boundary layer. Members with years in parentheses by the ensemble member were only part of the ensemble for that year. Aside from the control NSSL-WRF member and _GFS member, members are initialized using 3 h SREF member forecasts initialized at 2100Z for the initial conditions and lateral boundary conditions.

Ensemble Member	ICs/LBCs	Microphysics	PBL	Radiation	Land-surface
1	00Z NAM	WSM6	MYJ	RRTM/Dudhia	Noah
2	00Z GFS	WSM6	MYJ	RRTM/Dudhia	Noah
3	21Z em_ctl	WSM6	MYJ	RRTM/Dudhia	Noah
4	21Z nmb_ctl	WSM6	MYJ	RRTM/Dudhia	Noah
5	21Z nmb_p1	WSM6	MYJ	RRTM/Dudhia	Noah
6	21Z nmm_ctl	WSM6	MYJ	RRTM/Dudhia	Noah
7	21Z nmm_n1	WSM6	MYJ	RRTM/Dudhia	Noah
8	21Z nmm_p1	WSM6	MYJ	RRTM/Dudhia	Noah
9 (2015)	21Z nmb_n1	WSM6	MYJ	RRTM/Dudhia	Noah
10 (2015)	21Z nmb_p2	WSM6	MYJ	RRTM/Dudhia	Noah
11 (2014)	21Z em_n1	WSM6	MYJ	RRTM/Dudhia	Noah
12 (2014)	21Z em_p1	WSM6	MYJ	RRTM/Dudhia	Noah

Table 4.4.2 Area under the ROC curve statistics for ensemble-generated forecasts based on differing percentiles of STP. Bolded seasonally aggregated areas under the ROC curve are statistically significantly different from the SPC area under the ROC curve at α=.05. Numbers outside parentheses were verified using all tornadoes; within the parentheses used solely RM tornadoes. Percentages in the rightmost two columns may not add to 100 due to ties in ROC area, which occurred when both the SPC and the NSSL-WRF scored ROC areas of 0.5.

STP Percentile	Seasonally Aggregated ROC area	Percentage of Days NSSL-WRF ROC area > SPC ROC area	Percentage of Days SPC ROC area > NSSL-WRF ROC area
10th	0.845 (0.879)	33.5 (29.1)	36.3 (32.4)
25th	0.855 (0.889)	34.6 (32.4)	35.2 (29.1)
Median	**0.868** (**0.902**)	35.2 (33.0)	34.6 (28.6)
75th	**0.878** (**0.911**)	40.7 (37.4)	30.2 (25.3)
90th	**0.884** (**0.916**)	43.4 (39.6)	27.5 (23.1)
Maximum	**0.890** (**0.921**)	48.4 (40.7)	23.1 (22.0)
SPC	0.824 (0.865)	---	---

Table 4.4.3 Area under the ROC curve statistics for different methods of generating ensemble-based probabilities. Bolded seasonally aggregated areas under the ROC curve are statistically significantly different from the SPC area under the ROC curve at α=.05. Numbers outside parentheses were verified using all tornadoes; within the parentheses used solely RM tornadoes. Percentages in the rightmost two columns may not add to 100 due to ties in ROC area, which occurred when both the SPC and the NSSL-WRF scored ROC areas of 0.5.

Method	Seasonally Aggregated ROC area	Percentage of Days NSSL-WRF ROC area > SPC ROC area	Percentage of Days SPC ROC area > NSSL-WRF
2–5 km UH, Unfiltered	**0.867** (**0.900**)	39.6 (43.4)	23.6 (26.9)
0–3 km UH, Unfiltered	**0.889** (**0.919**)	50.0 (43.4)	22.5 (22.0)
2–5 km UH, Filtered by STP ≥ 1	0.810 (0.848)	29.7 (28.6)	37.4 (30.8)
STP-based, 10^{th} Percentile	0.845 (0.879)	33.5 (29.1)	36.3 (32.4)
SPC	0.824 (0.865)	---	---

Figures

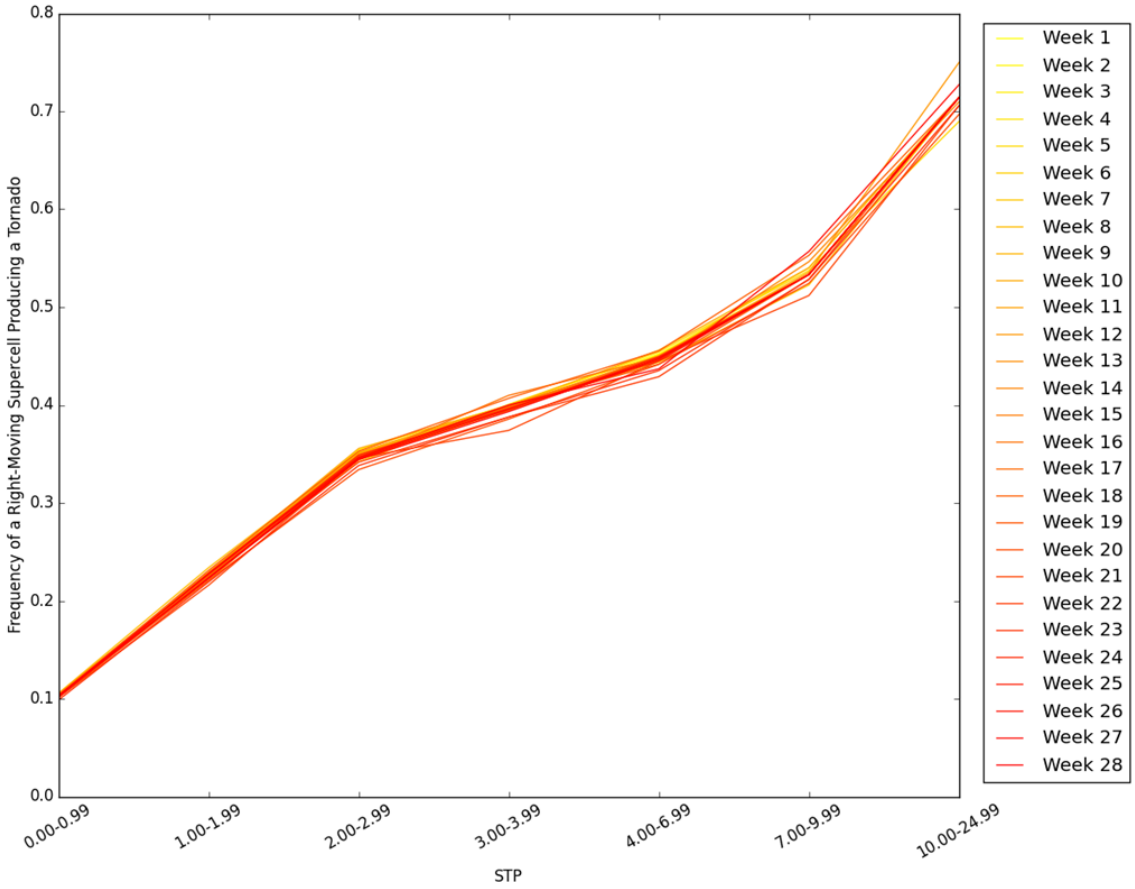

Figure 4.1 The climatological frequency of tornadoes given a right-moving supercellular storm associated with a LSR and a given modified fixed-layer STP using all data from 1 February 2014–31 December 2015 except the week indicated in the legend. Week 1 begins on 30 March 2014, week 14 begins on 29 June 2014, week 15 begins on 29 March 2015, and week 28 begins on 28 June 2015.

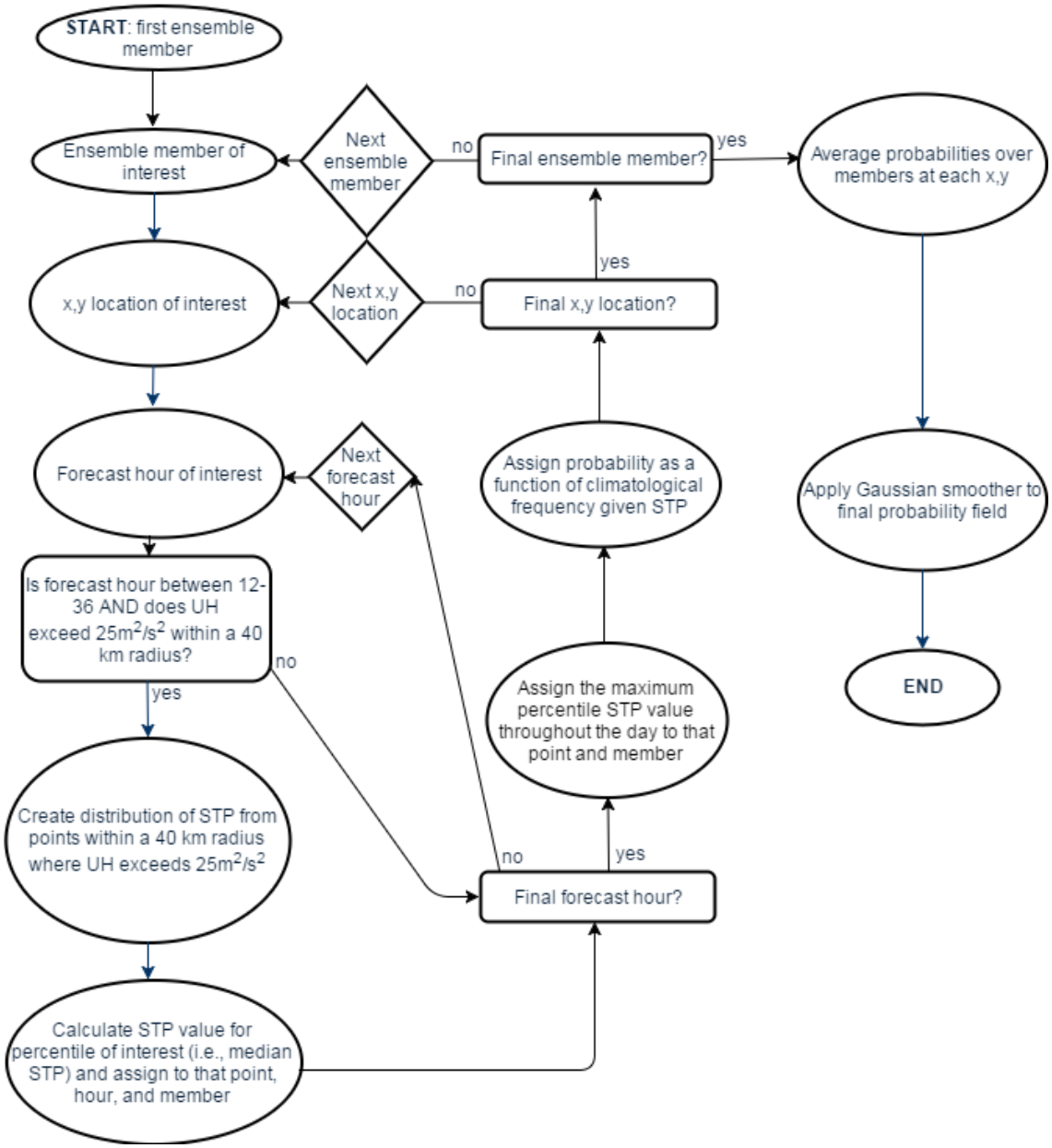

Figure 4.2 A schematic outlining the process of the probabilistic forecast generation. Rectangular boxes indicate decision points.

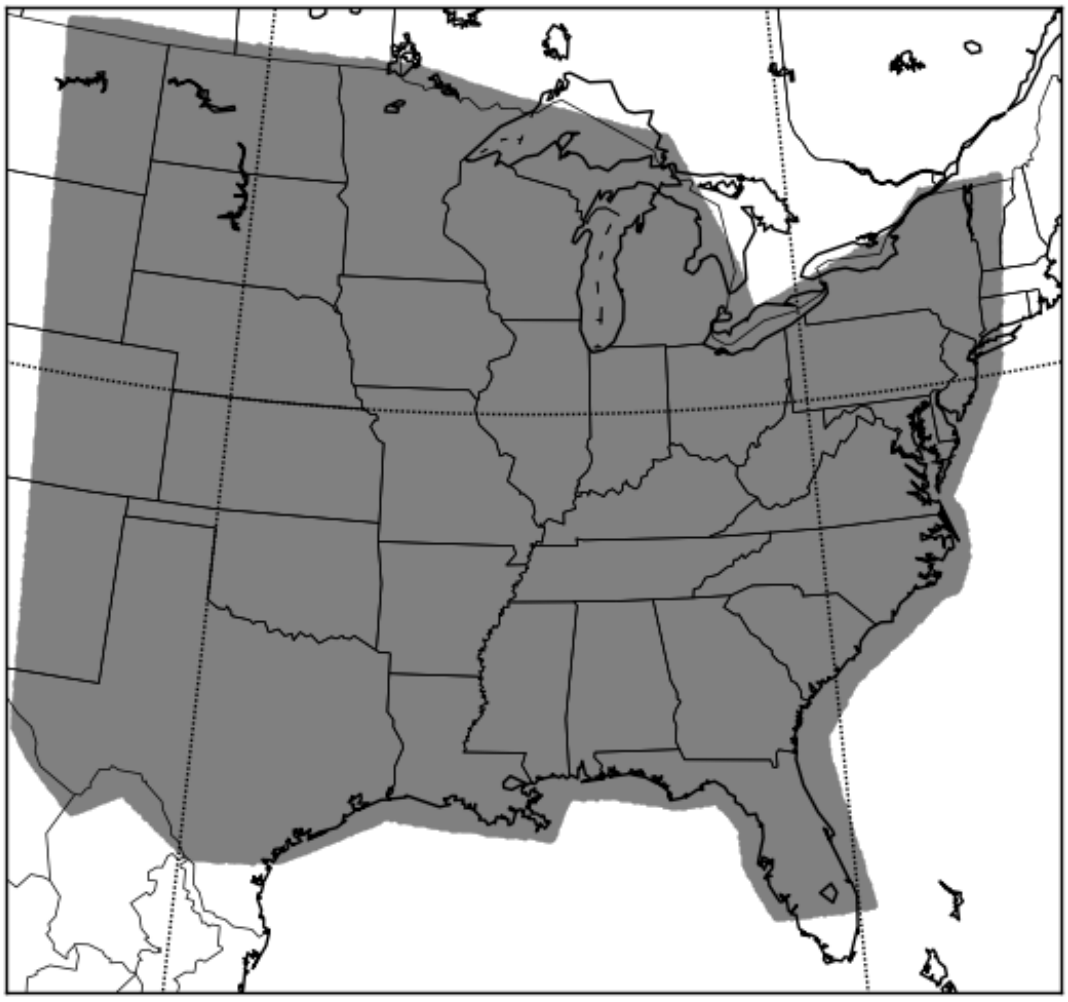

Figure 4.3 A subset of the model domain for the NSSL-WRF ensemble showing where objective verification measures were computed (shaded region).

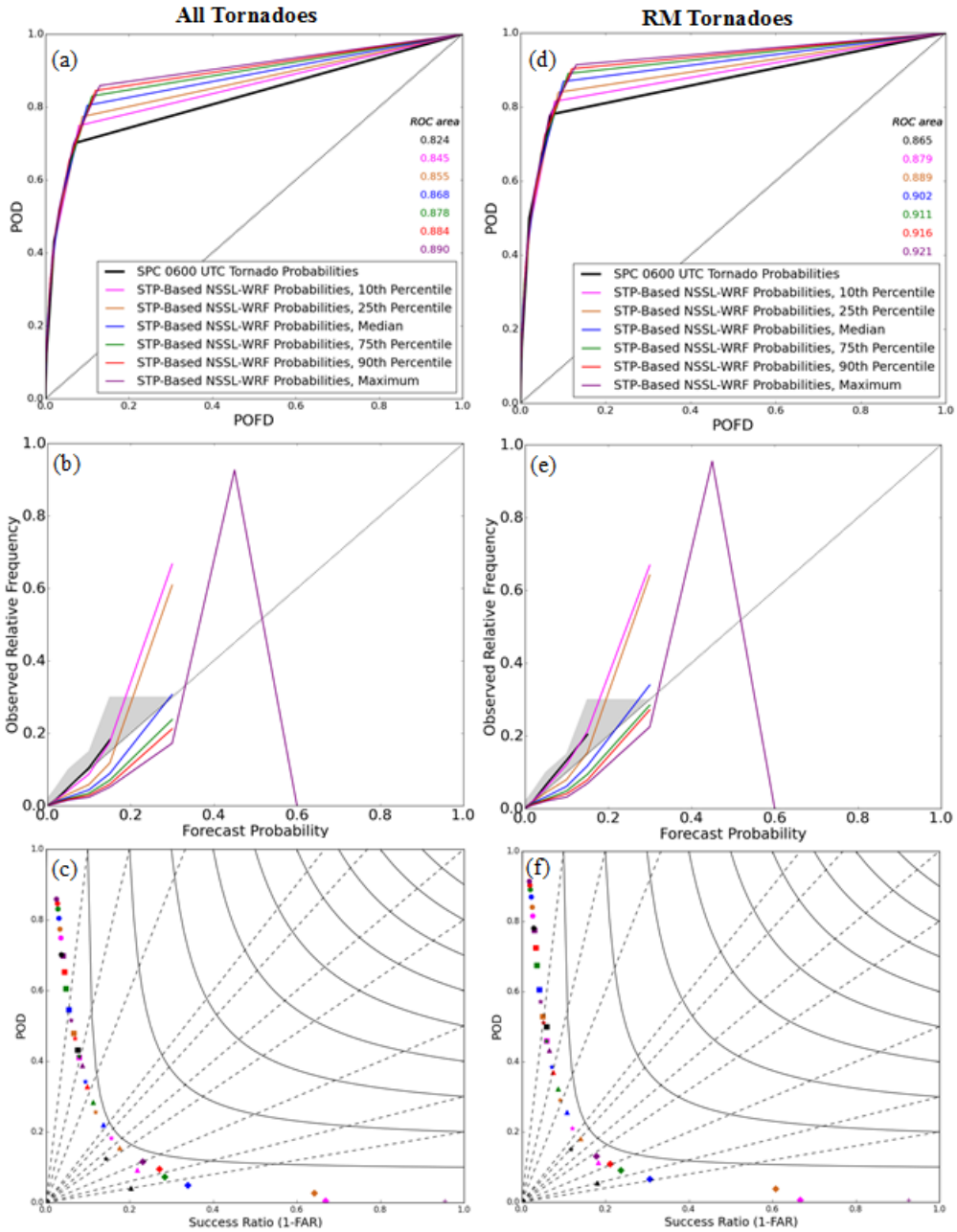

Figure 4.4 Summary statistics for different percentiles of STP used to calculated the STP-based NSSL-WRF ensemble probabilities: seasonally aggregated ROC curves for (a) all tornadoes and (d) RM tornadoes annotated with the areas under the ROC curve, reliability diagrams for (b) all tornadoes and (e) RM tornadoes, and performance diagrams for (c) all tornadoes and (f) RM tornadoes. Colors represent percentiles of STP used in probability generation. Black lines and symbols represent the SPC 0600 UTC forecasts. In (a) and (d), the thin black line indicates the performance of a random forecast, while in (b) and (e), it represents perfect reliability. In (c) and (f), the different symbols represent the different probability thresholds: Circles, squares, stars, triangles, and diamonds represent 2%, 5%, 10%, 15%, and 30%, respectively. Black dashed lines are lines of constant bias, while solid black lines are lines of constant CSI.

