UNIVERSITY OF OKLAHOMA

GRADUATE COLLEGE

RESOLUTION AND PHYSICS SENSITIVITIES IN CONVECTION-ALLOWING MODELS AND ENSEMBLES

A THESIS

SUBMITTED TO THE GRADUATE FACULTY

in partial fulfillment of the requirements for the

Degree of

MASTER OF SCIENCE IN METEOROLOGY

By

ERIC D. LOKEN
Norman, Oklahoma
2017

RESOLUTION AND PHYSICS SENSITIVITIES IN CONVECTION-ALLOWING MODELS AND ENSEMBLES

A THESIS APPROVED FOR THE
SCHOOL OF METEOROLOGY

BY

_____
Dr. Adam Clark, Chair

_____
Dr. Harold Brooks

_____
Dr. Ming Xue

_____
Dr. Jason Furtado

Acknowledgments

An old proverb declares: “it takes a village to raise a child.” I feel the same can be said for producing a master’s thesis. This work would not have been possible without help from so many others who have guided me along the way.

First and foremost, I thank my advisor, Dr. Adam Clark, for his excellent mentorship and guidance. Every time I wanted to discuss an aspect of this work, Adam made time to see me instantly, despite his busy schedule. His friendly, enthusiastic demeanor and constant encouragement (even in the face of my mistakes) filled me with excitement and inspiration and instilled in me a passion for research. To say that I look up to him is an understatement, for he is more than a model scientist—he is a model human being.

Second, I thank Drs. Harold Brooks, Ming Xue, and Jason Furtado for their guidance, feedback, and service on my master’s committee. Like Adam, Harold, Ming, and Jason all graciously took time out of their busy schedules to talk with me about my work, providing useful input and advice that has improved this thesis.

Third, I thank my parents, Rhonda and Richard Loken, who have always been there for me, offering their support, encouragement, and unconditional love. When I demonstrated an interest in the natural world at a young age, they actively fostered that interest by taking me to museums, planetariums, and exhibits. Throughout my entire life, they have been my advocate and my rock. For these reasons, I feel that this thesis belongs as much to them as it does to me.

Fourth, I thank my co-workers and officemates in Norman, who have extended me their friendship and support and who have given me invaluable feedback on my work. In particular, I thank Jon Labriola and Elizabeth Smith for constantly going out of

their way to offer helpful advice and feedback, and I thank Tyler Bell, whose technical support and coding expertise benefited me a great deal.

Last but not least, I thank my past teachers, coaches, and mentors; this thesis is just one byproduct of their instruction and guidance. I especially thank: Mrs. Chris Cook, my first grade teacher, whose meteorology unit ignited my lifelong passion for meteorology; Mrs. Kit Rittman, my seventh grade science teacher, who went out of her way to nourish my passion for meteorology; Mr. Brady Nichols—my high school cross country coach—and Dr. Michael Leckrone—my undergraduate marching band director—who both taught me work ethic and discipline; and Dr. Jonathan Martin, my undergraduate synoptic meteorology professor, who constantly inspired me with his passion for the atmospheric sciences, love of teaching, and command of the English language.

This work was made possible by a Presidential Early Career Award for Scientists and Engineers (PECASE). Additional support was provided by NOAA/Office of Oceanic and Atmospheric Research under NOAA-University of Oklahoma Cooperative Agreement NA11OAR4320072, U.S. Department of Commerce. In addition to NOAA CSTAR support, CAPS simulations received supplementary support from NSF Grant ATM-0802888.

Table of Contents

List of Tables

List of Figures

Abstract

In recent years, convection-allowing models (CAMs) and ensembles have become more prominent in both research and operational settings. However, it largely remains unclear how to leverage computing resources to maximize forecast quality and value at convection-allowing resolution. In this thesis, two research components are designed to address questions regarding convection-allowing ensemble design and the optimal use of computing resources.

The first component uses data from the 2010 and 2011 NOAA Hazardous Weather Testbed Spring Forecasting Experiments (HWT SFEs) to compare next-day probabilistic severe weather forecasts derived from simulated updraft helicity (UH) from three Advanced Research Weather Research and Forecasting (ARW-WRF) model configurations: a 4-km deterministic CAM; an equivalently-configured 1-km deterministic CAM; and an 11-member, 4-km convection-allowing ensemble. Results from this component suggest that creating a convection-allowing ensemble at relatively coarse grid-spacing may be a better use of computing resources than reducing the grid-spacing of a deterministic CAM.

The second research component uses data from the 2016 Community Leveraged Unified Ensemble (CLUE), which was assembled during the 2016 HWT SFE, to compare the spread and skill of mixed- and single-physics convection-allowing ensemble forecasts of 2-m temperature, 2-m dewpoint temperature, 500-mb geopotential height, and hourly accumulated precipitation at a variety of spatial scales. Up to 36-hour forecasts are analyzed. Results from this component indicate that, although the mixed-physics ensemble tends to produce forecasts with greater spread and slightly greater skill, the differences between the two ensemble forecasts are generally

small, especially at larger spatial scales and when the ensembles are well-calibrated. Model developers may therefore wish to consider implementing a single- rather than a mixed-physics ensemble operationally, given the similar performance but smaller maintenance costs of the single-physics ensemble.

Chapter 1: General Introduction

1. Introduction

Recent advances in computing power have led to the implementation of high-resolution numerical weather prediction (NWP) models capable of explicitly simulating convection without a convective parameterization scheme. These models, known as "convection-allowing models" (CAMs), are frequently used in the arena of severe storm forecasting, where they have been found to give forecasters useful information regarding storm mode, initiation, and evolution (Kain et al. 2006; Done et al. 2004).

Since 2004, the NOAA Hazardous Weather Testbed Spring Forecasting Experiment (HWT SFE; formerly called the NOAA/NSSL Spring Program) has provided a venue for researchers and forecasters to discuss and test the utility of CAMs and CAM applications. For example, HWT SFEs have explored the impact of horizontal grid spacing on model skill and value (e.g., Kain et al. 2008; Schwartz et al. 2009), investigated the use of simulated updraft helicity as a severe weather proxy (Kain et al. 2008, Sobash et al. 2011), and evaluated the overall usefulness of CAMs and convection-allowing ensembles to severe weather forecasters (e.g., Kain et al. 2006, Coniglio et al. 2010). These HWT SFEs—as well as studies from outside the annual HWT SFEs (e.g., Done et al. 2004, Schwartz et al. 2015b, Schwartz et al. 2017, etc.)—have collectively demonstrated that CAMs and convection-allowing ensembles provide skillful and useful forecast guidance.

Nevertheless, questions about how to optimally design convection-allowing models and ensembles (given the limitations of current computational resources) remain. For example: How much forecast quality and value is gained by further

decreasing the grid spacing of a deterministic CAM beyond 4 km (e.g., Kain et al. 2008; Schwartz et al. 2009, Schwartz et al. 2017)? How do relatively-coarse resolution CAM ensemble forecasts compare to relatively fine-resolution deterministic CAM forecasts? What is the optimal configuration of a CAM ensemble (Roebber et al. 2004, Duda et al. 2014, Johnson and Wang 2017), and how do ensemble specifications (e.g., the presence or absence of multiple microphysics schemes within an ensemble) impact ensemble spread and skill?

The purpose of this thesis is to utilize datasets from the 2010, 2011, and 2016 NOAA Hazardous Weather Testbed Spring Experiments (HWT SFEs) to determine how best to leverage computational resources and to deduce how CAM and CAM ensembles may be configured for optimal use in short-term (i.e., up to next-day) weather forecasting.

2. Research background

a) A brief history of electronic NWP: From ENIAC- to CAM-derived forecasts

Electronic NWP dates to the early 1950s, when Charney et al. (1950) integrated the barotropic vorticity equations forward in time using the Electronic Numerical Integrator and Computer (ENIAC). These early forecasts were slow (a 24-hour forecast took approximately 24 hours to produce) and operated at relatively coarse spatial and temporal resolutions: they used 736-km horizontal grid spacing and time steps of up to 3 hours (Charney et al. 1950). Moreover, strictly speaking, the "forecasts" were hindcasts, since they were made for past atmospheric states. However, rapid increases in computing power allowed the Swedish Military Weather Service to implement the first

real-time forecasts in Stockholm in 1954 (e.g., Bolin 1955; Bergthorsson et al. 1955).

As computing power increased further, forecasts were made with shorter time steps, increasingly sophisticated physical parameterizations, and finer horizontal and vertical resolution (Bushby 1986). By the mid-1960s, NWP became capable of analyzing fields other than pressure and vertical velocity; for example, Bushby and Timpson (1967) used a research model with 40-km horizontal grid spacing and 10 vertical levels to study precipitation and dynamic processes near fronts. Operational models, too, gradually began to function with lower horizontal grid spacing and increasing complexity. For example, in the 1970s, they gained the ability to integrate the full equations of motion (Lynch 2008; Bauer et al. 2015); meanwhile, the grid spacing of the National Centers for Environmental Prediction (NCEP) operational model decreased from 190.5 km in the 1970s (Limited-area Fine-mesh Model; Petersen and Stackpole 1989) to 30 km in 1994 (National Meteorological Center Mesoscale Eta Model; Black 1994).