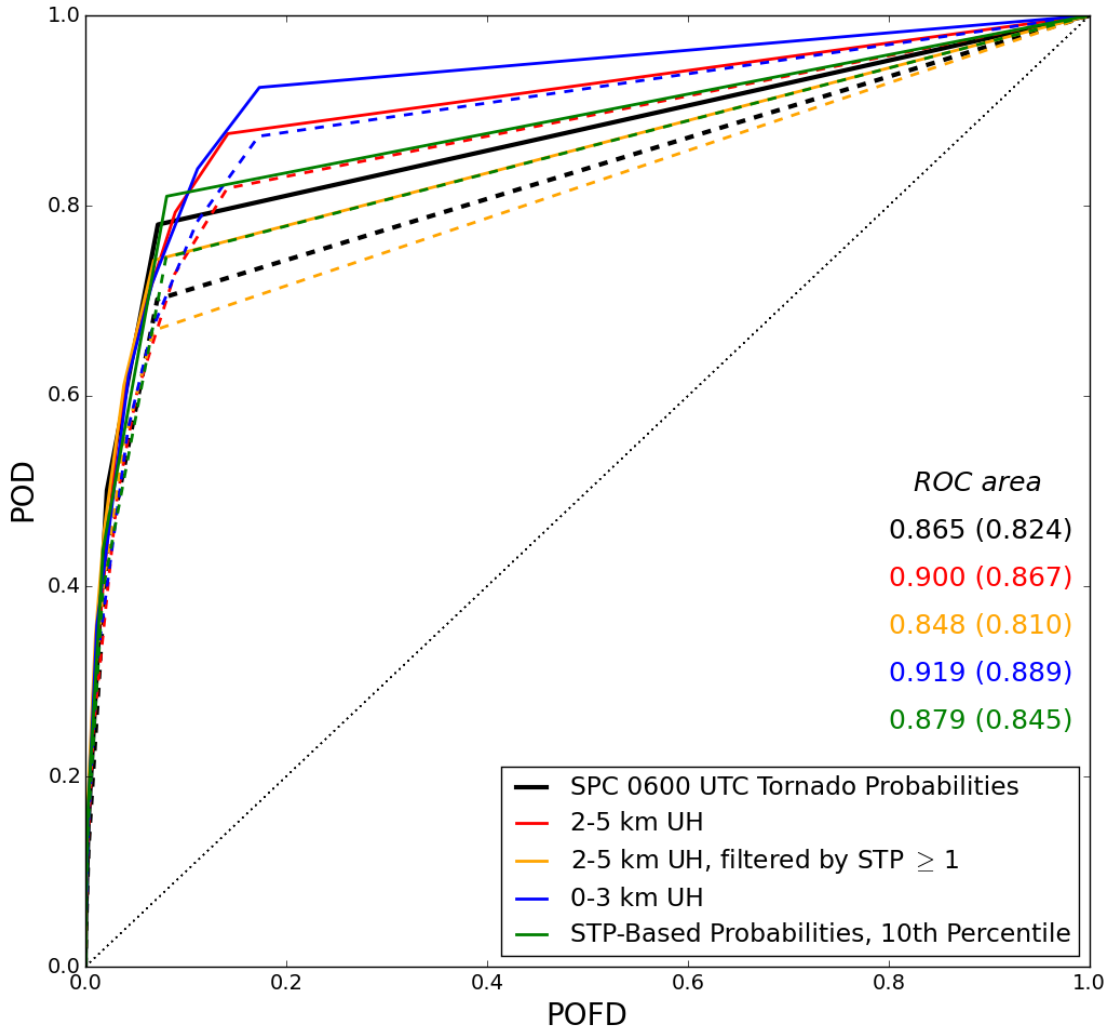

Figure 4.5 ROC curves for different probabilistic tornado forecasting methods, annotated with the area under the ROC curve for RM tornadoes (all tornadoes). Different colors represent the different methods. Solid lines are verified using only RM tornadoes, while dashed lines are verified using all tornadoes. The dotted black line indicates the ROC area of a random forecast.

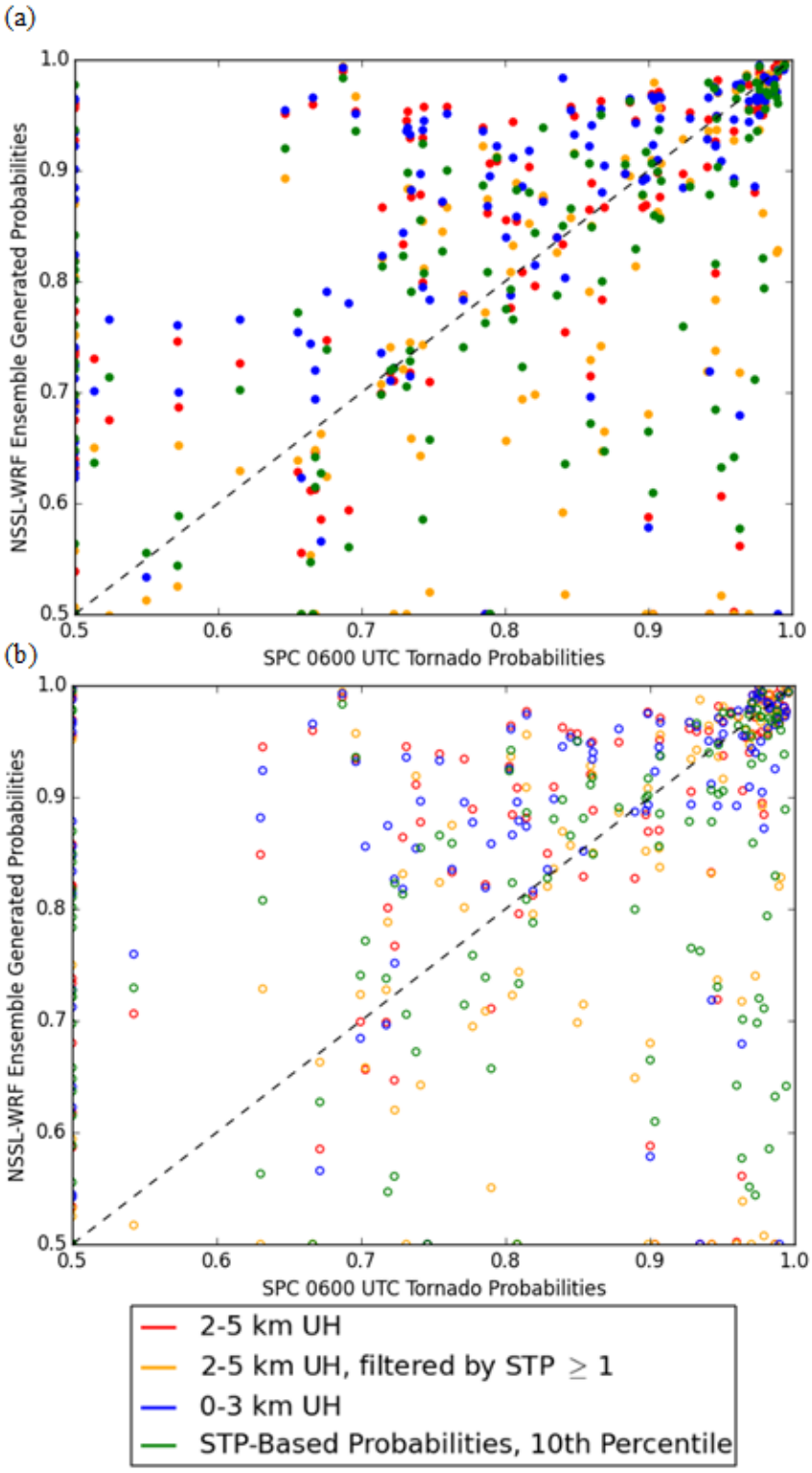

Figure 4.6 Daily ROC areas for the 0600 UTC tornado probabilities and NSSL-WRF ensemble-generated tornado forecasts using various methods of probability composition for (a) all tornadoes and (b) RM tornadoes. Each color represents a different method. The dashed line indicates equivalent scores for the SPC and the NSSL-WRF ensemble.

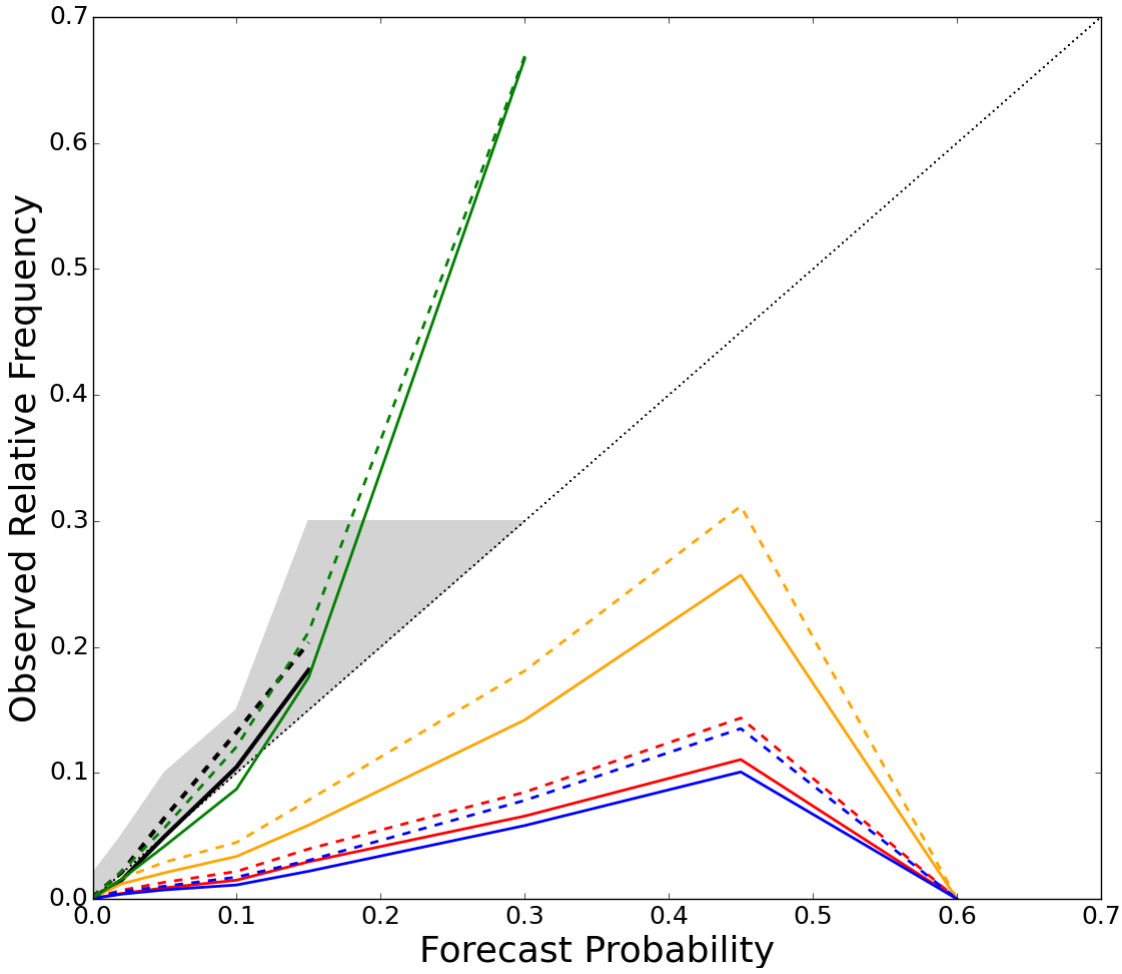

Figure 4.7 Reliability diagrams for different probabilistic tornado forecast methods. Different colors represent the different methods. Dashed lines are verified on all tornadoes and solid lines are verified solely on RM tornadoes. The dotted black line indicates perfect reliability. The shaded region represents where categorical forecasts currently issued by the SPC are reliable (e.g., the 2% forecast encompasses areas from 2%-4.99%).

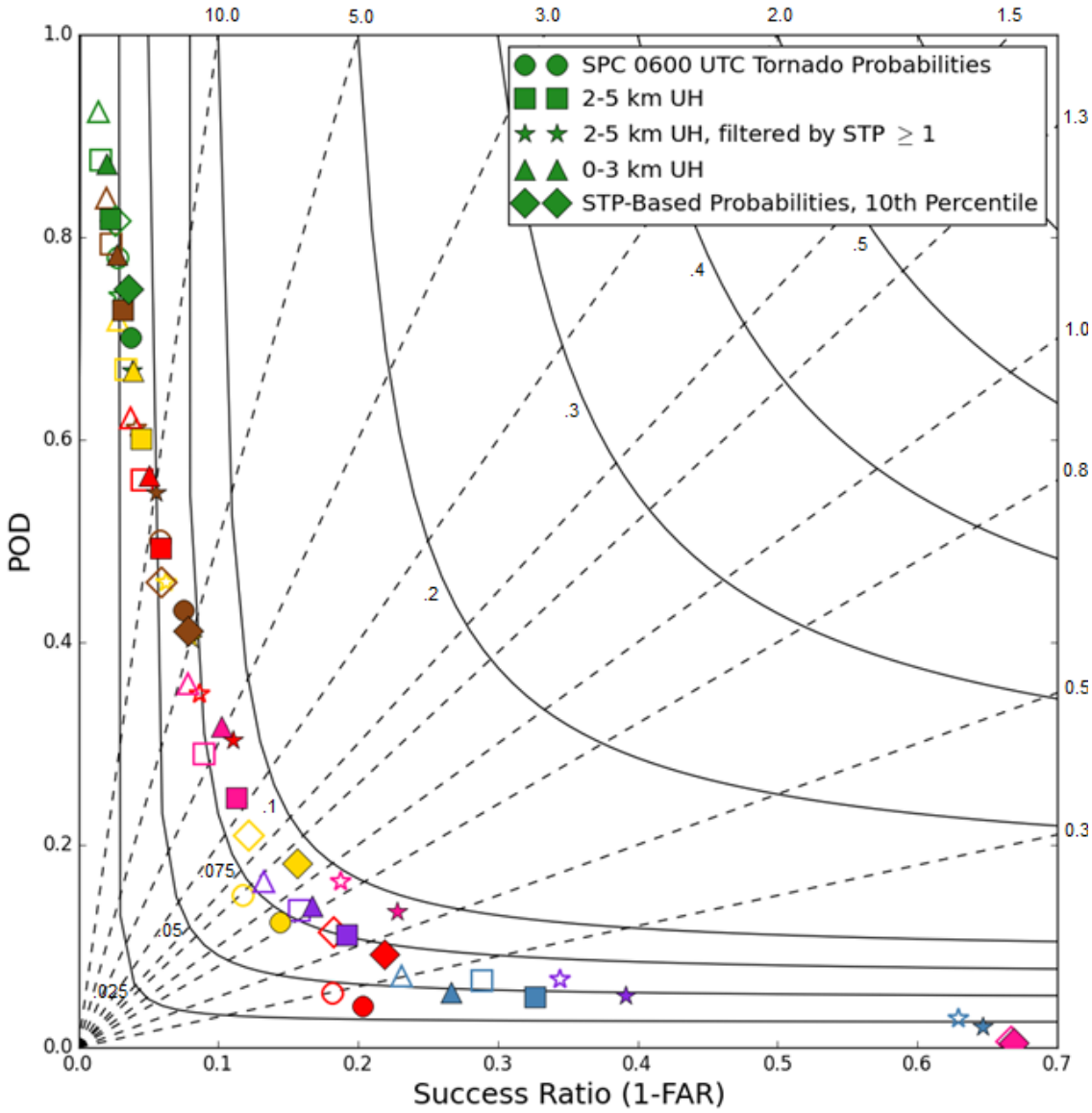

Figure 4.8 Performance diagrams for the forecast tornado probabilities. Different colors indicate different probability thresholds. Green, brown, yellow, red, pink, purple, and blue represent 2%, 5%, 10%, 15%, 30%, 45%, and 60%, respectively. Filled shapes are verified on all tornadoes; hollow shapes are verified on RM tornadoes. Black dashed lines are lines of constant bias, while solid black lines are lines of constant CSI.

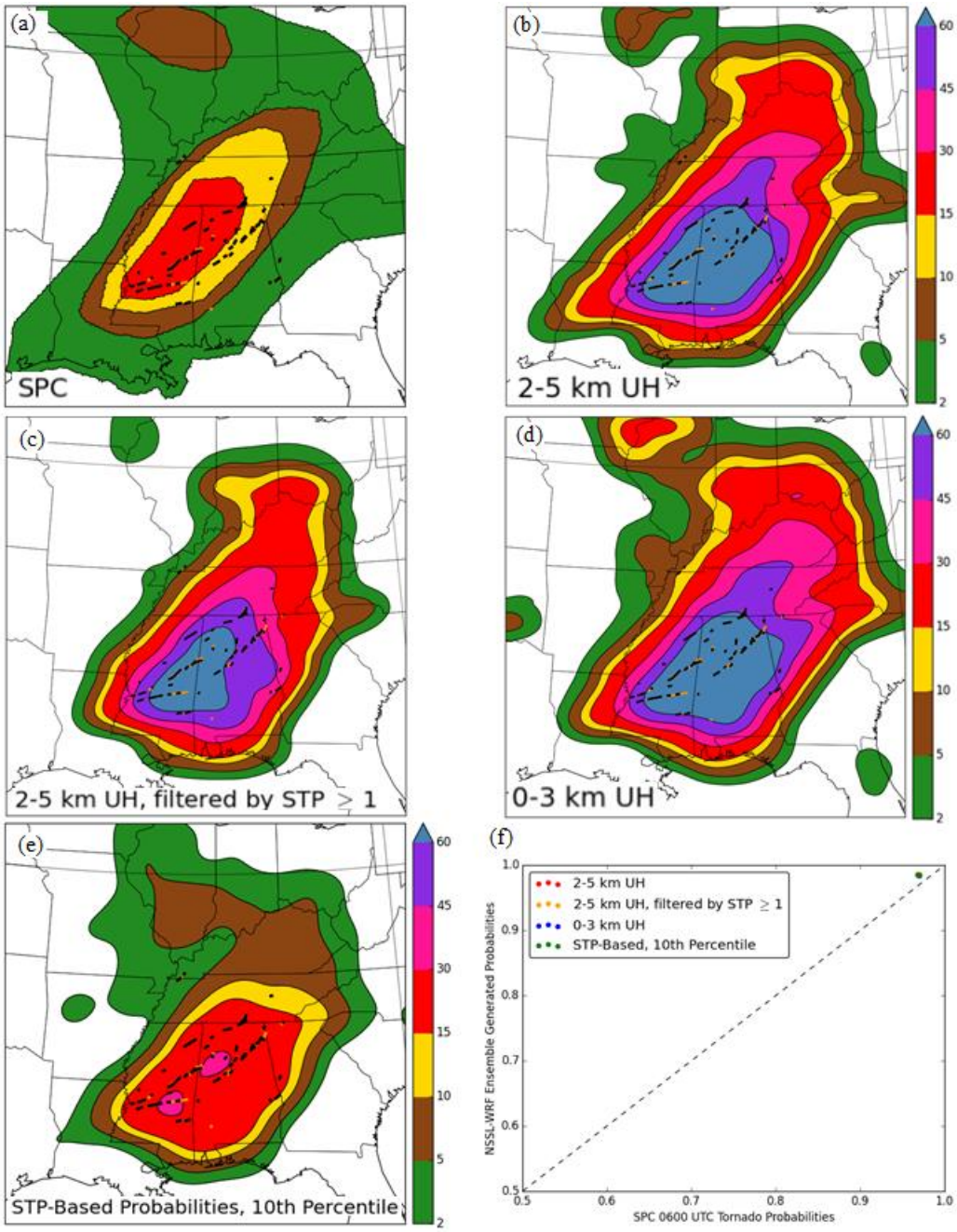

Figure 4.9 Forecast tornado probabilities for 28 April 2014 (a) issued at 0600 UTC by the SPC and generated with the NSSL-WRF ensemble, using (b) 2–5 km UH ≥ 75 m^2s^{-2}, (c) 2–5 km UH ≥ 75 m^2s^{-2} moving into an environment with STP ≥ 1, (d) 0–3 km UH ≥ 33 m^2s^{-2}, and (e) the 10^{th} percentile of STP from the hour previous to 2–5 km UH ≥ 25 m^2s^{-2}. All (orange) and RM (black) tornado paths are overlaid. (f) Daily ROC areas for the SPC and NSSL-WRF ensemble probabilities using the median STP on 28 April 2014. Different colors represent different methods. The dashed line indicates equivalent scores for the SPC and the NSSL-WRF ensemble. Filled circles are verified on all tornadoes and hollow circles are only verified on RM tornadoes.

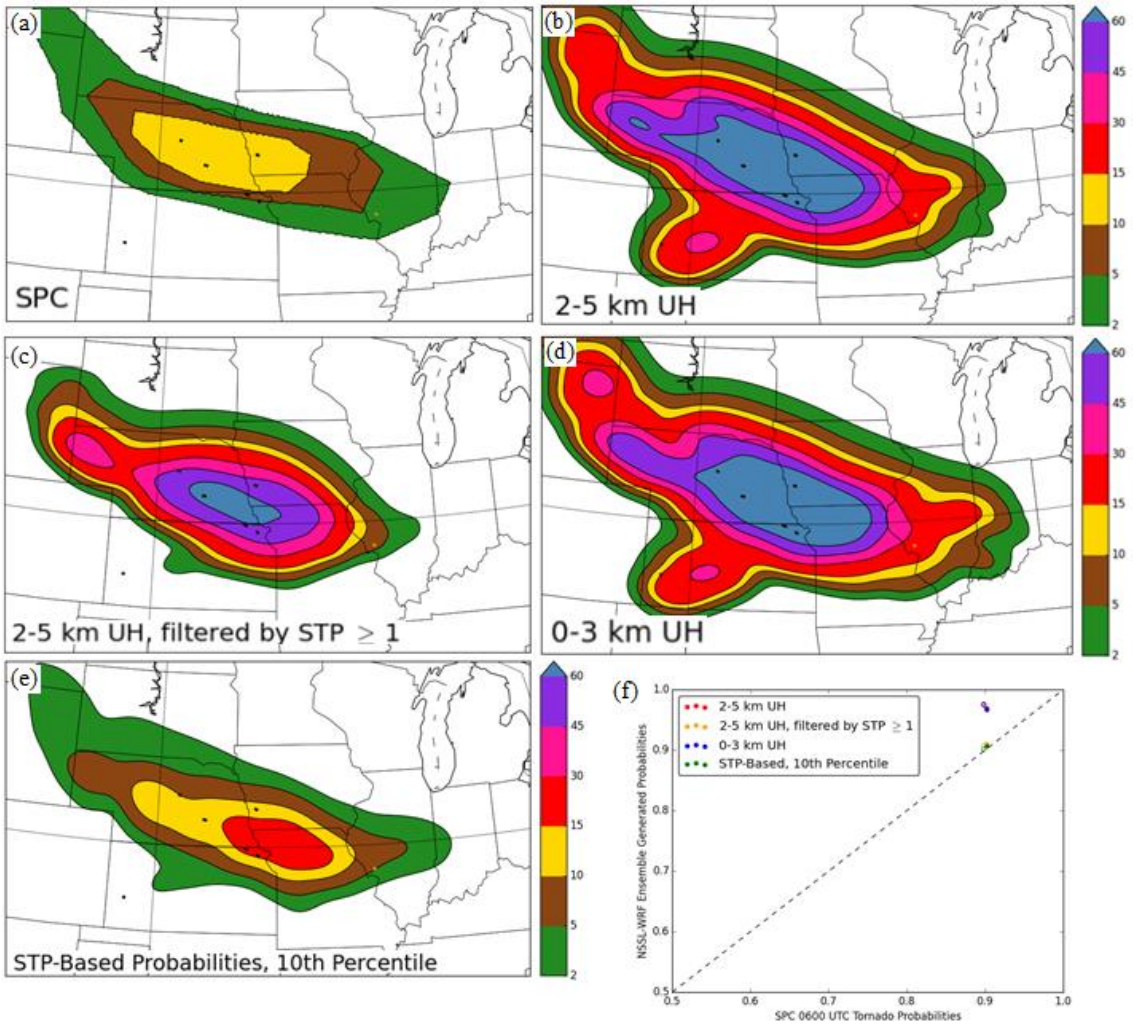

Figure 4.10 Forecast tornado probabilities for 3 June 2014 (a) issued at 0600 UTC by the SPC and generated with the NSSL-WRF ensemble, using (b) 2–5 km UH ≥ 75 m^2s^{-2}, (c) 2–5 km UH ≥ 75 m^2s^{-2} moving into an environment with STP ≥ 1, (d) 0–3 km UH ≥ 33 m^2s^{-2}, and (e) the 10^{th} percentile of STP from the hour previous to 2–5 km UH ≥ 25 m^2s^{-2}. All (orange) and RM (black) tornado paths are overlaid. (f) Daily ROC areas for the SPC and NSSL-WRF ensemble probabilities using the median STP on 03 June 2014. Different colors represent different methods. The dashed line indicates equivalent scores for the SPC and the NSSL-WRF ensemble. Filled circles are verified on all tornadoes and hollow circles are only verified on RM tornadoes.

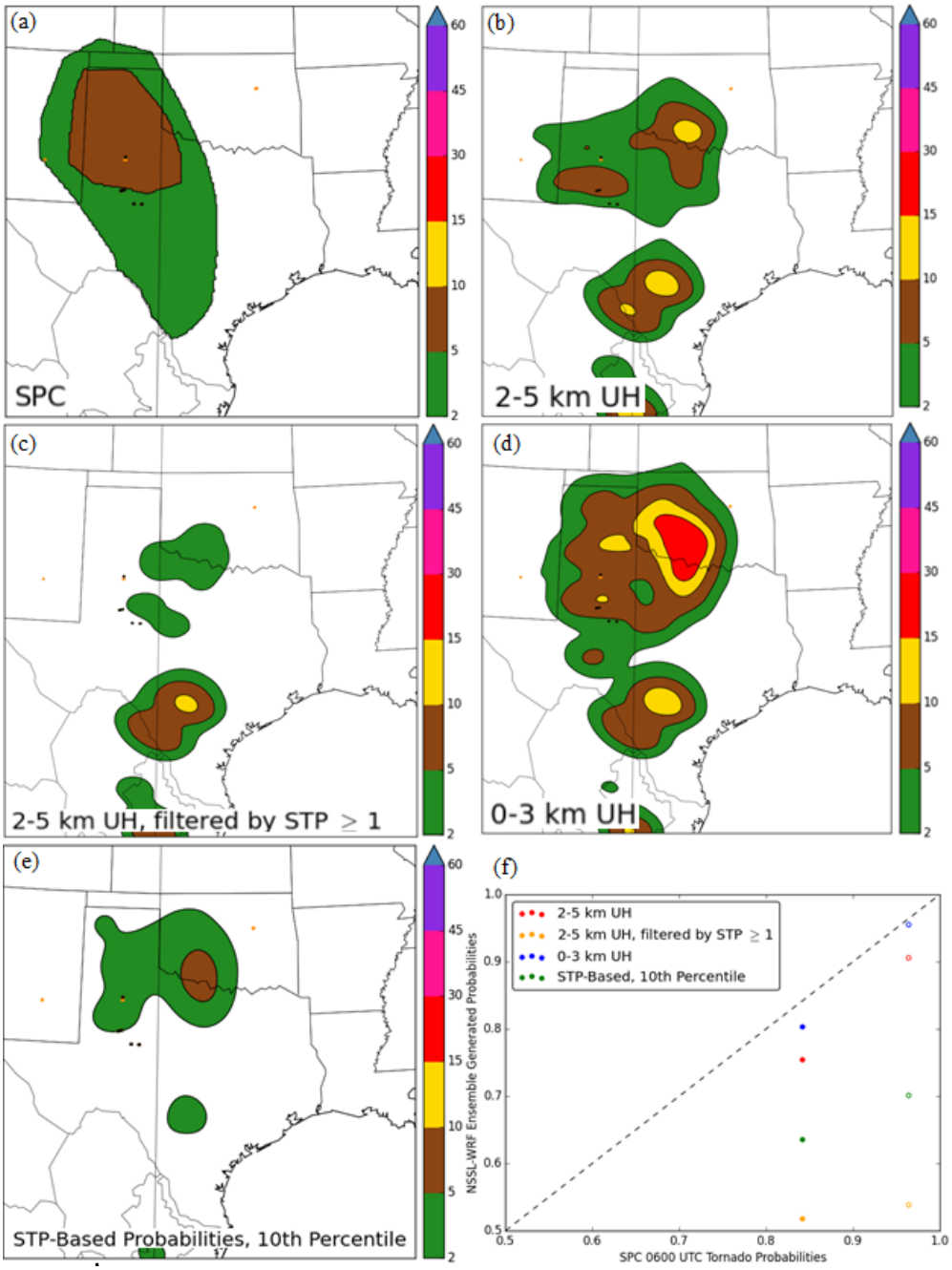

Figure 4.11 Forecast tornado probabilities for 5 May 2015 (a) issued at 0600 UTC by the SPC and generated with the NSSL-WRF ensemble, using (b) 2–5 km UH ≥ 75 m^2s^{-2}, (c) 2–5 km UH ≥ 75 m^2s^{-2} moving into an environment with STP ≥ 1, (d) 0–3 km UH ≥ 33 m^2s^{-2}, and (e) the 10^{th} percentile of STP from the hour previous to 2–5 km UH ≥ 25 m^2s^{-2}. All (orange) and RM (black) tornado paths are overlaid. (f) Daily ROC areas for the SPC and NSSL-WRF ensemble probabilities using the median STP on 5 May 2015. Different colors represent different methods. The dashed line indicates equivalent scores for the SPC and the NSSL-WRF ensemble. Filled circles are verified on all tornadoes and hollow circles are only verified on RM tornadoes.

Chapter 5: The Impact of Updraft Helicity Timing on Ensemble-Derived Tornado Probabilities

A paper to be submitted to *Weather and Forecasting*

Burkely T.Gallo[1], Adam J. Clark[2], Bryan T. Smith[3], Richard L. Thompson[3], Israel Jirak[3], and Scott R. Dembek[2,4]

[1]School of Meteorology, University of Oklahoma, Norman, Oklahoma

[2]NOAA/OAR National Severe Storms Laboratory, Norman, Oklahoma

[3]NOAA/NWS/NCEP Storm Prediction Center, Norman, Oklahoma

[4]Cooperative Institute for Mesoscale Meteorological Studies, University of Oklahoma, Norman, Oklahoma

Abstract

Probabilistic ensemble-derived tornado forecasts generated from convection-allowing models often use hourly maximum updraft helicity (UH) alone or in combination with environmental parameters as a proxy for right-moving (RM) supercells. However, large false alarm areas can occur from UH swaths associated with nocturnal mesoscale convective systems (MCSs), which climatologically produce fewer tornadoes than RM supercells. This study incorporates UH occurrence and timing with the forecast near-storm significant tornado parameter (STP) to calibrate the probability of a tornado. To generate the probabilistic forecasts, observed climatological frequencies of a tornado given a RM supercell and STP value are applied to the model output in three ways, two of which incorporate UH timing information. One method

uses the observed climatological frequency for a given 3-hr window to generate the probabilities. Another normalizes the observed climatological frequency by the number of hail, wind, and tornado reports observed in that 3-hr window compared to the maximum number of reports in any 3-hr window. The final method is independent of the time of UH occurrence and uses the observed climatological frequency encompassing all hours. The normalized probabilities reduce the false alarm area compared to the other methods, but have a smaller area under the ROC curve and require a much higher percentile of the STP distribution to be used in probability generation to become reliable. A case study demonstrates that the normalized probabilities focus on the most likely area for RM supercellular tornadoes, decreasing the false alarm generated by UH associated with nocturnal MCSs.

5.1 Introduction

The addition of convection-allowing model (CAM) ensembles to the suite of available numerical guidance provides severe convective forecasters guidance on convective mode when generating forecasts (Kain et al. 2008; Clark et al. 2012a). Indeed, as computing power increases, ever more guidance is becoming available to forecasters (Gallo et al. 2017a). As such, summary products for severe convective forecasters have been developed using storm-scale metrics alone and in combination with environmental information (Sobash et al. 2011; Gallo et al. 2016; Loken et al. 2017; Gagne et al. 2017). Many of the products include a measure of hourly maximum updraft helicity (UH; Kain et al. 2010), a storm-scale rotation metric which indicates a

forecasted midlevel mesocyclone and is often used as a proxy for a right-moving (RM) supercell (Naylor et al. 2012).

Since supercells produce many severe convective storm reports, UH has been a focus in forecasting severe convection (Sobash et al. 2011; Sobash et al. 2016a; Loken et al. 2017). Efforts have recently expanded from forecasting any type of severe convection to specific hazards (Gallo et al. 2016; Gagne et al. 2017) by including environmental parameters. One such parameter, the significant tornado parameter (STP), was developed by Thompson et al. (2003) and adapted by Thompson et al. (2012) to reflect environmental parameters important to tornadogenesis. STP was formulated using reanalysis soundings, but is a common environmental parameter in numerical weather prediction forecasts.

Smith et al. (2012) and Thompson et al. (2017) developed climatologies of tornado occurrence given a RM supercell and a STP value. These climatologies were used by Gallo et al. (2017b) to generate probabilistic tornado forecasts using a 10-member CAM ensemble with varying initial and lateral boundary conditions, based on a 4-km experimental version of the Weather Research and Forecasting (WRF; Skamarock et al. 2008) model run at the National Severe Storms Laboratory (NSSL), known as the NSSL-WRF ensemble (Gallo et al. 2016; Clark 2017). These probabilities were calibrated by empirical climatological frequencies, resulting in skillful forecasts of tornadoes from RM supercells (RM tornadoes) that overforecast tornado occurrence slightly (Gallo et al. 2017b). This probability generation method was more reliable and skillful than other methods of probabilistic forecast generation that treat the UH occurrence and the STP value as thresholds to be exceeded, rather than treating each

point probabilistically. Those forecasts also generated large false alarm areas linked to mesoscale convective systems (MCSs), which are less likely to produce tornadoes than supercells (Smith et al. 2012). This work attempts to use the observed climatology and timing of UH occurrence to reduce the false alarm areas from UH associated with MCSs.