In the 1990s, the prospect of explicitly resolving thunderstorms numerically became conceivable (e.g., Lilly 1990). The results of Weisman et al. (1997), who simulated squall lines using varying model grid spacing, suggested that convective systems could be explicitly simulated (i.e., simulated without the use of a convective parameterization scheme) at a grid spacing as coarse as 4 km. During the Bow-Echo and Mesoscale Convective Vortex Experiment (BAMEX), it was shown that 4-km grid spacing models run without convective parameterization could be useful for predicting convection operationally (Done et al. 2004). Indeed, it was found that the explicit 4-km forecasts of convection better predicted convective mode and the number of daily

mesoscale convective systems compared to forecasts created using a 10-km grid spacing model with parameterized convection (Done et al. 2004). Given the results of Done et al. (2004), the 2004 Storm Prediction Center–National Severe Storms Laboratory Spring Program decided to test the value of CAMs to human forecasters (Kain et al. 2006). It was found that the CAMs helped forecasters better predict convective initiation, evolution, and mode (Kain et al. 2006). Due to these demonstrated benefits of CAMs (e.g., Kain et al. 2006; Done et al. 2004), CAMs have recently been implemented operationally (e.g., the High Resolution Rapid Refresh (HRRR) model; Benjamin et al. 2016). However, while CAMs offer skillful and useful forecast guidance for fields related to convection (e.g., simulated low-level reflectivity, UH, etc.), deterministic CAMs provide no information about forecast uncertainty. By contrast, ensembles aim to account for uncertainties in model parameterizations and initial conditions (Roebber et al. 2004).

b) Weather prediction using ensembles

Early conceptions of ensemble forecasting can perhaps be traced to Poincare (1914), who recognized that, for non-linear systems, adding small perturbations to a forecast's initial conditions could drastically alter the forecast; indeed, Poincare (1914) postulated this this behavior could be responsible for limiting predictability (Bauer et al. 2015). One of the first to study forecast uncertainty in the context of NWP was Thompson (1957), who conducted an analysis of how initial forecast errors grow with time. In a similar vein, Lorenz (1963) found that if the present and past states of a non-periodic system are not completely known, the skill of a forecast will deteriorate with

time. Using a 28-variable atmospheric model, Lorenz (1965) showed that small errors in initial conditions will grow to large errors over time. Given his finding, Lorenz (1965) conceived of running multiple simulations, each with slightly different initial conditions, to determine the range of possible future atmospheric states. Building on the work of Lorenz (1965), Epstein (1969a) advocated an ensemble approach to forecasting to account for initial condition uncertainty; he noted that the time series of the ensemble mean forecast behaved differently than the time series of any individual ensemble member. Epstein (1969b) found that, relative to a deterministic forecast, an ensemble forecast had a lower mean square error, extended the range of time for which the forecast was useful, and provided probabilistic information that could help convey the forecast's uncertainty. Leith (1974) used random perturbations (i.e., a Monte Carlo procedure) to create ensemble forecasts and, like Epstein (1969b), found that ensemble mean forecasts outperformed any individual ensemble member.

The promise of ensemble forecasting ultimately led to the use of ensembles operationally. Indeed, both the European Center for Medium-Range Weather Forecasts (ECMWF) and the U.S. National Meteorological Center (NMC) started operational forecasting in December 1992 (Toth and Kalnay 1993; Tracton and Kalnay 1993; Molteni et al. 1996). The implementation of ensembles operationally signaled a transition from a deterministic to a probabilistic forecasting approach as well as a shift in the goal of NWP: instead of merely optimizing forecast skill, the aim was to additionally optimize the utility of NWP products (Tracton and Kalnay 1993). While ensembles in the 1990s and early 2000s were generally used to create medium- to long-range forecasts, ensembles gradually began to operate at finer resolution and shorter

time scales (Roebber et al. 2004). For example, in 2007, the Center for Analysis and Prediction of Storms (CAPS) began running a real-time, 10-member convection-allowing ensemble with 4-km grid spacing out to 33 hours as part of the 2007 NOAA HWT SFE (Xue et al. 2007; Schwartz et al. 2010). These 4-km convection-allowing forecasts showed promise for convective-related fields, such as simulated composite reflectivity (Xue et al. 2007), accumulated precipitation, and probability of precipitation (Schwartz et al. 2010). Indeed, using data from the 2007 HWT SFE, Clark et al. (2009) found that a 5-member convection-allowing ensemble produced better precipitation forecasts than a 15-member convection-parameterizing ensemble.

The promise of fine-resolution convection-allowing ensembles has fostered ideas of their use in operations. For example, Stensrud et al. (2009) conceived of using convection-allowing ensembles to produce probabilistic warnings for tornadoes, hail, flash floods, and damaging winds as part of the Warn-On-Forecast initiative. Currently, experimental convection-allowing ensembles, such as the Storm Scale Ensemble of Opportunity (SSEO; Jirak et al. 2012) and the National Center for Atmospheric Research Ensemble Prediction System (NCAR EPS; Schwartz et al. 2015a) are being run and evaluated for eventual use in operations. A recent study by Schwartz et al. (2017) supports the eventual use of 1-km grid spacing convection-allowing ensemble forecasts once sufficient computing power becomes available.

Nevertheless, ensembles generally remain under-dispersive (i.e., the observation routinely falls outside of the "envelope" of ensemble member solutions), and questions about optimal ensemble configuration remain (e.g., Roebber et al. 2004; Duda et al. 2014; Johnson and Wang 2017). A first attempt to address these questions was made

during the 2016 HWT SFE with the Community Leveraged Unified Ensemble (CLUE; Clark et al. 2016), a convection-allowing ensemble made up of 65 members contributed from a variety of organizations, including the National Severe Storms Laboratory (NSSL), the Center for Analysis and Prediction of Storms (CAPS), the University of North Dakota, NOAA's Earth Systems Research Laboratory/Global Systems Division (ESRL/GSD), and the National Center for Atmospheric Research (NCAR). All ensemble members had similar specifications (e.g., 3-km horizontal grid spacing, 0000 UTC initialization on weekdays, domain covering the contiguous United States (CONUS), etc.) to allow for controlled experiments with various ensemble subsets. One aim of this thesis is to use data from the 2016 CLUE to determine how the inclusion of multiple microphysics and planetary boundary layer (PBL) schemes impact the spread and skill of a convection-allowing ensemble at various spatial scales.

3. Research questions and hypotheses

Two research components have been designed and implemented to meet the goal of the thesis, which is to determine how convection-allowing models and ensembles may be optimally used and designed. The first component uses data from 63 days of the 2010-2011 NOAA HWT SFEs to compare next-day probabilistic severe weather forecasts derived from three convection-allowing WRF-ARW model configurations: a deterministic configuration with 4-km horizontal grid spacing, a deterministic configuration with 1-km horizontal grid spacing, and an 11-member ensemble with 4-km grid spacing. As in Sobash et al. (2011), large values of simulated updraft helicity are used as a proxy for severe weather reports, and these severe weather proxies are

spatially smoothed to create probabilistic severe weather forecasts. The second component uses data from 23 days of the 2016 NOAA HWT SFE to compare the spread and skill of two 3-km grid spacing convection-allowing ensembles: a 9-member "mixed-physics" ensemble containing multiple microphysics and planetary boundary layer (PBL) parameterizations and a 10-member "single-physics" ensemble containing only one microphysics scheme and one PBL scheme. For both the mixed- and single-physics ensembles, spread is examined at varying spatial scales and for four different variables, including: 2-m temperature, 2-m dewpoint temperature, 500-mb height, and hourly accumulated precipitation. Meanwhile, ensemble skill is also evaluated at varying spatial scales for hourly and 6-hourly precipitation forecasts.

The primary research question (Q1) associated with the first research component is:

Q1: For next-day, all-hazards severe weather forecasts derived from simulated UH, which of the following two approaches results in forecasts with higher quality and value: reducing the horizontal grid spacing of a deterministic CAM from 4 km to 1 km, or adding members to create a 4-km, 11-member CAM ensemble?

The hypothesis (H1) associated with Q1 is as follows:

H1: While both the 1-km deterministic CAM and the 11-member, 4-km ensemble will provide greater forecast quality relative to the 4-km deterministic CAM, more quality and value will be gained by creating the 4-km ensemble than by reducing the horizontal

grid spacing from 4 km to 1 km.

To test H1, probabilistic next-day severe weather forecasts from each of the three model configurations are tested and evaluated using metrics such as area under the relative operating characteristic curve (AUC), attributes diagrams, and performance diagrams. Individual-day forecasts are also evaluated objectively and subjectively.

The primary research questions (Q2 and Q3) associated with the second component are:

Q2: For each of the four variables mentioned above (i.e., hourly accumulated precipitation, 2-m temperature, 2-m dewpoint temperature, and 500-mb height), will the spread (i.e., variance) of the mixed-physics ensemble forecasts be greater than that of the single-physics ensemble forecasts at any/all spatial scales?

Q3: Will the mixed-physics ensemble produce more skillful hourly precipitation forecasts relative to the single-physics ensemble at any/all spatial scales?