Section 5.2.1 briefly describes how this study adapts the methodology of Gallo et al. (2017b) using normalization techniques, and section 5.2.2 describes the data and verification metrics used. Section 5.3.1 shows the aggregated statistical results, while section 5.3.2 gives an example case study. Finally, section 5.4 presents conclusions and ideas for future work.

5.2 Data and Methodology

5.2.1 Probabilistic Forecast Generation

Probabilistic forecasts were generated following the technique of Gallo et al. (2017b), which incorporates empirical environmental frequencies of a tornado given a RM supercell and a modified STP value (Fig. 5.1; black line). The modified STP is defined by:

$$STP = \left(\frac{SBCAPE}{1500 Jkg^{-1}}\right) * \left(\frac{SHR6}{20 ms^{-1}}\right) * \left(\frac{SRH1}{150 m^2 s^{-2}}\right) * \left(\frac{2000m - SBLCL}{1000m}\right) * \left(\frac{200 Jkg^{-1} + SBCIN}{150 Jkg^{-1}}\right) \quad (5.1)$$

where SBCAPE, SBCIN, and SBLCL are the convective available potential energy (CAPE), convective inhibition (CIN), and lifted condensation level (LCL) at the surface, respectively, SHR6 is the 0-6 km shear, and SRH1 is the 0–1 km storm-relative helicity. This STP utilizes the capping functions from the effective-layer STP [e.g., if SHR6 < 12.5 m s^{-1}, that term is set to 0; Thompson et al. (2012)] while recognizing the

inability to efficiently calculate the effective inflow layer at every point within the CAM ensemble. The tornado occurrence frequencies utilize 1202 tornado reports and 5422 hail or wind reports occurring from February 2014–December 2015 (Thompson et al. 2017; Gallo et al. 2017b) and are calculated by dividing the number of tornado reports from RM supercells [using Smith et al. (2015)'s RM supercell definition] by the number of hail, wind, and tornado reports from RM supercells in each STP bin.

To apply these frequencies to a CAM ensemble, NSSL-WRF ensemble forecasts initialized at 0000 UTC and extending to 36 hours were used. Following Gallo et al. (2017b), hourly forecast values of UH and STP were extracted. For each member and forecast hour between 1200 UTC and 1200 UTC the following day (spanning forecast hours 12–36), each gridpoint was checked for UH exceeding 25 m^2s^{-2} anywhere within a 40 km radius as a proxy for a RM supercell. If UH exceeded this threshold, the point STP value from the previous hour was added to a distribution of UH at that gridpoint. Next, a percentile of the distribution was selected as the representative STP value for that forecast hour. The daily maximum STP value from this process was then input into the empirical climatological frequencies, resulting in a probability for that point and member. An average of the probabilities at each grid point was taken across all members, and smoothed using a Gaussian kernel to generate a final probabilistic field similar to the SPC's probabilistic forecasts.

When this methodology used the observed climatological frequencies generated independent of time, often UH swaths associated with nocturnal MCSs produced false alarm areas (Gallo et al. 2017b). While RM tornado reports show a steep peak during the afternoon hours, overnight hours contain only a small fraction of reports (Fig. 5.1).

However, the diurnal UH cycle maintains UH throughout the evening hours at even high thresholds. Thus, to reduce false alarm two methods of incorporating timing information through the climatological frequencies were applied. While Gallo et al. (2017b) calculated the probabilities using an equation independent of the report occurrence time, this study broke apart the climatological frequencies using a moving three-hour time window centered on the hour of interest (Fig. 5.2a). This approach will be known as the non-normalized time-dependent method. The other method incorporated the frequency at each hour *and* the total number of hail, wind, and tornado reports occurring in that window as compared to the maximum three-hour window by weighting each hour according to the number of reports occurring therein. The three-hour window containing the most reports (2300 UTC) had a weight of one (Fig. 5.2b). This approach will be known as the normalized time-dependent method. The final approach follows the method of Gallo et al. (2017b), utilizing frequencies calculated for the entire day, and will be called the daylong method. Additionally, the daylong probabilities interpolated between STP bins, whereas the time-dependent probabilities did not due to a smaller sample size for the three-hour windows.

5.2.2 Verification Metrics and Data

Ensemble-generated forecasts were verified alongside the 0600 UTC forecasts from the Storm Prediction Center (SPC), since the ensemble probabilities are designed as operational first-guess tornado forecasts and ideally would behave comparably to the SPC forecasts. Verification metrics used include the area under the receiver operating curve (ROC area; Mason 1982), reliability diagrams, and performance diagrams (Roebber 2009). The ROC area describes how forecasts discriminate areas of event

occurrence from areas of event non-occurrence by plotting the probability of detection (POD) vs. the probability of false detection (POFD), but contains no bias information. Reliability diagrams plot the observed frequency vs. the forecast probability, complementing the ROC areas. Performance diagrams visualize four different contingency-table-based metrics, including the bias, the success ratio (SR), the POD, and the critical success index (CSI), which is often used as a rare-event score. (Wilks 2011). Statistics were generated at each of the probability thresholds forecast by the SPC: 2%, 5%, 10%, 15%, 30%, 45%, and 60%. Verification statistics were computed across 182 days in the 2014 and 2015 spring seasons, defined as April–June, over approximately the eastern 2/3 of the CONUS. Observed tornado path data were regridded to the 4 km NSSL-WRF ensemble grid prior to verification, and treated as yes/no events. A yes event occurred if a tornado passed within 40 km of a point, consistent with the SPC's forecast probabilities.

5.3 Results

5.3.1 Seasonal Performance Statistics

The most reliable probabilities with sufficiently high ROC areas were compared between each method. ROC areas for all percentiles of the daylong and the non-normalized time-dependent probabilities were higher than the SPC forecasts, while all of the normalized time-dependent probabilities had lower ROC areas than the SPC (in most cases because the POD was lower than the SPC with a very similar POFD; not shown). Differences between ROC areas of STP percentiles within each forecast method were minimal, contrasting with the reliability, which varied greatly within

methods and between methods. Thus, the most reliable percentiles for each method were chosen for comparison: the 10th percentile for the daylong (Fig. 5.3a) and non-normalized time-dependent (Fig. 5.3b) probabilities, and the 75th percentile for the normalized time-dependent probabilities (Fig. 5.3c). The ROC areas were very similar between methods, excepting the normalized time-dependent probabilities which had a lower ROC area than the other methods due to both decreased POD and POFD (Fig. 5.3d). The reliability of these four methods was also similar, and all reliably forecasted RM tornadoes up to the 15% threshold (Fig. 5.3e).

A performance diagram shows that the CSI of the normalized time-dependent probabilities is consistently higher than the other methods, despite a lower ROC area (Fig. 5.4). For example, at the 2% threshold, its POD is much lower than the other methods, with a slight increase in SR. At the 5% threshold, the SPC has the highest POD of any forecast, but also has a lower SR than either set of time-dependent forecast probabilities. At the higher-impact 10% and 15% probabilities all methods have similar PODs, but the first-guess probabilities have less false alarm than the SPC forecasts. The 10% forecast threshold also has the highest CSIs of any forecast threshold.

While the seasonally aggregated statistics show highly similar forecast methods, the aim of incorporating the time of UH occurrence is to reduce the nocturnal MCS-associated false alarm. To determine the impact of the timing information, probabilities were also generated using all methods for each hour and averaged across the domain. The diurnal cycle of the daylong and the non-normalized time-dependent probabilities maintained areas of probability throughout the nocturnal hours, while the normalized time-dependent probabilities showed a sharp decrease from the afternoon peak that

resulted in nearly zero probability overnight (Fig. 5.5). The non-normalized time-dependent probabilities increased the afternoon probabilities as compared to the daylong probabilities, and the normalized time-dependent probabilities increased the afternoon probabilities even further, likely due to the different percentiles of STP used to generate the probabilities. The peaks of the average probabilities are offset from the peak report time, and this is speculated to be due to the probabilities being evaluated only over the spring season, while the diurnal cycle of reports utilized data from February 2014 through December 2015.

5.3.2 Case Study: 29 June 2014

A case study illustrates the forecast improvement provided by including the time of UH occurrence in the probability generation, particularly in reducing the threat from nocturnal MCSs. 29 June 2014 had a surface low-pressure center evolving across the south-central High Plains, with ample low-level moisture ahead of the main low. The 0600 UTC convective outlook from the SPC mentions appreciable uncertainty in the storm coverage and timing, making this a case where forecasters could use first-guess tornado guidance that reduces false alarm from non-favorable convective modes. The SPC highlighted a 10% tornado threat across the Iowa/Missouri border, with a broad 5% extending north through Wisconsin and west to the middle of Nebraska (Fig. 5.6a). A few initial supercells developed near a residual outflow boundary, but a complex storm evolution with multiple mergers ensued and a MCS developed around 0300 UTC. The tornado threat was primarily associated with the supercellular storms; twelve RM tornadoes occurred out of fourteen total tornado reports. All ensemble-generated probabilities have the same magnitude as the 0600 UTC SPC forecasts: 10% (Fig. 5.6b-

d). However, the placement and extent of the 10% probabilities differ. The daylong probabilities and the normalized time-dependent probabilities both have a broad swath of probabilities extending into Illinois and a secondary area of probabilities across Kentucky, whereas the normalized time-dependent probabilities correctly eliminate this area because the UH was occurring at 0300–0600 UTC (Fig. 5.6e). The normalized time-dependent probabilities actually also increased the probabilities where tornadoes occurred, resulting in a better forecast on this day.

5.4 Summary and Discussion

Probabilities were developed that consider the time of UH occurrence within an ensemble and the climatological frequency of a tornado given the existence of a right-moving supercell. These probabilities address a shortcoming of prior first-guess forecasts, which often had false alarm associated with UH produced by nocturnal MCSs. Weighting the timing information by the overall number of reports during a given three-hour window further lessens the nocturnal false alarm, as the most heavily weighted time occurs in the same window as the majority of reports: around 0000 UTC. The normalized time-dependent probabilities had lower ROC areas than any other method, likely because the reduction in area covered by the probabilities decreased the POD. Since tornadoes are rare events, missed events greatly affect the statistical scores. The CSI of the normalized time-dependent probabilities suffered less from missed events, reflecting the improvement in reducing false alarm. Overall, the normalized time-dependent probabilities performed well, particularly at high probabilistic thresholds, which often have larger potential impacts than the lower, more common

thresholds. At these higher thresholds, the normalized time-dependent probabilities maintained as high or higher PODs than other forecast methods, while also maintaining high SRs. The diurnal cycle of the normalized time-dependent probabilities more accurately reflects the diurnal report cycle than the other probabilities do, decreasing the nocturnal false alarm area compared to the UH occurrence. Reducing the false alarm generated by UH from nocturnal MCSs via the timing of UH occurrence focuses the forecast on areas at a risk of supercellular tornadoes, remaining true to the underlying climatological frequencies used to generate the probabilistic forecast while providing forecasters with a skillful and reliable first guess tornado forecast.

Acknowledgements

The authors thank Chris Melick and Robert Hepper of the SPC for providing regridded SPC forecasts, and Andrew Dean of the SPC for obtaining the data for the climatological frequency calculations. This material is based upon work supported by the NSF Graduate Research Fellowship under Grant No. DGE-1102691, Project #A00-4125. BTG, AJC and SRD were provided support by NOAA/Office of Oceanic and Atmospheric Research under NOAA-University of Oklahoma Cooperative Agreement #NA11OAR4320072, US Department of Commerce. AJC also received support from a Presidential Early Career Award for Scientists and Engineers.

Figures

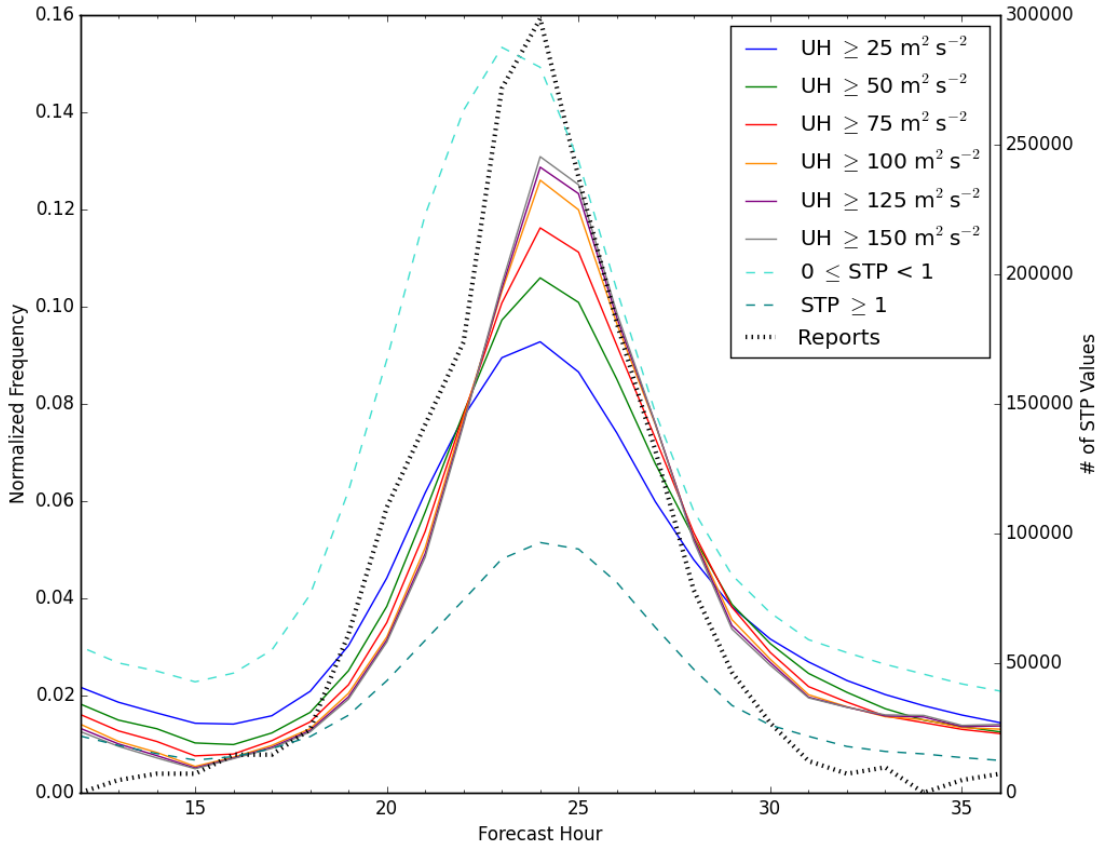

Figure 5.1 Report, UH, and STP diurnal distributions. Plots begin at forecast hour 13, corresponding to 1300 UTC on the day of the forecasts and end at forecast hour 36, corresponding to 1200 UTC on the following day.

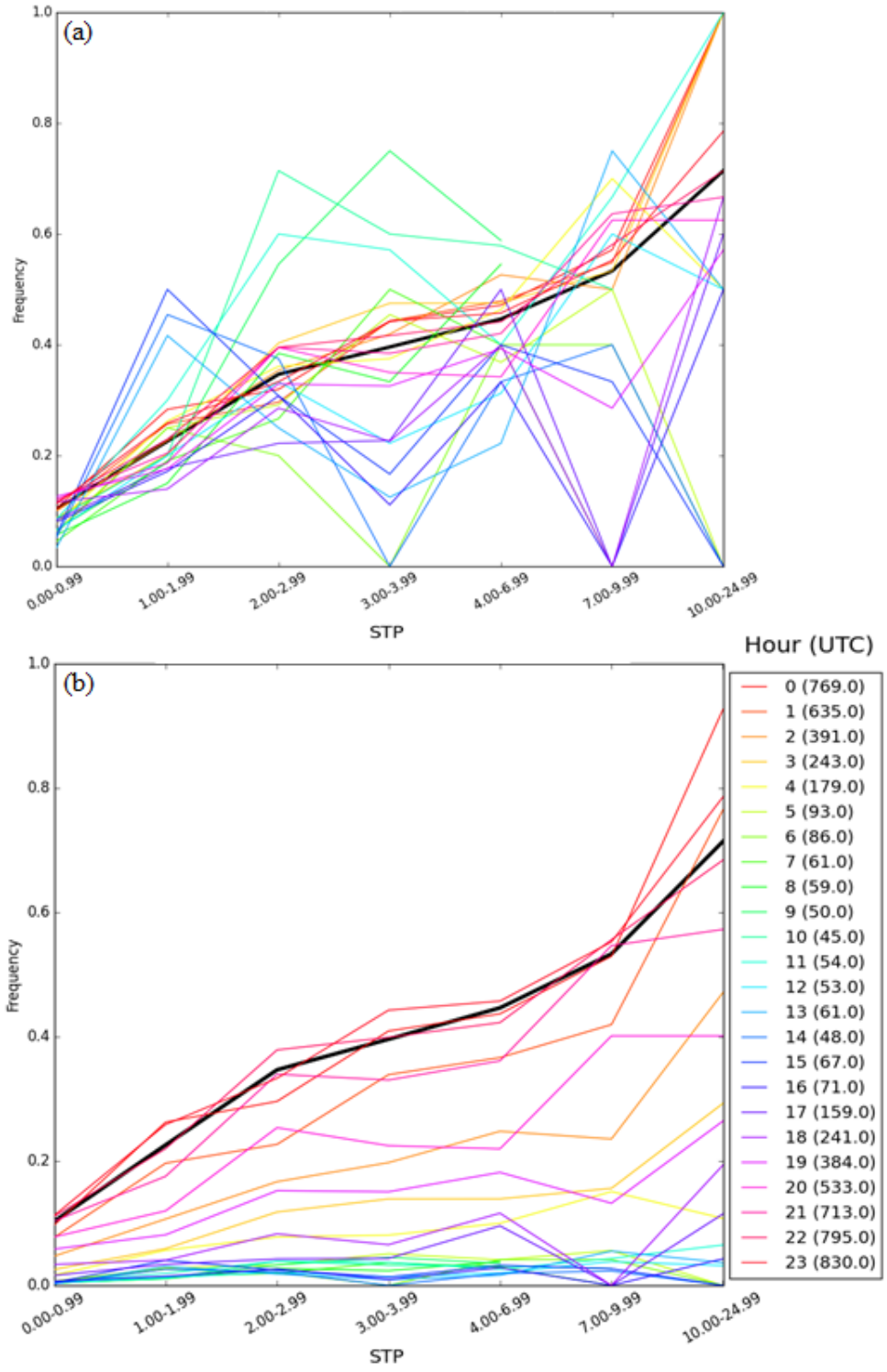

Figure 5.2 (a) Climatological frequency of tornado occurrence given a RM supercell, time of day, and STP value based on data from February 2014–December 2015. Each colored line represents the center of a 3-hour time window. (b) The climatological frequencies of tornado occurrence normalized by the maximum number of hail, wind, and tornado reports in a given three-hour window.

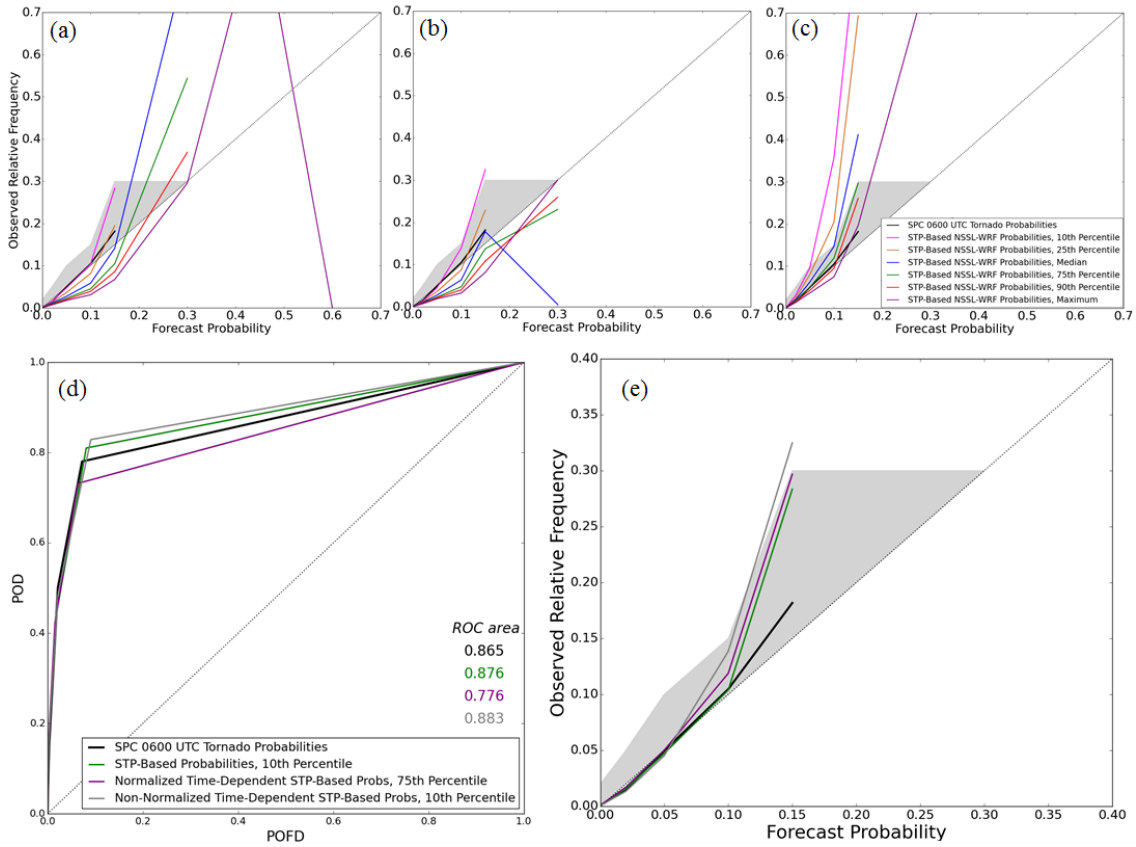

Figure 5.3 Reliability diagrams for different percentiles of STP used to formulate (a) the daylong probabilities, (b) non-normalized time-dependent probabilities, and (c) the normalized time-dependent probabilities. The diagonal represents perfect reliability, and the shaded area shows where SPC forecasts can be considered reliable. (d) ROC curve, with the diagonal representing a forecast with no skill, and (e) reliability diagram for the SPC and selected percentiles of each probabilistic forecast generation method, with the shading and diagonal as in (a) – (c).

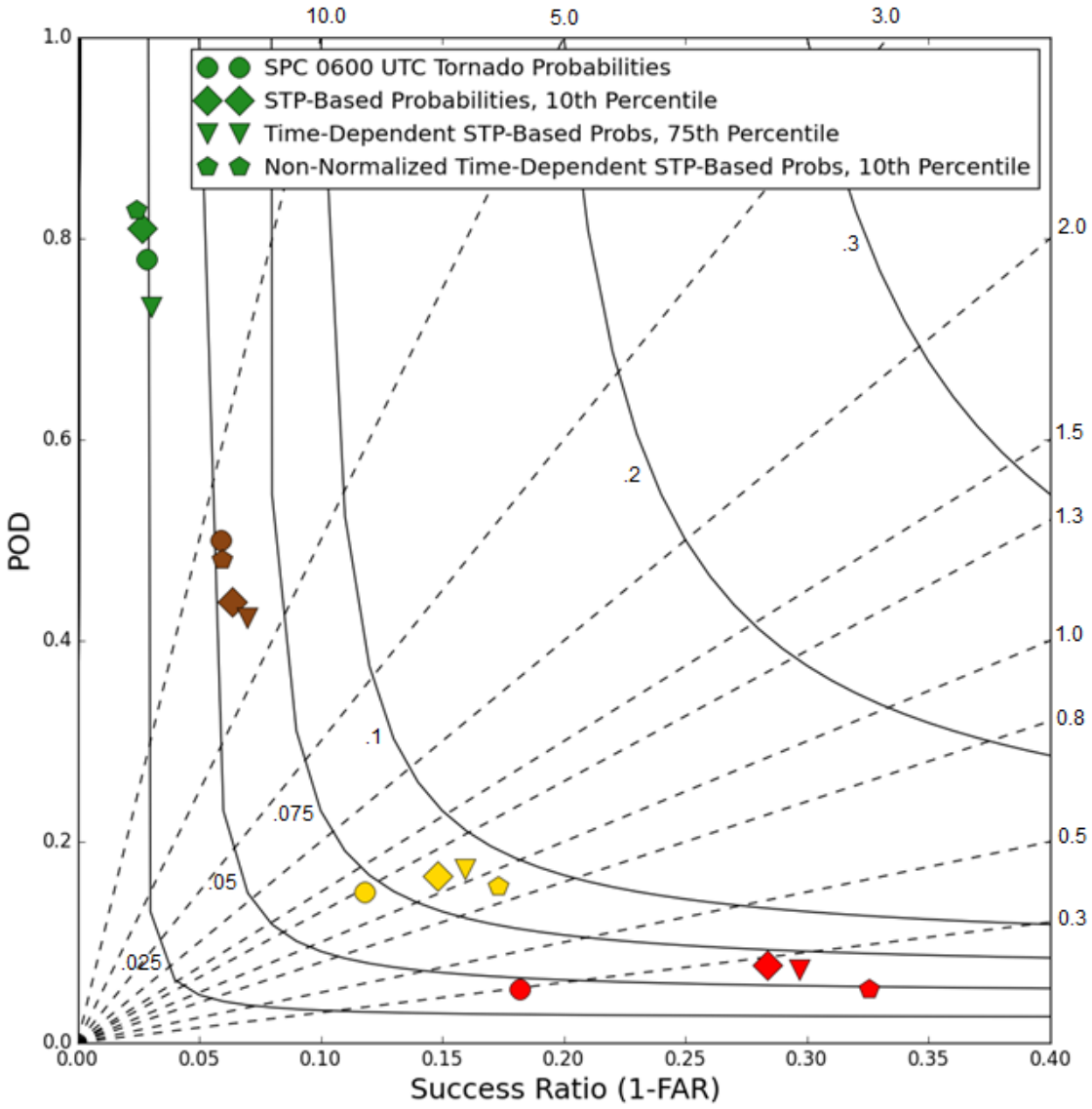

Figure 5.4 Performance diagram for the three different methods of probability generation and the SPC. Green, brown, yellow, and red shapes represent the 2%, 5%, 10%, and 15% forecast threshold, respectively. Dashed lines are of constant bias, and solid curved lines are lines of constant CSI. FAR stands for the false alarm ratio.

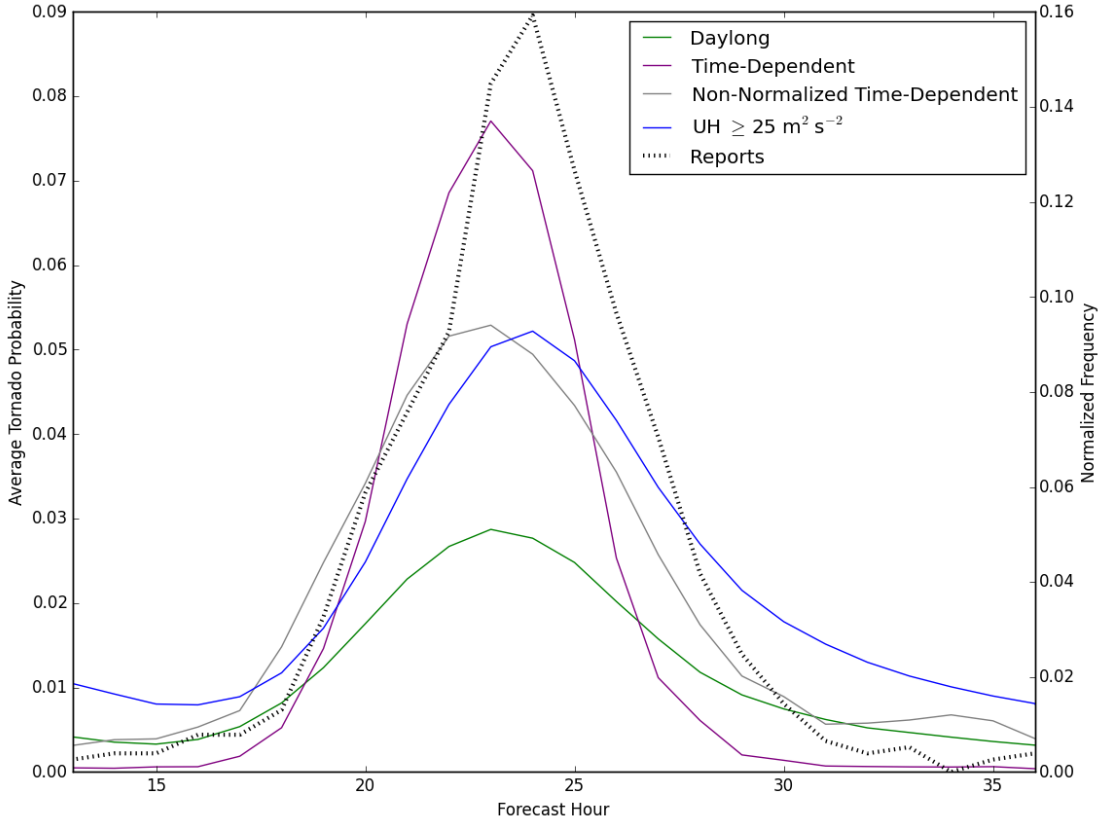

Figure 5.5 The diurnal cycle of report frequency, UH frequency, and average probability over the verification domain for each forecast method.

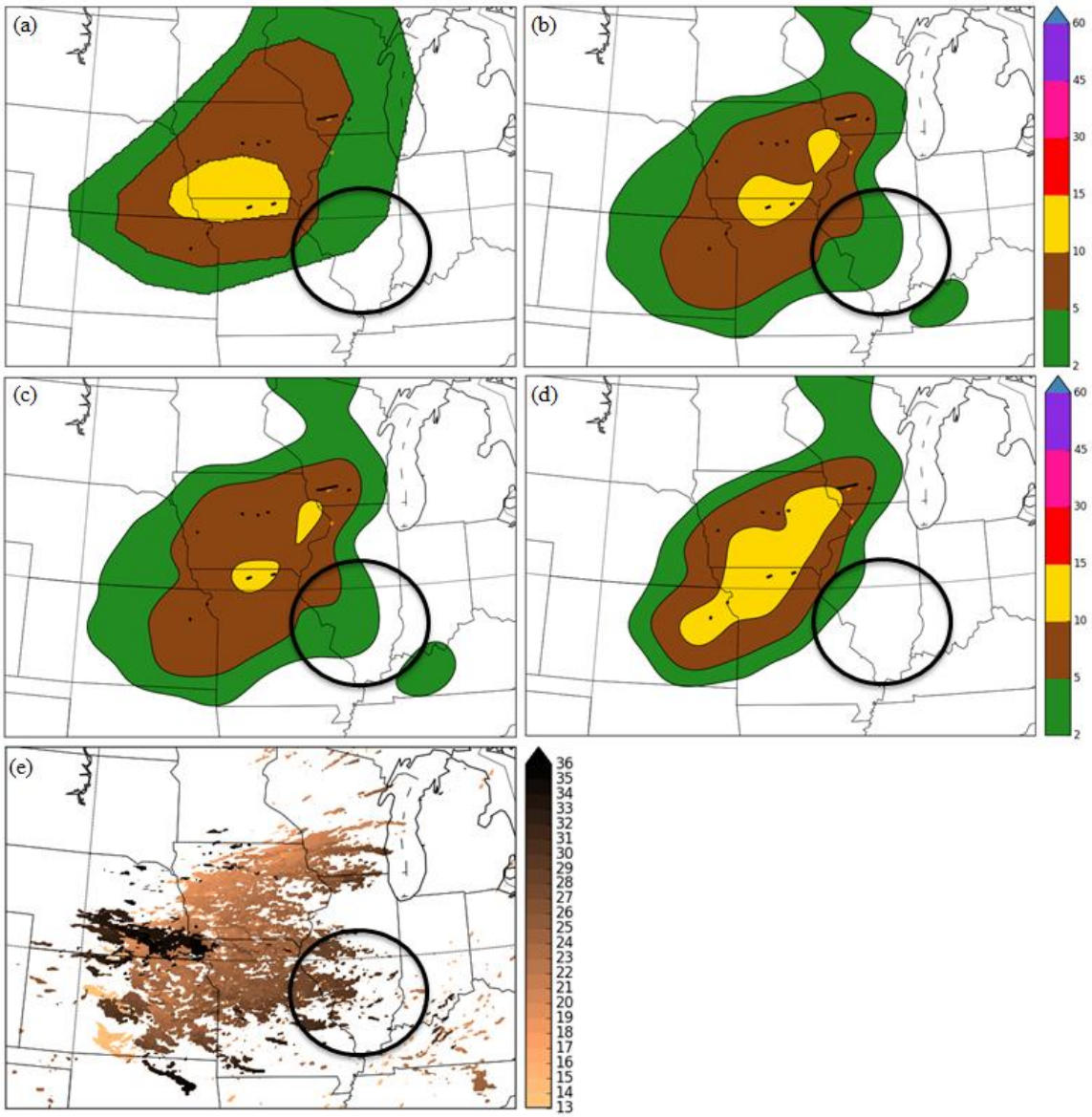

Figure 5.6 Tornado forecasts for 29 June 2014 from (a) the SPC, (b) the daylong probabilities, (c) the non-normalized time-dependent probabilities, and (d) the normalized time-dependent probabilities. Black lines show the tracks of RM tornadoes, while orange lines represent non-RM tornadoes. (e) Ensemble 2–5 km UH $\geq 25 m^2 s^{-2}$, color-coded by hour of UH occurrence. The circle highlights a large area of nocturnal UH reduced by the normalized probabilities.

Chapter 6: Conclusions and Future Work

6.1 General Conclusions

As convection-allowing models (CAMs) and ensembles proliferate, the amount of information available to forecasters continually increases. However, forecasters must still follow a strict operational schedule for product issuance, and may not be able to incorporate all of the available information from NWP into their forecasts. Since CAMs have been shown to provide useful guidance regarding convective initiation, evolution, and mode (Kain et al. 2008; Clark et al. 2012a), forecasters focused on severe convection should be able to easily utilize the information from increasingly sophisticated CAMs.

In particular, the availability of convective mode information via storm-scale rotation metrics such as updraft helicity (UH) has the potential to greatly benefit forecasters, as it has been shown to provide reliable guidance regarding the occurrence of severe convective weather in the form of wind, hail, or tornadoes (Sobash et al. 2011). Since tornadoes have a large societal and economic impact (Simmons and Sutter 2011), being able to differentiate the tornado threat from the general severe convective threat would benefit forecasters by enabling them to better prepare the general public and partners such as emergency managers for potential impacts. The storm-scale metrics unique to CAMs can be combined with more traditional environmental fields used in an ingredients-based method of forecasting (Doswell et al. 1996), adding to a forecaster's tornado prediction toolbox.