The hypotheses (H2 and H3) corresponding to Q2 and Q3 are:

H2: In general, the variance of the mixed-physics ensemble forecasts will be greater than the variance of the single-physics ensemble forecasts for the low-level variables (i.e., 2-m temperature and 2-m dewpoint temperature) and hourly accumulated precipitation but not for 500-mb height. However, as the spatial scale increases, the

variance of the mixed- and single-physics forecasts will become increasingly similar for all four variables.

H3: Because of its greater member diversity, the mixed-physics ensemble will produce more skillful 6-hourly precipitation forecasts than the single-physics ensemble at smaller spatial scales. Additionally, the mixed-physics ensemble will demonstrate skill at a smaller scale relative to the single-physics ensemble. As the spatial scale increases, the skill of the mixed- and single-physics ensemble forecasts will be increasingly similar.

To test H2, the variance of each ensemble is computed for forecast hours 1-36 at spatial scales ranging from 3 km to 720 km for each of the four variables mentioned in H2. To test H3, 6-hourly precipitation forecasts from each ensemble are created and evaluated for forecast hours 1-36 at spatial scales ranging from 3 km to 720 km using fractions skill score (FSS), AUC, and attributes diagrams.

4. Thesis Organization

Two papers are developed to test the above research questions and hypotheses. Each paper is assigned to a separate thesis chapter and is designed to function as a standalone journal article. The first paper, *Comparison of Next-Day Probabilistic Severe Weather Forecasts from Coarse- and Fine-Resolution CAMs and a Convection-Allowing Ensemble*, has been accepted for publication by *Weather and Forecasting* and makes up Chapter 2. The second paper, *Spread and Skill in Mixed- and Single-Physics Convection Allowing Ensembles at Different Spatial Scales*, is intended to form the

basis of a future journal article and is assigned to Chapter 3. Finally, Chapter 4 provides a general discussion of both papers, directly addresses the research questions and hypotheses posed earlier in the introduction, and offers suggestions for future work.

Chapter 2: Comparison of Next-Day Probabilistic Severe Weather Forecasts from Coarse- and Fine-Resolution CAMs and a Convection-Allowing Ensemble

Eric D. Loken[1,2,4], Adam J. Clark[4], Ming Xue[2,3], Fanyou Kong[3]

[1]Cooperative Institute for Mesoscale Meteorological Studies, The University of Oklahoma, Norman, Oklahoma
[2]School of Meteorology, University of Oklahoma, Norman, Oklahoma
[3]Center for Analysis and Prediction of Storms, The University of Oklahoma, Norman, Oklahoma
[4]NOAA/OAR National Severe Storms Laboratory, Norman, Oklahoma

Abstract

Given increasing computing power, an important question is whether additional computational resources would be better spent reducing the horizontal grid spacing of a convection-allowing model (CAM) or adding members to form CAM ensembles. The present study investigates this question as it applies to CAM-derived next-day probabilistic severe weather forecasts created by using forecast updraft helicity as a severe weather proxy for 63 days of the 2010 and 2011 NOAA Hazardous Weather Testbed Spring Forecasting Experiments. Forecasts derived from three sets of Weather Research and Forecasting model configurations are tested: a 1-km deterministic model; a 4-km deterministic model; and an 11-member, 4-km ensemble. Forecast quality is evaluated using relative operating characteristic (ROC) curves, attributes diagrams, and performance diagrams, and forecasts from five representative cases are analyzed to investigate their relative quality and value in a variety of situations.

While no statistically significant differences exist between the 4-km and 1-km deterministic forecasts in terms of area under ROC curves, the 4-km ensemble forecasts offer weakly significant improvements over the 4-km deterministic forecasts over the entire 63-dataset. Further, the 4-km ensemble forecasts generally provide greater

forecast quality relative to either of the deterministic forecasts on an individual day. Collectively, these results suggest that, for purposes of improving next-day CAM-derived probabilistic severe weather forecasts, additional computing resources may be better spent on adding members to form CAM ensembles than on reducing the horizontal grid spacing of a deterministic model below 4 km.

1. Introduction

The prospect of increasing a numerical weather prediction model's forecast skill by decreasing its horizontal grid spacing has interested scientists for some time (e.g., Lilly 1990; Brooks et al. 1992; Weygandt and Seaman 1994; Mass et al. 2002). This interest is evidenced, in part, by the decrease in horizontal grid spacing of the United States operational North American Mesoscale (NAM) Model from 80-km in 1993 to 12-km in 2001 (Kain et al. 2008) and the advent of the 3-km grid spacing High Resolution Rapid Refresh (HRRR) model (Benjamin et al. 2016). As increasing computing power has permitted models to operate with finer horizontal resolution, it has become clear that 4-km is about the maximum grid spacing that can still produce the dominant circulations in mid-latitude mesoscale convective systems without having to use convective parameterization (e.g., Weisman et al. 1997, Done et al. 2004). These models run without convective parameterization are typically referred to as convection-allowing models, or CAMs.

Decreasing horizontal grid spacing has generally led to clear improvements in forecast skill at convection-parameterizing resolutions (e.g., Mass et al. 2002). Decreasing from convection-parameterizing to convection-allowing grid spacing has

also led to clear improvements, especially for fields related to convection (e.g., Clark et al. 2009, 2010, 2012b; Weisman et al. 2008; Done et al. 2004). However, further decreasing the grid spacing of a convection-allowing model has provided mixed results. For example, Kain et al. (2008) compared 4-km and 2-km Weather Research and Forecasting-Advanced Research WRF (WRF-ARW) model forecasts for simulated lowest-level reflectivity, hourly precipitation, and hourly updraft helicity fields during the 2005 NOAA Hazardous Weather Testbed Spring Forecasting Experiment (HWT SFE), finding that while the finer resolution forecasts tended to produce more convective detail, they added no significant quality or value relative to the coarser resolution forecasts. Schwartz et al. (2009) obtained a similar result when comparing 4- and 2-km WRF-ARW simulated 1-km above ground level reflectivity and 1-h accumulated precipitation forecasts during the 2007 NOAA HWT SFE. Likewise, Clark et al. (2012b) noted that participants in the 2010 HWT SFE gave similar subjective ratings to 1-km and 4-km CAM forecasts of deep convection.

Meanwhile, Johnson et al. (2013), who studied 4- and 1-km 30-h forecasts of 1-h accumulated precipitation over 91 days of the 2009-2011 NOAA HWT SFEs, found that the 1-km forecasts had a significantly greater median of maximum interest relative to the 4-km forecasts but noted that the two sets of forecasts had similar object-based threat scores. When the 1-km forecasts were mapped onto a 4-km grid, the two forecast configurations had similar verification scores, suggesting that the 1-km forecasts were superior predominantly on scales not fully resolvable with 4-km grid spacing (Johnson et al. 2013). Interestingly, Roberts and Lean (2008), who used the Met Office Unified Model to compare precipitation forecasts from runs with 12-, 4-, and 1-km horizontal

grid spacing, found that the 1-km forecasts outperformed the 4- and 12-km forecasts for all scales greater than 15 km. With this said, Roberts and Lean (2008) used a modified form of convective parameterization for their 4-km run, and they focused on forecast time periods of 7 hours or less, when differences between the 4- and 1-km forecasts may not yet have been dominated by large-scale errors (e.g., Schwartz et al. 2009; Potvin and Flora 2015). VandenBerg et al. (2014) compared storm motion forecasts from models with 1- and 4-km horizontal grid spacing and concluded that the 1-km storm motion forecasts may offer some improvements over the 4-km forecasts, noting that—at least when viewed relative to environmental flow and for short-lived storms—mean storm speeds produced by the 1-km model were significantly closer to the observed mean storm speeds. Potvin and Flora (2015) studied the impact of varying horizontal model resolution on idealized supercells, concluding that—at least in an idealized framework and at short (i.e., on the order of 1-h) time scales—4-km horizontal grid spacing was too coarse to reliably resolve key supercell processes, since storms tended to decay prematurely and have large track errors. However, Potvin and Flora (2015) noted that their 3-km grid spacing simulations typically resolved important operational features, such as low-level rotation tracks, while their 1-km simulations had the ability to resolve rapid changes in low-level rotation. For a single case of convection over the central U.S. on 26 May 2008, Xue et al. (2013) found that 1-km grid spacing forecasts subjectively outperformed the corresponding 4-km forecasts in terms of storm structure and intensity.

Given mixed findings on the benefits of decreasing horizontal grid spacing beyond about 4 km, an open question is whether additional computing power should be

spent on increasing horizontal resolution or on adding ensemble members to improve forecasting skill. Indeed, an ensemble may provide an advantage over a similarly-configured deterministic model by accounting for forecast uncertainties related to errors in initial conditions and model parameterizations (e.g., Wandishin et al. 2001). However, it is currently unknown how the skill of an ensemble at coarser—but still convection-allowing—resolution would compare to that of a similarly-configured deterministic model with finer grid spacing. The present study seeks to address this question as it applies to next-day probabilistic severe weather forecasts derived from forecast updraft helicity (UH).