By distilling environmental characteristics conducive to tornadogenesis and storm-scale rotation metrics, this work generated first-guess tornado probabilities aimed

at operational forecasters. Testing the probabilities in a real-time, operational framework such as the Spring Forecasting Experiments (SFEs) conducted in NOAA's Hazardous Weather Testbed allowed for instant feedback from forecasters, as well as model developers and researchers. This work began with a simple evaluation of how UH, a mesocyclone-scale rotation diagnostic, could be used as a coarse proxy for tornadoes. From that initial step, environmental information was added through increasingly complex methodologies. The initial attempt at limiting the influence of UH to regions with favorable environmental parameters to tornadogenesis shifted after feedback from forecasters in the SFEs indicated that the magnitudes of the probabilities were too high. The probabilities resulting from that feedback treated each point as having a probability of tornado occurrence based on the model storm environment attributes, rather than assuming a tornado once a threshold of UH occurred. A second round of feedback affecting this work occurred in SFE 2017, when many participant comments indicated that the updated probabilities were too high in nocturnal MCS situations. That feedback motivated the incorporation of timing information into the probabilities, to focus the probabilities on right-moving (RM) supercells, which more often occur during the afternoon and evening. The research-to-operations, operations-to-research framework was fundamental in developing the hypotheses in this dissertation, which are resolved as follows:

Adding high-resolution information to constrain tornado probabilities to areas that are environmentally favorable to tornadogenesis will result in more skillful probabilities than solely using 2–5 km UH.

Subjectively and objectively, forecasts that incorporate environmental information are found to be more successful than forecasts using only UH in all metrics except for the area under the ROC curve. This hypothesis was tested by comparing forecasts that used UH $\geq$ 75 m^2s^{-2} alone, forecasts that used UH $\geq$ 75 m^2s^{-2} only where the environment in the previous hour had an LCL height less than 1500 m and a SBCAPE/MUCAPE ratio of .75 or greater, forecasts that used UH $\geq$ 75 m^2s^{-2} only where the environment in the previous hour had a STP of one or greater, and forecasts that used UH $\geq$ 25 m^2s^{-2} and assigned a probability of a tornado based on empirical climatological frequencies. These thresholds were chosen to (1) rule out elevated and high-based storms, which were less likely to produce a tornado, (2) focus on areas with favorable conditions for tornadogenesis, and (3) make use of observed tornado frequencies and look beyond a threshold exceedance paradigm.

Objectively, the addition of environmental information improved the reliability of forecasts over the 182 cases examined. Improvement in the reliability occurred whether the information was incorporated as an additional threshold criterion or was incorporated probabilistically. The CSI generally also saw an improvement when environmental information was used. The decrease in the ROC area compared to using UH information only is largely due to a decrease in POD. This decrease in ROC area was less prevalent when focusing on RM supercellular tornadoes. However, ROC areas of all sets of probabilities remained above the 0.7 threshold of a skillful forecast. Subjectively, participants in SFE 2015 often rated the probabilities incorporating environmental information higher than the probabilities without environmental information, as the probabilities that only used UH often had magnitudes that were too

high. Since UH indicates mesocyclones, the high false alarm generated by using a threshold of UH as a coarse tornado proxy was expected.

Incorporating observed tornado frequencies given a right-moving supercell will provide more accurate and reliable probabilities than those generated solely using model-derived information.

Similar to the previous hypothesis, this hypothesis was tested by comparing probabilistic forecasts generated using UH $\geq$ 75 m^2s^{-2} where STP $\geq$ 1 in the previous hour and forecasts generated using UH $\geq$ 25 m^2s^{-2} where a probability of tornado occurrence was assigned based on the STP value in the previous hour. Empirical frequencies of a tornado given a RM supercell and a Local Storm Report (LSR) were used to assign the probability of a tornado at every point in the second set of probabilities, limiting the magnitude to observed frequencies. This approach therefore decreased some of the over-forecasting problem that occurred in the approaches that solely used model-derived information, resulting in more reliable forecasts than the forecasts that required a threshold of UH and STP. Other statistical metrics showed better scores for the probabilistic approach as well. The area under the ROC curve for the forecasts using the empirical frequencies was higher than the ROC area for the forecasts using thresholds of UH and STP. This result shows that not only were the forecasts that used empirical information more reliable, they also better discerned between areas of tornado occurrence and non-occurrence. In other words, the probabilities that solely used thresholds of UH and STP had less accurate forecasts and more over-forecasting than the probabilities that incorporated the empirical frequencies.

Tornado probabilities generated using a convection-allowing ensemble can be used operationally as first-guess tornado forecasts and have similar verification statistics to initial probabilistic tornado forecasts issued by the Storm Prediction Center (SPC) at 0600 UTC.

To test this hypothesis, six methods of probability generation were compared to initial probabilistic tornado forecasts issued by the SPC: three methods used solely model-derived information, while three other methods incorporated empirical frequencies. Of the three methods that incorporated empirical frequencies, two of those methods used information about the time of UH occurrence. Probabilities were evaluated only at probabilistic forecast thresholds used by the SPC, enabling an apples-to-apples comparison between the methods. The SPC forecasts were extremely reliable at all forecast thresholds and maintained skillful ROC areas with high PODs and low POFDs. Of the first-guess forecasts evaluated herein, the forecasts with the highest ROC areas also suffered from severe over-forecasting, to the point where they were not useful as first-guess probabilities for SPC forecasters. In particular, the UH-only forecast methods and the method that incorporated a threshold of STP over-forecast drastically. The methods incorporating the environmental frequencies performed more similarly to the SPC forecasts, with ROC areas that were statistically the same as the SPC's, and had comparable reliabilities. The similarity in the statistics between the SPC forecasts and the probabilities that incorporated empirical frequencies support the hypothesis that these probabilities could be used operationally as first-guess tornado forecasts. For actual operational usage, the probabilities will need to be incorporated

into an ensemble that produces first-guess forecasts prior to 0600 UTC, and this is the subject of ongoing work.

Incorporating temporal information regarding UH occurrence will reduce areas of false alarm linked to nocturnal MCSs, which often produce UH in NWP but do not often produce tornadoes.

The incorporation of timing information via weighting of probabilities reduced nocturnal false alarms, supporting this hypothesis. By taking into account the timing of the UH as well as the distribution of hail, wind, and tornado LSRs, normalized probabilities were created that reduced nocturnal false alarm. This result was shown from generating the probabilities for each hour and plotting the diurnal cycle of average probability for each forecast method. A reduction in nocturnal probability was found in the normalized probabilities, and no such reduction occurred in the probabilities that did not incorporate UH timing information. The non-normalized method of incorporating UH timing information, which did not utilize the diurnal report distribution, amplified the afternoon peak in probability compared to the daylong probabilities, but maintained probability overnight. The normalized probability method is considered a workaround for the mode problem posed by nocturnal MCSs, but it will dampen signals from nocturnal supercells that may be tornadic.

6.2 Directions for Future Research

The analysis herein focused on the most common severe weather season in the Great Plains of the United States: April through June. Thus, the applicability of these probabilities across seasons is unknown and should be the focus of future work.

Verification of the probabilities also encompassed the entire eastern two-thirds of the CONUS, and smaller regions were not considered. Therefore, future work could also examine how these probabilities perform in different regions, as Sobash and Kain (2017) did for probabilities of any type of severe convective hazard.

As mentioned previously, the probabilities herein were applied to the NSSL-WRF ensemble, so forecasts were only available after the initial 0600 UTC SPC forecasts had been issued. While the probabilities were then able to inform updates to the forecast, having the probabilities available from the start of the forecast process would ultimately be more useful. As such, work is ongoing to incorporate these probabilities into the High-Resolution Ensemble Forecast, version 2 (HREFv2), an operationalized version of the SPC's Storm Scale Ensemble of Opportunity (SSEO). When applied to the HREFv2, the probabilities can be updated twice daily, and become available prior to the initial forecasts. Besides being adaptable to different ensemble configurations, the probabilities could also be applied to deterministic forecasts. Since the ensemble provides some smoothing to the probabilities herein, a different Gaussian kernel may be needed for deterministic forecasts to achieve probabilities at a similar resolution to the current SPC forecasts.

Additionally, while the normalized probabilities provide a passable workaround for the mode problem, more explicit detections of convective mode could be incorporated into future forecasts. Specifically, object-based verification and detection methods could be applied to reflectivity and UH fields to determine a forecast convective mode, and the probabilities could be applied accordingly. Since these probabilities are based upon and designed to forecast RM supercells, detection of mode

would ensure that they are applied solely to simulated RM supercells. Furthermore, the database from which the climatological frequencies are generated contains modes besides RM supercells; additional probabilities could be generated for MCS objects based on the frequency of a tornado given an MCS and a certain value of STP.

Besides STP, there are many other environmental variables that could influence tornadogenesis. One way to find the key ingredients to forecasting tornadoes from CAM ensembles is to feed a multitude of variables through a machine-learning algorithm to determine which variables have the most influence on the number of tornadoes in any given day. Since the variables in question would be known meteorological variables, this process could give forecasters insight into which model-produced environmental or storm-scale attributes could have an influence on simulated storms. Such an approach could provide a fruitful partnership between statistical models and operational forecasting.

Since some of the variables produced by CAM ensembles are difficult to directly observe (e.g., UH), verification of these fields is difficult. More work is needed to understand the link between UH and observed mesocyclone-scale rotation strength, particularly since grid spacing has a large influence on UH intensity. Extreme values of UH tend to draw the eye of those using CAMs and CAM ensembles, and therefore the link between UH and observed storms needs to be more closely studied to gain a better understanding of how the simulated correlates to the observed.

Finally, the continuous feedback between researchers and forecasters in developing this work was a crucial component that should be incorporated into future studies. Developing forecast products that are skillful and operationally useful is an

imperative task in the field of meteorology, and soliciting forecaster opinions throughout the research process strengthens the final products immeasurably.

References

Adams-Selin, R., C. Ziegler, and A. J. Clark, 2014: Forecasting hail using a one-dimensional hail growth model inline within WRF. In Proceedings, 27th Conference on Severe Local Storms, Madison, WI, Amer. Met. Soc., 11B.2.

Adams-Selin, R., and C. Ziegler, 2016: Forecasting hail using a one-dimensional hail growth model within WRF. *Mon. Wea. Rev.*, **144**, 4919-4939. doi: 0.1175/MWR-D-16-0027.1.

Alexander, C. R., S. S. Weygandt, T. G. Smirnova, S. Benjamin, P. Hofmann, E. P. James, and D. A. Koch, 2010: High Resolution Rapid Refresh (HRRR): Recent enhancements and evaluation during the 2010 convective season. *Preprints*, 25th Conf. on Severe Local Storms, Denver, CO, Amer. Meteor. Soc., 9.2.

Alexander, C. R., S. Weygandt, S. Benjamin, D. C. Dowell, M. Hu, T. Smirnova, J. B. Olson, J. Kenyon, G. Grell, E. P. James, J. M. Brown, and H. Lin, 2015: The 2015 Operational Upgrades to the Rapid Refresh (RAP) and High-Resolution Rapid Refresh (HRRR). *27th Conf. on Weather Analysis and Forecasting/23rd Conf. on Numerical Weather Prediction*. Chicago, IL. 2A.2.

Aligo, E., B. S. Ferrier, J. Carley, E. Rogers, M. Pyle, S. J. Weiss, and I. L. Jirak, 2014: Modified Microphysics for Use in High-Resolution NAM Forecasts. *27th Conf. on Severe Local Storms*. Madison, WI. 16A.1.

Anderson, J. L., 2001: An Ensemble Adjustment Kalman Filter for Data Assimilation. *Mon. Wea. Rev.*, **129**, 2884–2903. doi: 10.1175/1520-493(2001)129<2884:AEAKFF>2.0.CO;2

Anderson, J. L., 2003: A local least squares framework for ensemble filtering. *Mon. Wea. Rev.*, **131**, 634-642.

Anderson, J., T. Hoar, K. Raeder, H. Liu, N. Collins, R. Torn, and A. Avellano, 2009: The Data Assimilation Research Testbed: A Community Facility. *Bull. Amer. Meteor. Soc.*, **90**, 1283–1296. doi: 10.1175/2009BAMS2618.1.

Anderson-Frey, A., 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.

Benjamin, S. G., and Coauthors, 2016: A North American hourly assimilation and model forecast cycle: The Rapid Refresh. *Mon. Wea. Rev.*, **144**, 1669–1694, doi:10.1175/MWR-D-15-0242.1.

Blair, S.F., J.M. Laflin, D.E. Cavanaugh, K.J. Sanders, S.R. Currens, J.I. Pullin, D.T. Cooper, D.R. Deroche, J.W. Leighton, R.V. Fritchie, M.J. Mezeul II, B.T. Goudeau, S.J. Kreller, J.J. Bosco, C.M. Kelly, and H.M. Mallinson, 2017: High-

Resolution Hail Observations: Implications for NWS Warning Operations. *Wea. Forecasting*, **32**, 1101–1119, doi: 10.1175/WAF-D-16-0203.1.

Bothwell, P.D., J.A. Hart, and R.L. Thompson, 2002: An integrated three-dimensional objective analysis scheme, *Preprints*, 21st Conf. on Severe Local Storms, San Antonio, TX.

Bougeault, P., and P. Lacarrère, 1989: Parameterization of orography-induced turbulence in a mesobeta-scale model. *Mon. Wea. Rev.*, **117**, 1872–1890, doi:10.1175/1520-0493(1989)117,1872: POOITI.2.0.CO;2.

Boustead, J. M., B. E. Mayes, W. Gargan, J. L. Leighton, G. Phillips, and P. N. Schumacher, 2013: Discriminating environmental conditions for significant warm sector and boundary tornadoes in parts of the Great Plains. *Wea. Forecasting*, **28**, 1498–1523, doi:10.1175/WAF-D-12-00102.1.

Brewster, K.A., D.R. Stratman, and R. Hepper, 2016: 4-D Visualization of storm-scale forecasts using VAPOR in the Hazardous Weather Testbed Spring Forecasting Experiment. *Preprints, 28th Conf. on Severe Local Storms*. Portland, OR. 15B.6.

Brimelow, J.C., G. W. Reuter, and E. R. Poolman, 2002: Modeling maximum hail size in Alberta thunderstorms. *Wea. Forecasting*, **17**, 1048-1062.

Brooks, H. E., C. A. Doswell III, and L. J. Wicker, 1993: STORMTIPE: A forecasting experiment using a three-dimensional cloud model. *Wea. Forecasting*, **9**, 352-362.

Brooks, H. E., M. Kay, and J. A. Hart, 1998: Objective Limits on Forecasting Skill of Rare Events. *Preprints, 19th Conf. on Severe Local Storms*. Minneapolis, MN. P14.2.

Brooks, H. E., and C. A. Doswell III, 2001: Some aspects of the international climatology of tornadoes by damage classification. *Atm. Res.*, **56**, 191–201.

Brooks, H. E., and C. A. Doswell III, 2002: Deaths in the 3 May 1999 Oklahoma City Tornado from a Historical Perspective. *Wea. Forecasting*, **17**, 354–361.

Brooks, H. E., C. A. Doswell III, and M. P. Kay, 2003: Climatological Estimates of Local Daily Tornado Probability for the United States. *Wea. Forecasting*, **18**, 626-640.

Bryan, G.H., J.C. Wyngaard, and J.M. Fritsch, 2003: Resolution Requirements for the Simulation of Deep Moist Convection. *Mon. Wea. Rev.*, **131**, 2394–2416.

Buizza, R., A. Hollingsworth, F. Lalaurette, and A. Ghelli, 1999: Probabilistic predictions of precipitation using the ECMWF Ensemble Prediction System. *Wea. Forecasting*, **14**, 168–189.

Bunkers, M. J., B. A. Klimowski, J. W. Zeitler, R. L. Thompson, and M. L. Weisman, 2000: Predicting supercell motion using a new hodograph technique. *Wea. Forecasting*, **15**, 61–79.

Carley, J. R., B. R. J. Schwedler, M. E. Baldwin, R. J. Trapp, J. Kwiatkowski, J. Logsdon, and S. J. Weiss, 2011: A proposed model-based methodology for feature-specific prediction for high-impact weather. *Wea. Forecasting*, **26**, 243–249.

Chen, F., and J. Dudhia, 2001: Coupling an advanced land-surface–hydrology model with the Penn State–NCAR MM5 modeling system. Part I: Model description and implementation. *Mon. Wea. Rev.*, **129**, 569–585.

Cintineo, J. L., T. M. Smith, V. Lakshmanan, H. E. Brooks, and K. L. Ortega, 2012: An Objective High-Resolution Hail Climatology of the Contiguous United States. *Wea. Forecasting*, **27**, 1235–1248.

Clark, A.J., W. A. Gallus Jr., M. Xue, and F. Kong, 2009: A Comparison of Precipitation Forecast Skill between Small Convection-Allowing and Large Convection-Parameterizing Ensembles. *Wea. Forecasting*, **24**, 1121–1140.

Clark, A. J., S.J. Weiss, J.S. Kain, I.L. Jirak, M. Coniglio, C.J. Melick, C. Siewert, R. A. Sobash, P.T. Marsh, A.R. Dean, M. Xue, F.Y. Kong, K.W. Thomas, Y.H. Wang, K. Brewster, J.D. Gao, X.G. Wang, J. Du, D.R. Novak, F.E. Barthold, M.J. Bodner, J.J. Levit, C.B. Entwistle, T.L. Jensen, and J. Correia, 2012a: An overview of the 2010 Hazardous Weather Testbed Experimental Forecast Program Spring Experiment. *Bull. Amer. Meteor. Soc*., **93**, 55-74.

Clark, A.J., J.S. Kain, P.T. Marsh, J. Correia, M. Xue and F. Kong, 2012b: Forecasting Tornado Pathlengths Using a Three-Dimensional Object Identification Algorithm Applied to Convection-Allowing Forecasts. *Wea. Forecasting*, **27**, 1090-1113.

Clark, A.J., J. Gao, P. Marsh, T. Smith, J. Kain, J. Correia, M. Xue, and F. Kong, 2013: Tornado pathlength forecasts from 2010 to 2011 using ensemble updraft helicity. *Wea. Forecasting*, **28**, 387–407, doi:10.1175/WAF-D-12-00038.1.