UH has been identified as an important severe weather forecasting parameter. For example, Kain et al. (2008) used large values of hourly UH to successfully identify mesocyclones during the 2005 SFE, and Kain et al. (2010) developed a strategy for calculating temporal maximum UH by tracking the largest values occurring at any time step between model output times, thus accounting for the rapid evolution in convective storms. Hereafter, UH refers to the hourly maximum quantity (i.e., the maximum UH value at any time step between hourly output times). Sobash et al. (2011), inspired by the perceived correspondence between large values of UH and severe weather reports during the 2008 SFE, treated "extreme" values of simulated UH as "surrogate" severe weather reports. Applying a spatial smoother to these surrogate reports, Sobash et al. (2011) created a field of surrogate severe probabilistic forecasts (SSPFs) that provided skillful and useful guidance for severe weather forecasters. Clark et al. (2012c) and Clark et al. (2013) investigated whether simulated UH track lengths corresponded with observed tornado track lengths, finding that simulated UH track lengths showed some

skill as proxies for tornado track lengths particularly during the spring months and particularly when the storm environment was used to filter the UH tracks associated with elevated and/or high-based storms. While most previous research has focused on deterministic UH forecasts, Sobash et al. (2016b) investigated the effect of using a 30-member CAM ensemble to create day-1 and day-2 SSPFs over a 32-day period coinciding with the Mesoscale Predictability Experiment (Weisman et al. 2015). Sobash et al. (2016b) found that the ensemble SSPFs were more skillful and reliable relative to the deterministic SSPFs on the mesoscale but not necessarily for larger scales.

This paper builds on the work of Sobash et al. (2011) and Sobash et al. (2016b) by investigating how a reduction in grid spacing from 4- to 1- km and the creation of a 4-km ensemble influences the quality and value (e.g., Murphy 1993) of next-day SSPFs derived from forecast UH fields. The study is organized as follows: section 2 details the model specifications and the methodology used for this study; section 3 provides the results and examines five case study days; section 4 summarizes and discusses the results; and section 5 outlines potential future work.

2. Methods

a) Model specifications

During the 2010 and 2011 HWT SFEs, the Center for Analysis and Prediction of Storms (CAPS) ran a 26-member Storm-Scale Ensemble Forecast (SSEF) system with 4-km grid spacing (Clark et al. 2012a). The present study analyzes model output from the control member of this SSEF system, an equivalent 1-km version of this control member, and a subset of 11 ensemble members for a total of 63 days (38 days from

2010 and 25 days from 2011) over the 2010 and 2011 SFEs (Table 2.1)[1]. Both the 4-km and 1-km deterministic models and all 11 members of the ensemble subset use the ARW-WRF model (Skamarock et al. 2008) dynamic core and have 51 vertical levels. The domain of all models covers the contiguous United States, although the analysis domain is restricted to the eastern two-thirds of the United States (Fig. 2.1). Analyses from the 0000 UTC 12-km NAM are used as the background for both deterministic models. Then, WSR-88D data are assimilated along with surface and upper air observations using the Advanced Regional Prediction System three-dimensional variational data assimilation and cloud analysis system (ARPS 3DVAR; Xue et al. 2003, Gao et al. 2004). The subset of 11 SSEF members was chosen for the ensemble because hourly maximum updraft helicity was available from these members (7 of the 26

[1] Three of the days in the 63-day dataset—26 May 2010, 27 April 2011, and 29 April 2011—contain missing data from at least one ensemble member. Each of the three days is assessed on a case-by-case basis to determine how to appropriately handle the analysis for each day. 26 May 2010 has one forecast hour missing from two different ensemble members. Data from the 28th forecast hour is missing from the arw_m5 member, and data from the 26th forecast hour is missing from the arw_m6 member. Given that these forecast hours are late in the forecast period (and therefore likely do not contain the maximum UH over the entire period), and given that only one forecast hour is missing from only two of eleven ensemble members, the missing data is neglected. In the case of 27 April 2011, data is missing from all forecast hours for the arw_m13 member. Given that only one of 11 ensemble members is missing, the decision is made to include 27 April 2011 in the dataset but to evaluate the data as having come from a 10-member ensemble instead of an 11-member ensemble. In the case of 29 April 2011, two ensemble members, the arw_m5 and the arw_m12 members, contain missing data for the final six forecast hours. Given that only two SPC storm reports occurred in the contiguous U.S. on this day and that both occurred between 1900 UTC and 2000 UTC (i.e. forecast hours 7 and 8), the missing data are neglected.

members that did not use the ARW dynamic core did not produce hourly maximum UH), and these were the only members that accounted for both model and analysis error with mixed-physics and perturbed initial conditions and lateral boundary conditions (ICs/LBCs), respectively. The other members shared the same set of ICs/LBCs or LBCs to study various IC perturbation methods, the impact of radar data assimilation, and physics sensitivities. Thus, this set of members was less diverse and tended to cluster around the arw_cn member solution. The ensemble IC/LBC perturbations are derived from evolved (through 3-h) perturbations of 2100 UTC NCEP operational Short-Range Ensemble Forecast (SREF; Du et al. 2006) system members and added to the ARPS 3DVAR analyses. Corresponding SREF forecasts are used for LBCs. Full model specifications are provided in Table 2.2.

One notable difference between the 2010 and 2011 forecasts is forecast length: the models produced 30-h forecasts in 2010 but 36-h forecasts in 2011. Hence, for 2010 the 18-h period from 12z to 6z on the next day is investigated (12-30 h forecast times), while for 2011 the 24-h period from 12z to 12z on the next day is examined (12-36 h forecast times). Because the primary goal is to assess next-day severe weather forecast guidance—and because the Storm Prediction Center's (SPC's) Day 1 Convective Outlook forecasts span from 12z to 12z—output from the first twelve forecast hours after model initialization is ignored for both forecasts.

b) Producing SSPFs from UH

As in Sobash et al. (2011, 2016b), extreme values of 2-5 km UH are treated as surrogate severe weather reports (SSRs). 2-5 km UH is computed using the following

formula, as in Kain et al. (2008) and Sobash et al. (2011):

$$\mathrm{UH} = \sum_{z=2000\mathrm{m}}^{z=5000\mathrm{m}} \overline{w\zeta\Delta z} = \left(\overline{w\zeta_{2,3}} + \overline{w\zeta_{3,4}} + \overline{w\zeta_{4,5}}\right) \times 1000 \qquad (1),$$

where w is vertical velocity (in ms^{-1}), ζ is vertical vorticity (in s^{-1}), and Δz is the vertical distance between computation levels (here, 1000m). The subscripts indicate the bottom and top computational levels (in km), and the overbars denote an average over the layer between the two given computational levels.

SSPFs are derived from SSRs using the following methodologies for deterministic and ensemble simulations: First, the maximum UH value that occurred at each grid box over the entire 18-h (for 2010) or 24-h (for 2011) period of interest is found for the 4- and 1-km models each day. These maximum daily UH values are remapped to an 80-km grid using the maximum UH from all of the finer-resolution grid points falling within the 80-km grid boxes. Remapping to an 80-km grid is done to match the verification scales used by the SPC and to reduce the computational expense of creating the SSPFs. It should be noted that the SSPFs produced by remapping to an 80-km grid are very similar to those produced on the native 4- or 1-km grids using a 40-km radius of influence when the same 4- and 1-km thresholds are used on both the native and 80-km grids. Remapping to the coarser grid is therefore done to save computation time. After remapping, a UH threshold is applied to produce a binary field, with 1s assigned to points equal to or exceeding the UH threshold and 0s assigned to all other points. UH thresholds from 25 to 125 m^2s^{-2} in increments of 25 m^2s^{-2} are used for

the 4-km data since these represent typical extreme values based on subjective experience and previous research (e.g., Kain et al. 2008; Sobash et al. 2011, 2016b). Because UH is grid spacing dependent (i.e., increasing resolution results in higher UH), the same thresholds could not be similarly applied to the 1-km UH data. Instead, UH values at equivalent percentiles for each of the thresholds used for the 4-km data are found for the 1-km data. The percentiles are computed using the distributions of UH from all cases after remapping the maximum values to the 80-km grid. This procedure ensures that the number of SSRs in the 1- and 4-km forecasts are similar, thus minimizing the impact of differences in biases. The percentiles are computed separately for 2010 and 2011 data because it was thought that the forecast length difference between the two years (30-h in 2010 vs. 36-h in 2011) might cause the characteristics of the distributions to be slightly different. Indeed, the 1-km UH values from 2011 are usually higher than 2010. Table 2.3 lists the UH thresholds used for the 4-km data, their percentiles, and the UH value from the 1-km data at each of these percentiles for 2010 and 2011. Lastly, as in Sobash et al. (2011), a Gaussian kernel is applied to the binary field to produce forecast probabilities using the formula:

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

where f is the probability value at a given grid point, N is the total number of grid points containing a SSR, d_n is the distance from the grid point to the point of the nth SSR, and σ is the standard deviation of the Gaussian kernel (hereafter referred to as the spatial

smoothing parameter). Sobash et al. (2011) found that σ = 160-km and σ = 200-km produced the best reliability for the smallest UH thresholds tested. Herein, σ = 120-km is used because it produces reliable forecasts for some of the larger UH thresholds examined and because resolution is not sacrificed as much as with larger σ values (i.e., more frequent larger probabilities can occur). To produce SSPFs from the 11-member 4-km ensemble, a similar procedure is used. However, after UH from each member is remapped to the 80-km grid and a specified threshold is applied, the ratio of members that exceed the threshold is calculated for each point. Then, the Gaussian smoother is applied to produce the SSPF field. Note that creating the ensemble SSPF field using this procedure is identical to creating the field by averaging the individual SSPFs from each ensemble member (Sobash et al. 2016b). For the ensemble SSPFs, σ is varied from 60- to 120-km in 30-km increments to identify the optimal value of σ in the ensemble framework.

c) Verification

To verify the SSPFs, archived observed storm reports (OSRs) are obtained from the SPC website. These OSRs include reports of wind 58 miles per hour or greater, hail measuring 1 inch or greater in diameter, and tornadoes. The OSRs are filtered to include only those that fall within the designated forecast time periods. Hence, OSRs from 12z to 6z on the following day are considered for 2010, and OSRs from 12z to 12z on the following day are considered for 2011. As with the SSRs, a binary 80-km grid of OSRs is constructed. Grid boxes with at least one OSR are assigned a value of 1, while all other grid boxes are assigned a value of 0. Relative operating characteristic (ROC)

curves (Mason 1982), attributes diagrams (Hsu and Murphy 1986), and performance diagrams (Roebber 2009) are constructed to help evaluate the quality of the SSPFs.