Clark, A. J., I. Jirak, J. Correia, S. J. Weiss, J. Kain, C. Melick, P. Marsh, A. Dean, K. Knopfmeier, G. Carbin, M. Coniglio and B. Twiest, 2015. Spring Forecasting Experiment 2015 Program Overview and Operations Plan. Available online at: https://hwt.nssl.noaa.gov/Spring_2015/HWT_SFE_2015_OPS_plan_final.pdf

Clark, A. J., I. Jirak, C. Melick, J. Correia, S.J. Weiss, J. Kain, A. Dean, K. Knopfmeier, C. Karstens and B. Twiest, 2016. Spring Forecasting Experiment 2016 Program Overview and Operations Plan. Available online at: https://hwt.nssl.noaa.gov/Spring_2016/HWT_SFE2016_operations_plan_final.pdf

Clark, A.J., 2017: Generation of Ensemble Mean Precipitation Forecasts from Convection-Allowing Ensembles. *Wea. Forecasting*, **32**, 1569–1583. doi:10.1175/WAF-D-16-0199.1

Clyne, J., Mininni, P., Norton, A., and Rast, M., 2007: Interactive desktop analysis of high resolution simulations: application to turbulent plume dynamics and current sheet formation, *New Journal of Physics*, **9**, 301.

Colman, B. R., 1990: Thunderstorms above frontal surfaces in environments without positive CAPE. Part I: A climatology. *Mon. Wea. Rev.*, **118**, 1103–1122

Coniglio, M. C., K. L. Elmore, J. S. Kain, S. J. Weiss, M. Xue, and M. L. Weisman, 2010: Evaluation of WRF model output for severe weather forecasting from the 2008 NOAA Hazardous Weather Testbed Spring Experiment. *Wea. Forecasting*, **25**, 408–427.

Coniglio, M. C., D. A. Imy, C. D. Karstens, A. J. Clark, J. Correia Jr., and C. J. Melick, 2014: Evaluation of One-Hour Probabilistic Severe Weather Forecasts Issued during the 2014 NOAA Hazardous Weather Testbed Spring Forecasting Experiment. *27th Conf. on Severe Local Storms*, Madison, WI, Amer. Met. Soc., 47.

Corfidi, S.F., S. J. Weiss, J. S. Kain, S. J. Corfidi, R. M. Rabin, and J. J. Levit, 2010: Revisiting the 3–4 April 1974 Super Outbreak of Tornadoes. *Wea. Forecasting*, **25**, 465–510.

Craven, J.P., and H. E. Brooks, 2004. Baseline climatology of sounding derived parameters associated with deep moist convection. *Nat. Weath. Dig.*, **28**, 13–24.

Cummins, K. L., M. J. Murphy, E. A. Bardo, W. L. Hiscox, R. B. Pyle, and A. E. Pifer, 1998: A combined TOA/MDF technology upgrade of the U.S. National Lightning Detection Network. *J. Geophys. Res.*,**103** (D8), 9035–9044.

Davies, T., M. J. P. Cullen, A. J. Malcolm, M. H. Mawson, A. Staniforth, A. A. White, and N. Wood, 2005: A new dynamical core for the Met Office's global and regional modelling of the atmosphere. *Quart. J. Roy. Meteor. Soc.*, **131**, 1759–1782, doi:10.1256/qj.04.101.

Done, J., C.A. Davis, and M. Weisman, 2004: The next generation of NWP: explicit forecasts of convection using the weather research and forecasting (WRF) model *Atmos. Sci. Let.*, **5**, 110-117.

Doswell, C. A., H. E. Brooks, and R. A. Maddox, 1996: Flash Flood Forecasting: An Ingredients-Based Methodology. *Wea. Forecasting*, **11**, 560-581.

Doswell, C.A. and D.W. Burgess, 1988: On Some Issues of United States Tornado Climatology. *Mon. Wea. Rev.*, **116**, 495–501, https://doi.org/10.1175/1520-0493(1988)116<0495:OSIOUS>2.0.CO;2

Doswell, C. A., H. E. Brooks, and N. Dotzek, 2009: On the implementation of the enhanced Fujita scale in the USA. *Atmos. Res.*, **93**, 554-563.

Du, J., G. DiMego, B. Zhou, D. Jovic, B. Yang, B. Ferrier, G. Manikin, M. Pyle, E. Rogers, Y. Zhu, and S. Benjamin, 2014: NCEP regional ensemble update: current systems and planned storm-scale ensembles. *Preprints*, 26th Conf. on Wea. Forecasting, Atlanta, GA, Amer. Meteor. Soc., J1.4.

Duda, J.D., and W. A. Gallus Jr., 2010: Spring and Summer Midwestern Severe Weather Reports in Supercells Compared to Other Morphologies. *Wea. Forecasting*, **25**, 190–206.

Duda, J. D., X. Wang, F. Kong, and M. Xue, 2014: Using Varied Microphysics to Account for Uncertainty in Warm-Season QPF in a Convection-Allowing Ensemble. *Mon. Wea. Rev.*, **142**, 2198-2219.

Dudhia, J., 1989: Numerical study of convection observed during the Winter Monsoon Experiment using a mesoscale two-dimensional model. *J. Atmos. Sci.*, **46**, 3077–3107.

Ebert, E. E., 2001: Ability of a poor man's ensemble to predict the probability and distribution of precipitation. *Mon. Wea. Rev.*, **129**, 2461-2480.

Elsner and Schmertmann, 1994: Assessing Forecast Skill through Cross Validation. *Wea. Forecasting*, **9**, 619-624.

Evensen, G., 1994: Sequential data assimilation with a nonlinear quasi-geostrophic model using Monte Carlo methods to forecast error statistics. *J. Geophys. Rea.*, **99**, 10143-10162.

Evensen, G., 2003: The Ensemble Kalman Filter: theoretical formulation and practical implementation. *Ocean Dyn.*, **53**, 343-367.

Ferrier, B. S., 1994: A Double-Moment Multiple-Phase Four-Class Bulk Ice Scheme. Part I: Description. *J. Atmos. Sci.*, **51**, 249-280.

Finley, J. P., 1884: Tornado predictions. *Amer. Meteor. J.*, **1**, 85-88.

Fowle, M. and P. J. Roebber, 2003: Short-Range (0–48 h) Numerical Prediction of Convective Occurrence, Mode, and Location. *Wea. Forecasting*, **18**, 782–794.

Gagne, D.J., A. McGovern, S.E. Haupt, R.A. Sobash, J.K. Williams, and M. Xue, 2017: Storm-Based Probabilistic Hail Forecasting with Machine Learning Applied to Convection-Allowing Ensembles. *Wea. Forecasting*, **32**, 1819–1840.

Gallo, B. T., A. J. Clark, and S. R. Dembek, 2016: Forecasting Tornadoes Using Convection-Permitting Ensembles. *Wea. Forecasting*, **31**, 273–295. doi: http://dx.doi.org/10.1175/WAF-D-15-0134.1.

Gallo, B.T., and Coauthors, 2017a: Breaking New Ground in Severe Weather Prediction: The 2015 NOAA/Hazardous Weather Testbed Spring Forecasting Experiment. *Wea. Forecasting*, **32**, 1541–1568, doi: https://doi.org/10.1175/WAF-D-16-0178.1

Gallo, B. T., A. J. Clark, B. T. Smith, R. L. Thompson, I. Jirak, and S. R. Dembek, 2017b: Blended Probabilistic Tornado Forecasts: Combining Climatological Frequencies with NSSL-WRF Ensemble Forecasts. *Wea. Forecasting*, in review.

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.

Grünwald, S., and H. E. Brooks, 2011: Relationship between sounding derived parameters and the strength of tornadoes in Europe and the USA from reanalysis data. *Atmos. Res.*, **100**, 479–488.

Hamill, T.M., and S.J. Colucci, 1998: Evaluation of Eta-RSM Ensemble Probabilistic Precipitation Forecasts. *Mon. Wea. Rev*., **126**, 711-724.

Hamill, T.M., 1999: Hypothesis Tests for Evaluating Numerical Precipitation Forecasts. *Wea. Forecasting*, **14**, 155-167.

Hart, J. A., and W. D. Korotky, 1991: The SHARP workstation user's manual—v1.50. A skewt/hodograph analysis and research program for the IBM and compatible PC. NOAA/NWS Forecast Office, Charleston, WV, 62 pp.

Hitchens, N. M., and H. E. Brooks, 2012: Evaluation of the Storm Prediction Center's Day 1 Convective Outlooks. *Wea. Forecasting*, **27**, 1580-1585.

Hitchens, N. M., H. E. Brooks, and M. P. Kay, 2013: Objective Limits on Forecasting Skill of Rare Events. *Wea. Forecasting*, **28**, 525-534.

Hong, S.-Y., J. Dudhia, and S. H. Chen, 2004: A revised approach to ice microphysical processes for the bulk parameterization of cloud and precipitations. *Mon. Wea. Rev.*, **132**, 103-120. doi:10.1175/1520-0493(2004)132<0103:ARATIM>2.0.CO;2.

Hong, S.-Y., and J.-O. J. Lim, 2006: The WRF single-moment 6-class microphysics scheme (WSM6). *J. Korean Meteor. Soc*., **42**, 129–151.

Hong, S.-Y., S. Y. Noh, and J. Dudhia, 2006: A new vertical diffusion package with an explicit treatment of entrainment processes. *Mon. Wea. Rev.*, **134**, 2318-2341. doi: 10.1175/MWR3199.1.

Hu, M., M. Xue, and K. Brewster, 2006: 3DVAR and cloud analysis with WSR-88D level-II data for the prediction of Fort Worth tornadic thunderstorms. Part I: Cloud analysis and its impact. *Mon. Wea. Rev.*, **134**, 675-698.

Iacono, M. J., J. S. Delamere, E. J. Mlawer, M. W. Shephard, S. A. Clough, and W. D. Collins, 2008: Radiative forcing by long-lived greenhouse gases: Calculations with the AER radiative transfer models. *J. Geophys. Res.*, **113**, D13103. doi:10.1029/ 2008JD009944.

Janjić, Z. I., 1994: The step-mountain eta coordinate model: Further developments of the convection, viscous sublayer, and turbulence closure schemes. *Mon. Wea. Rev.*, **122**, 927–945, doi:10.1175/1520-0493(1994)122,0927:TSMECM.2.0.CO;2.

Janjić, Z. I., 2002: Nonsingular implementation of the Mellor-Yamada level 2.5 scheme in the NCEP Meso model. NCEP Office Note 437, NOAA/NWS, 61 pp.

Janjić, Z. I., and R. Gall, 2012: Scientific documentation of the NCEP Nonhydrostatic Multiscale Model on the B Grid (NMMB). Part 1: Dynamics. NCAR/TN-489+STR, 75 pp. [Available online at http://nldr.library.ucar.edu/repository/assets/technotes/TECH-NOTE-000-000-000-857.pdf.]

Jewell, R., and J. Brimelow, 2009: Evaluation of Alberta hail growth model using severe hail proximity soundings from the United States. *Wea. Forecasting*, **24**, 1592-1609.

Jirak, I.L., S. J. Weiss, and C. J. Melick, 2012a: The SPC Storm-scale Ensemble of Opportunity: Overview and Results from the 2012 Hazardous Weather Testbed Spring Forecasting Experiment. *Preprints*, 26th Conf. Severe Local Storms, Nashville, TN.

Jirak, I. L., C. J. Melick, A. R. Dean, S. J. Weiss, and J. Correia, Jr., 2012b: Investigation of an automated temporal disaggregation technique for convective outlooks during the 2012 Hazardous Weather Testbed Spring Forecasting Experiment. Preprints, *26th Conf. on Severe Local Storms*, Nashville, TN.

Jirak, I.L., C.J. Melick, and S.J. Weiss, 2014: Combining Probabilistic Ensemble Information from the Environment with Simulated Storm Attributes to Generate Calibrated Probabilities of Severe Weather Hazards. *Preprints*, 27th Conf. Severe Local Storms, Madison, WI.

Johnson, A., X. Wang, F. Kong, and M. Xue, 2013: Object-Based Evaluation of the Impact of Horizontal Grid Spacing on Convection-Allowing Forecasts. *Mon. Wea. Rev.*, **141**, 3413-3425.

Kain, J. S., P. R. Janish, S. J. Weiss, M. E. Baldwin, R. S. Schneider, and H. E. Brooks, 2003: Collaboration between forecasters and research scientists at the NSSL and SPC: The spring program. *Bull. Amer. Meteor. Soc.*, **84**, 1797-1806.

Kain, J. S., S. J. Weiss, D. R. Bright, M. E. Baldwin, J. J. Levit, G. W. Carbin, C. S. Schwartz, M. L. Weisman, K. K. Droegemeier, D. B. Weber, and K. W. Thomas, 2008: Some Practical Considerations Regarding Horizontal Resolution in the First

Generation of Operational Convection-Allowing NWP. *Wea. Forecasting*, **23**, 931–952.

Kain, J.S., S. R. Dembek, S. J. Weiss, J. L. Case, J. J. Levit, and R. A. Sobash, 2010: Extracting Unique Information from High-Resolution Forecast Models: Monitoring Selected Fields and Phenomena Every Time Step. *Wea. Forecasting*, **25**, 1536–1542.

Kain, J. S., S. Willington, A. J. Clark, S. J. Weiss, M. Weeks, I. L. Jirak, M. C. Coniglio, N. M. Roberts, C. D. Karstens, J. M. Wilkinson, K. H. Knopfmeier, H. W. Lean, L. Ellam, K. Hanley, R. North, and D. Suri, 2016: Collaborative Efforts between the U.S. and U. K. to Advance Prediction of High Impact Weather. *Bull. Amer. Meteor. Soc.*, **98**, 937-948.

Kalnay, E., 2003: Atmospheric Modeling, Data Assimilation and Predictability. Cambridge University Press, 364 pp.

Karstens, C. D., T. M. Smith, K. M. Kuhlman, A. J. Clark, C. Ling, G. J. Sutmpf, and L. P Rothfusz, 2014: Prototype Tool Development for Creating Probabilistic Hazard Information for Severe Convective Phenomena. *Second Symposium on Building a Weather-Ready Nation*. Atlanta, GA, Amer. Met. Soc., 2.2.

Karstens, C. D., G. J. Stumpf, C. Ling, L. Hua, D. Kingfield, T. M. Smith, J. Correia, Jr., K. M. Calhoun, K. L. Ortega, C. J. Melick, and L. P. Rothfusz, 2015: Evaluation of a probabilistic forecasting methodology for severe convective weather in the 2014 Hazardous Weather Testbed. *Wea. Forecasting*, **30,** 1551-1570.

Klees, A. M., Y. P. Richardson, P. M. Markowski, C. Weiss, J. M. Wurman, and K. K. Kosiba 2016: Comparison of the Tornadic and Nontornadic Supercells Intercepted by VORTEX2 on 10 June 2010. *Mon. Wea. Rev.*, **144**, 3201-3231.

Kong, F., M. Xue, Y. Jung, K. Brewster, K. Thomas, Y. Wang, F. Shen, I. Jirak, A. Clark, J. Correia Jr., S. Weiss, M. Coniglio, and C. J. Melick, 2015: An Overview of CAPS Storm-Scale Ensemble Forecast for the 2015 NOAA HWT Spring Forecasting Experiment. *27th Conf. Wea. Anal. Forecasting/23rd Conf. Num. Wea. Pred.*, Chicago, IL, AMS, Paper 32.

Kuchera, E., S.Rentschler, G. Creighton, and J. Hamilton, 2014: The Air Force weather ensemble prediction suite. 15th Annual WRF Users' Workshop, Boulder CO. Website:http://www2.mmm.ucar.edu/wrf/users/workshop s/WS2014/ppts/2.3.pdf.

Kumar, S. V., and Coauthors, 2006: Land Information System – An interoperable framework for high resolution land surface modeling. *Environ. Modell. Software*, **21**, 1402-1415.

Kumar, S.V., C. D. Peters-Lidard, J. L. Eastman, and W.-K. Tao, 2007: An integrated high-resolution hydrometeorological modeling testbed using LIS and WRF. *Environ. Modell. Software*, **23**, 169-181.

Lim, K.-S. S. and S.-Y. Hong, 2010: Development of an effective double-moment cloud microphysics scheme with prognostic Cloud Condensation Nuclei (CCN) for weather and climate models. *Mon. Wea. Rev.*, **138**, 1587-1612.

Line, W. E., T. J. Schmit, D. T. Lindsey, and S. J. Goodman, 2016: Use of geostationary super rapid scan satellite imagery by the Storm Prediction Center. *Wea. Forecasting*, **31**, 483–494. doi: http://dx.doi.org/10.1175/WAF-D-15-0135.1

Loken, E. D., A. J. Clark, M. Xue, and F. Kong, 2017: Comparison of Next-Day Probabilistic Severe Weather Forecasts from Coarse- and Fine-Resolution CAMs and a Convection-Allowing Ensemble. *Wea. Forecasting*, **32**, 1403-1421.

Markowski, P. M., E. N. Rasmussen, and J. M. Straka, 1998: The occurrence of tornadoes in supercells interacting with boundaries during VORTEX-95. *Wea. Forecasting*, **13**, 852–859.

Mason, I., 1982: A model for assessment of weather forecasts. *Aust. Meteorol. Mag.*, **30**, 291–303.

McBeath, K., P. R. Field, and R. J. Cotton, 2014: Using operational weather radar to assess high-resolution numerical weather prediction over the British Isles for a cold air outbreak case-study. *Q. Jour. Royal Met. Soc.*, **140**, 225-239.

Melick, C.J., I.L. Jirak, J. Correia Jr., A.R. Dean, and S.J. Weiss, 2014: Exploration of the NSSL Maximum Expected Size of Hail (MESH) Product for Verifying Experimental Hail Forecasts in the 2014 Spring Forecasting Experiment. Preprints, *27th Conf. Severe Local Storms*, Madison, WI.

Mellor, G. L., and T. Yamada, 1982: Development of a turbulence closure model for geophysical fluid problems. *Rev. Geophys.*, **20**, 851–875.

Menzel, W. P., 2001: Cloud Tracking with Satellite Imagery: From the Pioneering Work of Ted Fujita to the Present. *Bull. Amer. Meteor. Soc.*, **82**, 33–47. doi: http://dx.doi.org/10.1175/1520-0477(2001)082<0033:CTWSIF>2.3.CO;2

Mercer, A.E., C. M. Shafer, C. A. Doswell III, L. M. Leslie, and M. B. Richman, 2012: Synoptic Composites of Tornadic and Nontornadic Outbreaks. *Mon. Wea. Rev.*, **140**, 2590–2608.

Milbrandt, J. A. and M. K. Yau, 2005: A Multimoment Bulk Microphysics Parameterization. Part I: Analysis of the Role of the Spectral Shape Parameter. *J. Atmos. Sci.*, **62**, 3051-3064.

Miller, P. A., M. F. Barth, and L. A. Benjamin, 2005: An update on MADIS observation ingest, integration, quality control and distribution capabilities. *Preprints, 21st Int. Conf. on Interactive Information and Processing Systems,* San Diego, CA, Amer. Meteor. Soc., J7.12. [Available online at https://ams.confex.com/ams/pdfpapers/86703.pdf.]

Miller, P. A., M. F. Barth, L. A. Benjamin, R. S. Artz, and W. R. Pendergrass, 2007: MADIS support for UrbaNet. *Preprints, 14th Symp. on Meteorological Observation and Instrumentation/16th Conf. on Applied Climatology*, San Antonio, TX, Amer. Meteor. Soc., JP2.5. [Available online at http://ams.confex.com/ams/pdfpapers/119116.pdf.]

Mittermaier, M. P., 2014: A Strategy for Verifying Near-Convection-Resolving Model Forecasts at Observing Sites. *Wea. Forecasting*, **29**, 185-204.

Mlawer, E. J., S. J. Taubman, P. D. Brown, M. J. Iacono, and S. A. Clough, 1997: Radiative transfer for inhomogeneous atmosphere: RRTM, a validated correlated-k model for the long-wave. *J. Geophys. Res.*, **102** (D14), 16 663–16 682.

Morrison, H., and J. O. Pinto, 2005: Mesoscale modeling of springtime Arctic mixed-phase stratiform clouds using a new two-moment bulk microphysics scheme. *J. Atmos. Sci.*, **62**, 3683–3704.

Morrison, H., and J. O. Pinto, 2006: Intercomparison of bulk microphysics scheme in mesoscale simulations of springtime Arctic mixed phase stratiform clouds. *Mon. Wea. Rev.*, **134**, 1880–1900.

Morrison, H., and J. A. Milbrandt, 2015: Parameterization of Cloud Microphysics Based on the Prediction of Bulk Ice Particle Properties. Part I: Scheme Description and Idealized Tests. *J. Atmos. Sci.*, **72**, 287–311. doi: http://dx.doi.org/10.1175/JAS-D-14-0065.1

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

Nakanishi, M., and H. Niino, 2004: An improved Mellor-Yamada level-3 model with condensation physics: Its design and verification. *Bound.-Layer Meteor.*, **112**, 1-31. doi: 10.1023/B:BOUN.0000020164.04146.98.

Nakanishi, M., and H. Niino, 2006: An improved Mellor-Yamada level-3 model: Its numerical stability and application to a regional prediction of advection fog. *Bound.-Layer Meteor.*, **119**, 397-407, doi: 10.1007/s10546-005-9030-8.

Naylor, J., M. S. Gilmore, R. L. Thompson, R. Edwards, and R. B. Wilhelmson, 2012: Comparison of objective supercell identification techniques using an idealized cloud model. *Mon. Wea. Rev.*, **140**, 2090–2102, doi:10.1175/MWR-D-11-00209.1.

Orf, L., R. Wilhelmson, B. Lee, C. Finley, and A. Houston, 2017: Evolution of a Long-Track Violent Tornado within a Simulated Supercell. *Bull. Amer. Meteor. Soc.*, **98**, 45–68.

Ortega, K.L, T. M. Smith, K. L. Manross, K. A. Scharfenberg, A. Witt, A. G. Kolodziej and J. J. Gourley, 2009: The Severe Hazards Analysis and Verification Experiment. *Bull. Amer. Meteor. Soc.*, **90**, 1519-1530.

Rasmussen, E.N. and D. O. Blanchard, 1998: A Baseline Climatology of Sounding-Derived Supercell and Tornado Forecast Parameters. *Wea. Forecasting*, **13**, 1148–1164.

Rasmussen, E. N., S. Richardson, J. M. Straka, P. M. Markowski, and D. O. Blanchard, 2000: The association of significant tornadoes with a baroclinic boundary on 2 June 1995. *Mon. Wea. Rev.*, **128**, 174–191.

Rasmussen, E.N., 2003: Refined Supercell and Tornado Forecast Parameters. *Wea. Forecasting*, **18**, 530–535.

Roebber, P.J., 2009: Visualizing Multiple Measures of Forecast Quality. *Wea. Forecasting*, **24**, 601–608. doi: http://dx.doi.org/10.1175/2008WAF2222159.1.

Rothfusz, L., C. D. Karstens, and D. Hilderbrand, 2014: Forecasting a Continuum of Environmental Threats: Exploring Next-Generation Forecasting of High Impact Weather, EOS Transactions, American Geophysical Union, **95,** 325-326.

Satoh, M., T. Masuno, H. Tomita, H. Miura, T. Nasuno, and S. Iga, 2008: Nonhydrostatic icosahedaral atmospheric model (NICAM) for global cloud resolving simulations. *J. Comput. Phys.*, **227**, 3486–3514.

Schaefer, J. T., 1990: The critical success index as an indicator of warning skill. *Wea. Forecasting*, **5**, 570-575.

Schwartz, C. S., G. S. Romine, K. R. Smith, and M. L. Weisman, 2014: Characterizing and Optimizing Precipitation Forecasts from a Convection-Permitting Ensemble Initialized by a Mesoscale Ensemble Kalman Filter. *Wea. Forecasting*, **29**, 1295-1318.

Schwartz, C., G. Romine, M. Weisman, R. Sobash, K. Fossell, K. Manning, and S. Trier, 2015a: A real-time convection-allowing ensemble prediction system initialized by mesoscale ensemble Kalman filter analyses. *Wea. Forecasting*, **30**, 1158-1181. doi:10.1175/WAF-D-15-0013.1.

Schwartz, C. S., G. S. Romine, R. A. Sobash, K. R. Fossell, and M. L. Weisman, 2015b: NCAR's Experimental Real-Time Convection-Allowing Ensemble Prediction System. *Wea. Forecasting*, **30**, 1645-1654.

Simmons, K. M., and D. Sutter, 2011: *Economic and Societal Impacts of Tornadoes.* American Meteorological Society, 282 pp

Simmons, K., D. Sutter, and R. Pielke, 2013: Normalized tornado damage in the United States: 1950–2011. *Environ. Hazards*, **12**, 132–147.

Skamarock, W. C., and Coauthors, 2008: A description of the Advanced Research WRF Version 2, NCAR Tech Note, NCAR/TN-475+STR, 113 pp. [Available at: http://www.mmm.ucar.edu/wrf/users/docs/arw_v3.pdf.]

Skamarock, W. C., Klemp, J. B., Duda, M., Fowler, L. D., Park, S.-H., and T. Ringler, 2012: A multiscale nonhydrostatic atmospheric model using centroidal Voronoi tessellations and c-grid staggering. *Mon. Wea. Rev*, **140**, 3090–3105.

Smagorinsky, J., 1963: General circulation experiments with the primitive equations. *Mon. Weather Rev.*, **91**, 99–164.

Smirnova, T.G., J.M. Brown, and S.G. Benjamin, 1997: Performance of different soil model configurations in simulating ground surface temperature and surface fluxes. *Mon. Wea. Rev.*, **125**, 1870-1884, http://dx.doi.org/10.1175/15200493(1997)1252.0.CO;2.

Smirnova, T.G., J.M. Brown, and D. Kim, 2000: Parameterization of cold-season processes in the 551 MAPS land-surface scheme. *J. Geophys. Res.*, **105**, 4077-4086, doi: 10.1029/1999JD901047.

Smirnova, T., J. Brown, S. Benjamin, and J. Kenyon, 2016: Modifications to the Rapid Update Cycle Land Surface Model (RUC LSM) available in the Weather Research and Forecast (WRF) model. *Mon. Wea. Rev.*, **144**, 1851-1865. doi:10.1175/MWR-D-15-0198.1.

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.

Smith, B. T., R. L. Thompson, A. R. Dean, and P. T. Marsh, 2015: Diagnosing the conditional probability of tornado damage rating using environmental and radar attributes. *Wea. Forecasting*, **30**, 914–932, doi:10.1175/WAF-D-14-00122.1.

Smith, T. M., J. Gao, K. M. Calhoun, D. J. Stensrud, K. L. Manross, K. L. Ortega, C. Fu, D. M. Kingfield, K. L. Elmore, V. Lakshmanan, and C. Riedel, 2014: Examination of a Real-Time 3DVAR Analysis System in the Hazardous Weather Testbed. *Wea. Forecasting*, **29**, 63-77.

Sobash, R. A., J. S. Kain, D. R. Bright, A. R. Dean, M. C. Coniglio, and S. J. Weiss, 2011: Probabilistic forecast guidance for severe thunderstorms based on

the identification of extreme phenomena in convection-allowing model forecasts. *Wea. Forecasting*, **26**, 714–728.

Sobash, R. A., C. S. Schwartz, G. S. Romine, K. R. Fossell, and M. L. Weisman, 2016a: Severe weather prediction using storm surrogates from an ensemble forecasting system. *Wea. Forecasting*, **31**, 255–271.

Sobash, R.A., G. S. Romine, C. S. Schwartz, D. J. Gagne II, and M. L. Weisman, 2016b: Explicit Forecasts of Low-Level Rotation from Convection-Allowing Models for Next-Day Tornado Prediction. *Wea. Forecasting*, **31**, 1591-1614.

Sobash, R.A. and J.S. Kain, 2017: Seasonal Variations in Severe Weather Forecast Skill in an Experimental Convection-Allowing Model. *Wea. Forecasting*, **32**, 1885–1902, https://doi.org/10.1175/WAF-D-17-0043.1.

Steiner, J. T., 1973: A three-dimensional model of cumulus cloud development. *J. Atmos. Sci.*, **30**, 414-434.

Stensrud, D. J., and Coauthors, 2009: Convective-scale warn-on-forecast system: A vision for 2020. *Bull. Amer. Meteor. Soc.*, **90**, 1487–1499.

Surcel, M., I. Zawadzki, and M. K. Yau, 2014: On the Filtering Properties of Ensemble Averaging for Storm-Scale Precipitation Forecasts. *Mon. Wea. Rev.*, **142**, 1093-1105.

Thompson, G., R. M. Rasmussen, and K. Manning, 2004b: Explicit forecasts of winter precipitation using an improved bulk microphysics scheme. Part I: Description and sensitivity analysis. *Mon. Wea. Rev.*, **132**, 519–542.

Thompson, G., P. R. Field, R. M. Rasmussen, and W. D. Hall, 2008: Explicit forecasts of winter precipitation using an improved bulk microphysics scheme. Part II: Implementation of a new snow parameterization. *Mon. Wea. Rev.*, **136**, 5095–5115, doi:10.1175/2008MWR2387.1.

Thompson, G., and T. Eidhammer, 2014: A study of aerosol impacts on clouds and precipitation development in a large winter cyclone. *J. Atmos. Sci.*, **71**, 3636–3658, doi:10.1175/ JAS-D-13-0305.1.

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.

Thompson, R. L., R. Edwards, and C. M. Mead, 2004a: An update to the supercell composite and significant tornado parameters. *Preprints*, 22nd Conf. on Severe Local Storms, Hyannis, MA, Amer. Meteor. Soc., P8.1.

Thompson, R. L., C. M. Mead, and R. Edwards, 2007: Effective storm-relative helicity and bulk shear in supercell thunderstorm environments. *Wea. Forecasting*, **22**, 102–115.

Thompson, R.L., 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.

Thompson, R. L., B. T. Smith, A. R. Dean, and P. T. Marsh, 2013: Spatial distributions of tornadic near-storm environments by convective mode. *Electronic J. Severe Storms Meteor.*, **8** (5), 1–22.

Thompson, R. L., and Coauthors, 2017: Tornado damage rating probabilities derived from WSR-88D data. *Wea. Forecasting*, **32**, 1509-1528.

Towns, J., T. Cockerill, M. Dahan, I. Foster, K. Gaither, A. Grimshaw, V. Hazlewood, S. Lathrop, D. Lifka, G. D. Peterson, R. Roskies, J. R. Scott, N. Wilkins-Diehr, 2014: XSEDE: Accelerating Scientific Discovery, *Comput. Sci. Eng.*, **16(5)**, 62-74, doi:10.1109/MCSE.2014.80

Trapp, R. J., G. J. Stumpf, and K. L. Manross, 2005: A reassessment of the percentage of tornadic mesocyclones. *Wea. Forecasting*, **20**, 680–687.

Verbout, S.M., H. E. Brooks, L. M. Leslie, and D. M. Schultz, 2006: Evolution of the U.S. Tornado Database: 1954–2003. *Wea. Forecasting*, **21**, 86–93.

Wandishin, M. S., S. L. Mullen, D. J. Stensrud, and H. E. Brooks, 2001: Evaluation of a short-range multimodel ensemble system. *Mon. Wea. Rev.*, **129**, 729–747.

Wang, Y., Y. Jung, T. A. Supinie, and M. Xue, 2013: A hybrid MPI/OpenMP parallel algorithm and performance analysis for an ensemble square root filter suitable for dense observations. *J. Atmos. Ocean. Tech.*, **30**, 1382-1397.

Weisman, M. L. and J. B. Klemp. 1982. The dependence of numerically simulated convective storms on vertical wind shear and buoyancy. *Mon. Wea. Rev.*, **110**, 504–520.

Weisman, M. L., and J. B. Klemp, 1984: The structure and classification of numerically simulated convective storms in directionally varying wind shears. *Mon. Wea. Rev.*, **112**, 2479–2498.

Weisman, M. L., and J. B. Klemp, 1986: Characteristics of isolated convective storms. *Mesoscale Meteorology and Forecasting*, P. S. Ray, Ed., Amer. Meteor. Soc., 331–358.

Weisman, M. L., W. C. Skamarock, J. B. Klemp, 1997: The resolution dependence of explicitly modeled convective systems. *Mon. Wea. Rev.*, **125**, 527–548.

Weisman, M. L., and R. Rotunno, 2000: The use of vertical wind shear versus helicity in interpreting supercell dynamics. *J. Atmos. Sci.*, **57**, 1452–1472.

Weisman, M. L., C. Davis, W. Wang, K. W. Manning, and J. B. Klemp, 2008: Experiences with 0–36-h Explicit Convective Forecasts with the WRF-ARW Model. *Wea. Forecasting*, **23**, 407–437.

Weiss, S.J., D.R. Bright, J.S. Kain, J.J. Levit, M.E. Pyle, Z.I. Janjic, B.S. Ferrier, and J. Du, 2006: Complementary Use of Short-range Ensemble and 4.5 KM WRF-NMM Model Guidance for Severe Weather Forecasting at the Storm Prediction Center. *Preprints*, 23rd Conf. Severe Local Storms, St. Louis MO.

Wicker, L. J., M. P. Kay, and M. P. Foster, 1997: STORMTIPE-95: Results from a convective storm forecast experiment. *Wea. Forecasting*, **12**, 388–398.

Wilks, D. S., 2011: *Statistical Methods in the Atmospheric Sciences: An Introduction*. Academic Press, 676 pp.

Wilson, D.R. and Ballard, S.P., 1999. A microphysically based precipitation scheme for the UK Meteorological Office Unified Model. *Q.J.R. Meteorol. Soc.*, **125**, 1607-1636.

Wilson, C. J., K. L. Ortega, and V. Lakshmanan, 2009: Evaluating multi-radar, multi-sensor hail diagnosis with high resolution hail reports. *Preprints, 25th Conf. on Interactive Information Processing Systems*, Phoenix, AZ, Amer. Meteor. Soc., P2.9. [Available online at http://ams.confex.com/ams/pdfpapers/ 146206.pdf.

Witt, A., M. D. Eilts, G. J. Stumpf, J. T. Johnson, E. D. Mitchell, and K. W. Thomas, 1998: An enhanced hail detection algorithm for the WSR-88D. *Wea. Forecasting*, **13,** 286-303.

Xue, M., D.-H. Wang, J.-D. Gao, K. Brewster, and K. K. Droegemeier, 2003: The Advanced Regional Prediction System (ARPS), storm-scale numerical weather prediction and data assimilation. *Meteor. Atmos. Phy.*, **82**, 139-170.

Xue, M., M. Hu, and M. Tong, 2006: Assimilation of radar data and short-range prediction of thunderstorms using 3DVAR, cloud analysis and ensemble Kalman filter methods. *12th Conf. Aviation Range Aerospace Meteor.*, 2006, Atlanta, Georgia, Amer. Meteor. Soc.

Xue, M., F. Kong, D. Weber, K. W. Thomas, Y. Wang, K. Brewster, K. K. Droegemeier, J. S. K. S. J. Weiss, D. R. Bright, M. S. Wandishin, M. C. Coniglio, and J. Du, 2007: CAPS realtime storm-scale ensemble and high-resolution forecasts as part of the NOAA Hazardous Weather Testbed 2007 spring experiment. *22nd Conf. Wea. Anal. Forecasting/18th Conf. Num. Wea. Pred.*, Amer. Meteor. Soc., CDROM 3B.1.