ROC curves plot probability of detection (POD), defined as:

$$\text{POD} = \frac{\text{hits}}{\text{hits} + \text{misses}} \qquad (3),$$

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

$$\text{POFD} = \frac{\text{false alarms}}{\text{false alarms} + \text{correct negatives}} \qquad (4),$$

Herein, POD and POFD are computed at specified levels of probability: 1%, 2%, 5%, 10%, 15%, 20%, 25%, 30%, 35%, 40%, 45%, 50%, 55%, 60%, 65%, 70%, 75%, 80%, 85%, 90%, and 95% to create the ROC curves. Each probability level is used to convert the probabilistic forecasts into binary (e.g., "yes"/"no") forecasts; grid boxes meeting or exceeding the given probability level are considered to be "yes" forecasts at that probability level. One method of determining forecast quality from a ROC curve is by assessing the area under the ROC curve (AUC; e.g., Marzban 2004), a single-number metric that measures a forecast's ability to discriminate between the occurrence and non-occurrence of an observed event (e.g., Mason and Graham 2002). In the present case, an observed event is defined as the occurrence of an OSR within a given 80-km grid box. An AUC value of 1.0 indicates a perfect forecast, while an AUC value of 0.5 or less represents a random forecast. An AUC value of 0.70 is typically considered to

represent the lower limit of skill for probabilistic forecast systems (Buizza et al. 1999; Sobash et al. 2011). In the present study, AUC values are computed over the entire 63-day dataset as a means of evaluating the overall performance of each of the three forecasts. AUC values are also computed over individual days to evaluate the performance of each individual daily forecast. In both cases, a trapezoidal approximation is used to compute AUC (Wandishin et al. 2001).

While ROC curves and AUC values assess a forecast's ability to discriminate between the occurrence and non-occurrence of events, these metrics do not give information about a forecast's bias (e.g., Wilks 2001). For this reason, attributes diagrams, which contain information about forecast bias, make good complements to ROC curves. Attributes diagrams plot observed relative frequency against forecast probability; herein, the attributes diagrams are made using the same levels of probability used for the ROC diagrams. Reliable forecasts are those in which the forecast probabilities correspond to the observed relative frequencies; therefore, points that fall along a diagonal line of slope 1 from the lower-left to upper-right of the diagram (called the perfect reliability line) are said to have perfect reliability. Points that fall above (below) the perfect reliability line represent under- (over-) forecasts. In addition to the perfect reliability line, attributes diagrams display horizontal and vertical lines at the sample climatological frequency (abbreviated herein as "sample climatology"), which is found by taking the total number of "yes" observations (i.e., occurrences of severe weather) divided by the total number of forecasts in all forecast bins. The horizontal sample climatology line is also referred to as the no resolution line, since points along this line have no resolution. Attributes diagrams also contain a no

skill line, located halfway between the perfect reliability and no resolution lines. Points along the no skill line do not contribute to the Brier skill score for a reference forecast of climatology. Meanwhile, points falling between the vertical sample climatology line and the no skill line contribute positively to the Brier skill score, since these points are closer to perfect reliability line than they are to the no resolution line (Wilks 1995).

Performance diagrams plot POD against success ratio (SR), defined as:

$$\mathrm{SR} = 1 - \frac{\text{false alarms}}{\text{hits} + \text{false alarms}} \qquad (5),$$

In addition, performance diagrams give information about a forecast's bias, defined as:

$$\text{bias} = \frac{\text{hits} + \text{false alarms}}{\text{hits} + \text{misses}} \qquad (6),$$

and critical success index (CSI), defined as:

$$\mathrm{CSI} = \frac{\text{hits}}{\text{hits} + \text{misses} + \text{false alarms}} \qquad (7),$$

since POD and SR both depend on bias and CSI (Roebber 2009). Herein, POD, SR, bias, and CSI are computed at the same 21 probability levels mentioned previously to produce the performance diagrams. Indeed, each of the 21 probability levels from a given forecast is explicitly plotted on a performance diagram; therefore, the diagrams are useful for users who wish to determine the probability level that yields a certain

value of POD, SR, bias, or CSI for a given forecast. Performance diagrams are also useful for comparing multiple forecasts. Since POD, SR, bias, and CSI are all optimized at 1.0, points that lie closer to the upper right-hand corner of a performance diagram represent more skillful forecasts (Roebber 2009).

A resampling technique outlined by Hamill (1999) is used to test for significant differences in aggregate AUC between the three forecast sets over the entire dataset. A resampling significance test is chosen because the AUC depends on contingency table elements, and small changes in contingency table elements may produce large changes in the AUC (Hamill 1999). Therefore, more common significance tests, such as the paired t-test, may be inappropriate to use in this case. Conceptually, the resampling technique builds a null distribution of the difference in the aggregate AUC between two forecast sets (e.g., the 1-km deterministic forecast set and the 4-km deterministic forecast set) by repeated random sampling of the contingency table elements of those forecast sets (Hamill 1999). The actual difference in aggregate AUC (computed by subtracting the aggregate AUC of the second forecast from the aggregate AUC of the first forecast) is compared to the null distribution to determine whether or not the difference in aggregate AUC between the two sets of forecasts is significant.

To build a null distribution of aggregate AUC differences between two forecast sets, two separate lists of contingency table elements are created. To start, contingency table elements are obtained from each of the two forecast sets for each of the 63 days in the dataset. The elements of each of the 63 days from the first forecast set (forecast set 1) are assigned to list 1, while the elements of each of the 63 days from the second forecast set (forecast set 2) are assigned to list 2. Next, for each of the 63 days in the

dataset, it is randomly determined whether the two lists will exchange contingency table elements for a given day. After the procedure is completed for all 63 days, the aggregate AUC is computed for lists 1 and 2, respectively. Finally, the difference between the aggregate AUC from list 1 and the aggregate AUC from list 2 is computed. This entire procedure is repeated 1000 times in order to form a null distribution of aggregate AUC differences. Finally, the actual aggregate AUC difference is compared to the null distribution to determine significance. If the actual aggregate AUC difference exceeds the 97.5th percentile or falls beneath the 2.5th percentile of the null distribution, the AUC difference between the two forecast sets is deemed to be significant at the 95% level.

3. Results

a) Comparing 1-km and 4-km deterministic forecasts

ROC curves for the 1- and 4-km deterministic forecasts at each UH threshold (i.e., UH = 25, 50, 75, 100, and 125 m^2s^{-2} for the 4-km forecasts and the corresponding 1-km values for the 1-km forecasts) suggest that the lower UH threshold forecasts have greater forecast skill than the higher threshold forecasts. The 25 m^2s^{-2} 4-km forecasts and corresponding 249 m^2s^{-2} (or 281 m^2s^{-2} for 2011) 1-km forecasts have the greatest AUC values, while the 125 m^2s^{-2} 4-km forecasts and corresponding 1158 m^2s^{-2} (or 1098 m^2s^{-2} for 2011) 1-km forecasts have the lowest AUC values (Fig. 2.2). Hereafter, to simplify the analysis, the 4-km UH threshold values are used to additionally refer to the corresponding 1-km UH thresholds in the text and figures (refer to Table 2.3 for the equivalence). For a given UH threshold, the 1-km deterministic forecasts have greater AUC values than the 4-km deterministic forecasts; this pattern holds for all five

thresholds. The greatest difference between the 1-km deterministic AUC and the 4-km deterministic AUC occurs for the UH threshold of 25 m^2s^{-2}. However, for all five UH thresholds examined, the AUC differences are *not* significant at the 95% level (Table 2.4).

Varying the UH threshold directly influences the reliability of the forecasts. The 25 m^2s^{-2} forecasts represent over-forecasts at nearly all forecast probabilities; moreover, both the 1-km and 4-km 25 m^2s^{-2} forecasts fall slightly below the no skill line for many of the probabilities, indicating that the forecasts contribute negatively to the Brier skill score at these probabilities. The 50 m^2s^{-2} forecasts, by contrast, fall near the line of perfect reliability, although these forecasts slightly under-forecast at low probabilities and slightly over-forecast at higher probabilities. The 75 m^2s^{-2} forecasts slightly under-forecast at most probabilities, while the 100 m^2s^{-2} and 125 m^2s^{-2} forecasts display a greater degree of under-forecasting.

For a given UH threshold, the reliability of the 1-km deterministic forecast is generally quite similar to the reliability of the 4-km deterministic forecast (Fig. 2.3a). However, the 1-km and 4-km reliabilities diverge slightly for the two higher UH threshold forecasts (i.e., UH = 100, 125 m^2s^{-2}) at higher forecast probability bins. These bins contain relatively few forecasts and therefore must be interpreted cautiously, since a single data point can exert undue influence on the reliability values (Fig. 2.3b).

On the performance diagrams, the lower threshold forecasts fall closer to the upper-right corner of the diagram than the higher threshold forecasts, indicating that—for a given probability—the lower threshold forecasts have greater values of CSI relative to the higher threshold forecasts (Fig. 2.4). For a given UH threshold, the 1-km

deterministic forecast demonstrates slightly greater skill than the 4-km deterministic forecast, as evidenced by the 1-km forecast's higher POD, SR, and CSI compared to the corresponding 4-km forecast. Nonetheless, the differences are generally slight, consistent with the lack of significance in AUC between the 1-km and 4-km forecasts. The performance diagrams also show that, for a given model forecast, bias and CSI values are optimized at lower probability levels as the UH threshold is increased.

b.) Comparing the deterministic and 4-km ensemble forecasts

The forecast quality of the 1-km and 4-km deterministic models is compared to that of an 11-member 4-km ensemble to determine the effect of adding ensemble members on forecast skill. Because of the finding in the previous subsection that the greatest differences in AUC between the forecasts came from the 25 m^2s^{-2} UH threshold forecasts, only the 25 m^2s^{-2} forecasts are analyzed for the 4-km deterministic and ensemble comparison.

All deterministic and ensemble forecasts with a UH threshold of 25 m^2s^{-2} have similar ROC curves (Fig. 2.5). The AUC for each curve exceeds 0.80, indicating that all forecasts show considerable skill. The 4-km deterministic forecast has the lowest AUC (0.838), while the 4-km ensemble forecast with σ = 90 km has the greatest AUC (0.874). At the 95% level, a (weakly) significant difference in AUC exists between the 4-km deterministic forecast and the ensemble forecasts with σ = 90 km and σ = 120 km (Table 2.5). No statistically significant differences are found between the 1-km deterministic forecast and any of the ensemble forecasts.

While the three ensemble forecasts (i.e., σ = 60, 90, 120 km) have similar ROC

curves and AUC values, the ensemble forecasts do have notably different reliability. The ensemble forecast with σ = 120 km has the best reliability, as its curve in the attributes diagram lies closest to the perfect reliability line (Fig. 2.6a). This result makes sense, given that the σ = 120 km ensemble forecast benefits from a high degree of spatial smoothing as well as ensemble smoothing. As a result, the σ = 120 km ensemble forecast has fewer high (i.e., $\geq$ 0.60) probabilities compared to the other ensemble and deterministic forecasts (Fig. 2.6b), which helps to reduce over-forecasting bias. However, even the σ = 120 km ensemble forecast has a tendency to over-forecast at nearly all probabilities. All three ensemble forecasts' curves on the attributes diagram mostly reside above the no-skill line, representing at least a slight improvement over either of the deterministic forecasts.

The performance diagrams indicate that the three ensemble forecasts have similar skill levels, as the ensemble forecast points are clustered very close to each other and are located about the same distance from the upper right-hand corner of the plot (Fig. 2.7). The 1-km deterministic forecast has less skill than any of the ensemble forecasts but greater skill than the 4-km deterministic forecast. These results corroborate the implications of the ROC curves and the AUC analysis; namely, that over the entire 63-day dataset, the ensemble forecasts have greater skill than the 1-km deterministic forecast, which in turn has greater skill than the 4-km deterministic forecast. For the five sets of forecasts examined, the performance diagrams indicate that a probability level of 40-50% optimizes bias, while a probability level of 35-40% optimizes CSI.

c) Comparing the AUC for individual days

For 61 of the 63 days in the dataset, individual-day AUC is computed from the 1-km deterministic, 4-km deterministic, and the 4-km ensemble forecasts (σ = 90 km) using the 25 m^2s^{-2} UH threshold. No AUC is computed for 29 April 2011 or 4 May 2011 because no OSRs occurred inside of the analysis domain for those days, resulting in an indeterminate POD (i.e., $\frac{\text{hits}}{\text{hits+misses}} = \frac{0}{0}$). For each of the 61 days analyzed, the 4-km deterministic AUC is subtracted from the 1-km deterministic AUC and the 4-km ensemble AUC, respectively, to obtain two distributions, which describe how the 1-km deterministic and 4-km ensemble forecasts perform relative to the 4-km deterministic forecasts. When the daily 4-km deterministic AUC is subtracted from the corresponding daily 1-km deterministic AUC, the distribution peaks just to the right of the zero line, indicating that, for most days, the 1-km deterministic model gives a slightly better forecast (in terms of AUC) than the 4-km deterministic model (Fig. 2.8a). The distribution has a small left tail, suggesting that the 4-km deterministic model performs notably better than the 1-km deterministic model for only a handful of days; the vast majority of the data points are located to the right of the zero line. When the daily 4-km deterministic AUC is subtracted from the corresponding daily 4-km ensemble AUC, the distribution peaks to the right of the zero line and has a long right tail, indicating that the ensemble performs better—and sometimes substantially better— than the 4-km deterministic forecast on the vast majority of the days (Fig. 2.8b). Interestingly, when the daily 1-km deterministic AUC is subtracted from the corresponding daily 4-km ensemble AUC, the distribution looks similar to that in Fig. 2.8b: it peaks to the right of the zero line and has a right tail (Fig. 2.8c), suggesting that the ensemble performs objectively better than the 1-km deterministic forecast as well as the 4-km deterministic

forecast on the majority of days in the analysis period.

Five days that span the distributions given in Figs. 2.8a,b are chosen for individual analysis. These days include: 11 May 2010, 15 June 2010, 7 June 2011, 18 May 2011, and 27 April 2011. The analysis of these individual days offers insight into what (if any) additional forecast quality and/or value can be gained by either reducing horizontal grid spacing from 4-km to 1-km or by adding members to form 4-km convection-allowing ensembles on a given day. The five case study examples are presented below:

1) 11 MAY 2010

On 11 May 2010— the day for which the 4-km ensemble forecast (AUC = 0.805) performed best relative to the 4-km deterministic forecast (AUC 0.564) in terms of AUC—the threat of severe weather existed across multiple regions. A weakening surface cyclone, located over southern Iowa at 1200 UTC, tracked east-northeastward and brought storms to the Ohio Valley around 0000 UTC on 12 May 2010. Meanwhile, low-level southerly flow from the Gulf of Mexico helped to destabilize a broad region of the Central Plains. Despite weak large-scale forcing, several severe-hail-producing storms formed in western Oklahoma before 0200 UTC on 12 May 2010. A cluster of severe storms also formed in eastern Missouri and western Kansas around 0300 UTC along a warm front. Additionally, several storms formed in northeastern Colorado ahead of an eastward-moving upper-level low during the evening hours; however, no observed severe weather was associated with these storms.

Interestingly, the three model configurations produced drastically different

forecasts on this day: the 1-km deterministic forecast (AUC = 0.561) highlighted regions near Missouri, Oklahoma, and Colorado for severe weather; the 4-km deterministic forecast highlighted Colorado but focused its secondary threat area on Ohio and surrounding states; while the 4-km ensemble showed lower severe probabilities in Colorado and gave non-zero severe probabilities over a region extending from western Kansas to eastern Ohio (Fig. 2.9a-c). OSRs were located in southwestern Kansas, northwestern Oklahoma, central Kansas, central Ohio, and western Pennsylvania, while no OSRs occurred in Colorado. Of the three forecasts, the 4-km ensemble forecast produced the greatest AUC on this day, as it had the lowest probabilities in northeastern Colorado (which reduced its POFD) and had non-zero severe probabilities over the three main regions where reports did occur (which increased its POD). While it is important to realize that probabilistic forecasts should be evaluated over multiple cases, this one case does suggest that ensembles can offer enhanced forecast quality not only by identifying regions of potential severe weather missed by a deterministic model, but also by reducing the magnitude of over-done deterministic severe probabilities.

2) 15 JUNE 2010

On 15 June 2010—the day for which the 1-km deterministic forecast (AUC = 0.763) performed objectively best relative to the 4-km deterministic forecast (AUC = 0.669)—a mid-level trough propagated northeastward into central Illinois, where a warm, moist air mass coincided with an environment containing 25-40 knots of 1000-500-mb wind shear. The shortwave trough initiated convection in eastern Missouri

around 1730 UTC, and this convection subsequently moved northeastward, producing a multitude of severe wind and hail reports in Illinois, Indiana, and western Ohio. Meanwhile, broad southerly flow resulted in moist and unstable conditions throughout much of the Southeastern United States. Although large-scale forcing for ascent and vertical wind shear were both weak in this region, numerous pulse storms formed, resulting in many severe wind reports.

Two main differences existed between the three forecasts on this day: the distribution (and magnitude) of the higher-end severe probabilities in central Illinois and the spatial coverage of the non-zero severe probabilities in the southeastern U.S. (Fig. 2.9d-f). Relative to the 4-km deterministic forecast, the 1-km deterministic forecast had greater severe probabilities in Illinois and gave a more continuous threat area in central Illinois. Additionally, the 1-km deterministic forecast introduced more non-zero severe probabilities into portions of the Southeastern U.S. relative to the 4-km deterministic forecast. The 4-km ensemble forecast (AUC = 0.864), meanwhile, maintained > 0.50 severe probabilities over central Illinois but completely filled in the Southeastern U.S. with non-zero severe probabilities, perhaps as a result of ensemble smoothing. The ensemble forecast was therefore rewarded with a greater POD and AUC relative to either deterministic forecast (although the ensemble's lower-magnitude severe probabilities in the Texas Panhandle, where no OSRs were located, may have also helped to elevate the ensemble's AUC over that of the two deterministic models). This case suggests that the 4-km ensemble forecast can potentially offer improvements in forecast quality relative to the 1-km deterministic forecast even on days when the 1-km deterministic forecast performs well relative to the 4-km deterministic forecast.

3) 7 JUNE 2011

7 June 2011 marked the day on which the 4-km deterministic forecast (AUC = 0.784) performed best relative to the 1-km forecast (AUC = 0.657). In the northern Plains, a 500-mb shortwave trough and the left exit region of a 300-mb jet tracked northeastward through the Dakotas, producing severe weather in south-central North Dakota during the early evening. During the overnight hours, storms fired in southern Wisconsin ahead of a cold front associated with a surface-low tracking northeastward through northern Minnesota. Farther eastward, strong θ_e advection and an unstable environment helped sustain a pre-existing mesoscale convective system as it tracked southward through Ohio, West Virginia, Virginia, and North Carolina. Along the Gulf Coast, numerous pulse storms formed in an environment of abundant moisture and instability; these storms produced a number of severe wind and hail reports.

Two main differences existed between the forecasts on this day: the magnitude and orientation of the higher-end severe probabilities near Pennsylvania and the spatial extent of the severe probabilities in southern Texas (Fig. 2.9g-i).

Relative to the 4-km deterministic and 4-km ensemble (AUC = 0.797) forecasts, the 1-km deterministic forecast focused its higher-end severe probabilities in Pennsylvania farther east, where fewer storm reports occurred. The 4-km ensemble had lower-magnitude severe probabilities in Pennsylvania relative to the two deterministic models, but its orientation of 0.3 severe probabilities more closely matched the orientation of the observations. In southern Texas, where no storm reports were observed, the 1-km deterministic forecast produced a much larger area of non-zero

severe probabilities relative to the 4-km deterministic and 4-km ensemble forecasts. The three forecasts generally had similar forecast probabilities over the Southeastern U.S and the Upper Midwest.

Given that the three forecasts all highlighted similar regions for severe weather on this day, the forecasts likely all had similar value. It is notable, however, that the 4-km ensemble forecast had a slightly greater AUC than the 4-km deterministic forecast on the day when the 4-km deterministic forecast performed best (in terms of AUC) relative to the 1-km deterministic forecast.

4) 18 MAY 2011

18 May 2011 represented the day on which the 4-km ensemble forecast (AUC = 0.882) performed objectively *worst* relative to the 4-km deterministic forecast (AUC = 0.912). Two upper-level troughs dominated the flow pattern on this day: one in the western U.S. and one in the eastern U.S. During the late afternoon, storms formed in eastern Colorado, which was located downstream of an upper trough and in the left exit region of a 300-mb jet. Later in the evening, storms also fired in west-central Kansas as a weak local vorticity maximum pivoted northeastward through the state and strong southerly winds helped increase low-level moisture. Perhaps due to a lack of large-scale forcing for ascent, no storms ultimately formed along the dryline in the Texas Panhandle region. Downstream of the trough over the eastern U.S., storms formed in the early afternoon, producing severe weather in Pennsylvania, West Virginia, Maryland, and Virginia.

The biggest difference between the three forecasts on this day was the ensemble

forecast's large region of ≥ 0.02 probabilities in Texas, Missouri, and Arkansas (Fig. 2.9j-l), which perhaps resulted from ensemble smoothing. Since OSRs never occurred in these states, the ensemble forecast had a large POFD relative to the deterministic forecasts. The ensemble forecast also produced slightly lower magnitudes of severe probability in the northeastern U.S., where OSRs were located, which decreased its POD relative to either deterministic forecast.

Given the ensemble forecast's inferior POFD and POD, this case shows that the quality of an ensemble forecast does not always exceed that of a deterministic forecast. Nevertheless, objectively, the ensemble's AUC is only about 0.030 lower than the 4-km deterministic forecast's AUC on the "worst day" for the ensemble relative to the 4-km deterministic forecast. This difference is small compared to the 0.241 difference in AUC between the 4-km ensemble and 4-km deterministic forecasts on 11 May 2010.

5) 27 APRIL 2011

27 April 2011 is chosen for individual analysis because it represents a "high-impact" day. In fact, this was one of the longest and deadliest tornado outbreaks in U.S. history. According to the SPC, some 937 severe reports—including 292 tornado reports— occurred over the contiguous U.S. Storms formed downstream of an upper trough and ahead of a cold front in the Southeastern U.S., where strong low-level southerly flow coincided with abundant vertical wind shear.

All three forecasts demonstrated considerable skill, as all three forecasts had AUC values greater than 0.93. The 1-km and 4-km deterministic forecasts were nearly identical and matched the observations quite well (Fig. 2.9m-o). The ensemble forecast

differed slightly from the deterministic forecasts, most notably by extending the region of ≥ 0.40 severe probabilities farther southward into south-central Mississippi, Alabama, and Georgia. The ensemble's southward shift of the higher severe probabilities likely contributed to its higher AUC relative to the two deterministic forecasts' AUC.

Because the shape of the three forecasts' higher-end (i.e., ≥ 0.40) severe probabilities was very similar, the forecasts had similar quality—and likely similar value—on this day. This case is important because it illustrates that all three model configurations—including the ensemble—can produce high-sharpness forecasts.

4. Summary and discussion

Maximum hourly 2-5 km updraft helicity (UH) forecasts from a 4-km grid spacing model, an equivalently configured 1-km grid spacing model, and an 11-member 4-km grid spacing ensemble are remapped to an 80-km grid and used to produce next-day probabilistic severe weather forecasts for 63 days of the 2010 and 2011 NOAA HWT SFEs. As in Sobash et al. (2011), extreme values of UH are treated as surrogate severe weather reports (SSRs). SSRs are smoothed spatially using a two-dimensional isotropic Gaussian smoother to create probabilistic severe weather forecasts.

After testing a variety of 4-km UH values (i.e., UH = 25, 50, 75, 100, and 125 m^2s^{-2}) and their corresponding 1-km UH values as thresholds for SSRs, it is found that the 25 m^2s^{-2} threshold not only gives the largest AUC for each of the three forecast configurations but also produces the greatest difference in AUC between the three forecast configurations. Meanwhile, the 50 m^2s^{-2} threshold yields the most reliable

forecasts, with thresholds greater than (less than) 50 m^2s^{-2} producing under-forecasts (over-forecasts). These results are consistent with Sobash et al. (2011), who found that a 4-km UH threshold of approximately 34 m^2s^{-2} (the smallest tested) gave the largest AUC values while a threshold near 41 m^2s^{-2} produced the most reliable forecasts.

Ensemble surrogate severe weather probabilistic forecasts (SSPFs) are created by calculating, at each grid point, the fraction of ensemble members exceeding the specified UH threshold (always 25 m^2s^{-2} for the ensemble forecasts) and then smoothing these values spatially using a Gaussian smoother. Three values of σ are tested for the ensemble SSPFs (σ = 60, 90, and 120 km). The σ = 90 km forecast produces the best AUC, although varying the spatial smoothing parameter has only a slight impact on forecasts' AUC values for the three values analyzed. Varying the spatial smoothing parameter has a larger impact on forecasts' reliability values: the σ = 120 km forecast gives the best reliabilities but still over-forecasts. These results agree with Sobash et al. (2011, 2016b), who found that increasing the spatial smoothing parameter had little effect on a forecast's AUC but resulted in a general progression from over-forecasting, to near-perfect reliability, to under-forecasting. Such a finding suggests that a value of σ could be found to optimize the ensemble forecast's reliability. However, increasing the spatial smoothing parameter beyond 120 km is not attempted or analyzed here since the ensemble is presumed to account for some of the spatial uncertainty in the SSRs (e.g., Sobash et al. 2016b), eliminating the need to test spatial smoothing parameter values beyond those tested for the deterministic forecasts (i.e., σ = 120 km).

A 2-sided resampling hypothesis test is conducted to test for significance between the 4- and 1-km deterministic forecasts and between the 4-km deterministic

and 4-km ensemble forecasts, using a UH threshold of 25 m^2s^{-2}. Results suggest that while no significant difference in AUC exists between the 4- and 1-km deterministic forecasts, a weakly significant difference in AUC exists between the 4-km deterministic and 4-km forecasts (with the 4-km ensemble forecasts having the greater AUC). No significant difference is found between any of the three model configurations at the four higher UH thresholds examined (i.e., UH = 50, 75, 100, and 125 m^2s^{-2}), since, as the UH threshold is increased beyond 25 m^2s^{-2}, the three sets of forecasts produce increasingly similar AUC values.

The lack of significant difference between the 4-km and 1-km deterministic forecasts agrees with Kain et al. (2008), who, qualitatively, observed a lack of dramatic differences in the forecast UH fields between the 4-km and 2-km horizontal grid-spacing models in the 2005 NOAA HWT SFE. The results of the present study also agree with Schwartz et al. (2009), who found that the models with 4- and 2-km grid-spacing showed similar skill in forecasting heavy rainfall and convective evolution, but who noted that the 2-km forecasts generally contained more realistic-looking convective features than the 4-km forecasts.

The non-significant difference between the 4-km and 1-km deterministic forecasts appears to contradict the findings of Roberts and Lean (2008). However, as suggested in Schwartz et al. (2009), Roberts and Lean (2008)'s use of a modified convective parameterization scheme with their 4-km grid spacing model and their focus on time periods of 7 hours or less may help explain the differences in findings between that study and the present study. At greater than 7-hour time scales, for example, it is possible that large-scale errors may become more important, thus rendering the 4-km

and 1-km deterministic forecasts similar. Therefore, it is possible that a significant difference in AUC between the 4- and 1-km deterministic forecasts would exist at shorter forecast lead times while no significant difference in AUC exists for next-day forecasts.

The present study's finding of a weakly significant difference between the 4-km ensemble and 4-km deterministic forecasts generally agrees with the results of Sobash et al. (2016b), who found that SSPFs produced from a 30-member ensemble had significantly greater fractions skill scores (FSSs) compared to the SSPFs produced from either of two deterministic forecasts for smaller (i.e., mesoscale) spatial scales. While the present study does not analyze FSS at a variety of spatial scales, subjective inspection of the present study's individual case studies suggests that—despite some noteworthy exceptions—many of the differences between the 4-km deterministic, 1-km deterministic, and 4-km ensemble forecasts occur on the mesoscale. One potential reason for the ensemble forecasts' superior performance relative to the 4-km deterministic forecasts could be the two types of smoothing used to create the ensemble forecasts (i.e., spatial and ensemble smoothing) compared to just the spatial smoothing used to create the deterministic forecasts. Indeed, when no spatial smoothing is applied to the ensemble probabilities, the ensemble AUC is reduced to, approximately, 0.832 over the entire dataset (not shown), compared to an AUC of 0.874 produced by the σ = 90 km ensemble. Sensitivity tests (not shown) reveal that the AUC is maximized for the σ = 90 km ensemble forecasts.

To demonstrate the range of day-to-day variation in skill between the three sets of forecasts, five days were selected for individual case study analysis: 11 May 2010, 15

June 2010, 7 June 2011, 18 May 2011, and 27 April 2011. In general, it is found that the 4- and 1-km deterministic forecasts exhibit similar quality (as measured by individual-day AUC) to each other, while the 4-km ensemble forecasts routinely provide enhanced quality relative to either deterministic forecast. Notably, even on days when the ensemble forecast is inferior, the quality of the ensemble forecast tends to remain close to that of the deterministic forecasts. Interestingly, it is found that neither the number of daily SPC storm reports nor the SPC 1200 UTC day-1 convective outlook categories serve as good predictors for determining whether the 1-km deterministic or 4-km ensemble forecast will outperform the 4-km deterministic forecast on a given day (not shown).

Herein, only a single ensemble configuration is used to create the ensemble forecasts. While an ensemble with more members could have been used, previous research has found that increasing the number of members in an ensemble provides diminishing returns to the ensemble's skill (e.g., Clark et al. 2011, Sobash et al. 2016b, Schwartz et al. 2014), suggesting that adding members to the 11-member ensemble may not significantly improve forecast skill. It is more difficult to predict how the presence of multiple dynamic cores and/or multiple physics parameterizations influences the skill of an ensemble. Previous research has suggested that multi-core and multi-physics ensembles generally have enhanced spread and forecast skill relative to single-core and single-physics ensembles, respectively (e.g., Eckel and Mass 2005, Berner et al. 2011). Therefore, it is possible that a multi-core, multi-physics ensemble could perform better than the ensemble used herein relative to the two deterministic models, while a single-core, single-physics ensemble could perform worse. However, these results are certainly

not guaranteed; increasing an ensemble's spread by introducing members with multiple cores or physics parameterizations does not necessarily improve the ensemble's forecast skill or reliability, especially if the ensemble is not properly calibrated for bias (Eckel and Mass 2005). Moreover, the benefits of a mixed- vs. single-physics ensemble may depend on situational factors, such as the amount of large-scale forcing for ascent (Stensrud et al. 2000), or potentially even on the time and spatial scales of the forecast. More work is needed to determine the optimal configuration of CAM ensembles.

5. Future work

Given the abundance of remaining questions, many potential avenues exist for future work. One such avenue, as mentioned above, is to investigate how the configuration of the ensemble (e.g., multi- vs. single-core and mixed- vs. single-physics) influences the ensemble's skill relative to either deterministic forecast. Another path is to determine how specific regimes, mesoscale scenarios (e.g., dominant convective mode, convective trigger, etc.) and/or seasons influence the relative skill of the three forecast configurations. Such knowledge would be invaluable, as it could focus forecasters' attention on the forecast model(s) most likely to deliver the greatest quality for a given situation. Future work may additionally wish to examine the three forecast sets' relative skill at varying lead times and/or for other types of forecasts, such as those involving precipitation or specific severe weather hazards. Finally, future work may investigate more complex metrics for diagnosing the probability of next-day severe weather events. For instance, UH may be combined with other parameters to create a more skillful next-day forecast (e.g., Gallo et al. 2016), or a form of UH other than 2-5

km UH may be tested (e.g., Sobash et al. 2016a). It is possible that these other metrics may produce higher-quality forecasts, which in turn could alter the relative quality and value of the 1-km deterministic, 4-km deterministic, and 4-km ensemble forecasts.

Acknowledgements

This work was made possible by a Presidential Early Career Award for Scientists and Engineers (PECASE). Additional support was provided by NOAA/Office of Oceanic and Atmospheric Research under NOAA-University of Oklahoma Cooperative Agreement NA11OAR4320072, U.S. Department of Commerce. In addition to NOAA CSTAR support, CAPS simulations received supplementary support from NSF Grant ATM-0802888. All 2010 and 2011 CAPS forecasts were produced at the National Institute of Computational Sciences (NICS) at the University of Tennessee, and Oklahoma Supercomputing Center for Research and Education resources were used for ensemble post-processing. The authors would also like to thank the three anonymous reviewers, whose insightful comments and feedback helped dramatically improve the manuscript.

Month	2010	2011
April	28, 29, 30	27-29
May	03-07, 10-14, 17-21, 24-28, 31	04, 09, 12-13, 18-20, 22-28, 31
June	01-04, 07-11, 14-18	01, 03, 06-10

Table 2.1 Dates from the 2010-2011 NOAA HWT SFEs included in the dataset (63 total dates).

Ensemble member/ Model	ICs	LBCs	Micro-physics	Land surface model	Boundary layer
arw_cn (4-km)[@*]	0000 UTC ARPSa	0000 UTC NAMf	Thompson	Noah	MYJ
arw_cn (1-km)[@]	0000 UTC ARPSa	0000 UTC NAMf	Thompson	Noah	MYJ
arw_m5[#]	arw_cn + em-p1 + recur pert	em-p1	Morrison	RUC	YSU
arw_m6[*]	arw_cn + em-p1_pert	em-p1	Morrison	RUC	YSU
arw_m7[*]	arw_cn + em-p2_pert	em-p2	Thompson	Noah	QNSE
arw_m8[*]	arw_cn – nmm-p1_pert	nmm-p1	WSM6	RUC	QNSE
arw_m9[*]	arw_cn + nmm-p2_pert	nmm-p2	WDM6	Noah	MYNN
arw_m10[*]	arw_cn + rsmSAS-n1_pert	rsmSAS-n1	Ferrier	RUC	YSU
arw_m11[*]	arw_cn – etaKF-n1_pert	etaKF-n1	Ferrier	Noah	YSU
arw_m12[**]	arw_cn + etaKF-p1_pert	etaKF-p1	WDM6	RUC (2010)/ Noah (2011)	QNSE
arw_m13 (2010)[**]	arw_cn – etaBMJ-n1_pert	etaBMJ-n1	WSM6	Noah (2010)/ RUC (2011)	MYNN
arw_m14[*]	arw_cn + etaBMJ-p1_pert	etaBMJ-p1	Thompson	RUC	MYNN
arw_m13 (2011)[&]	arw_cn + rsm-p1_pert	rsm-p1	M-Y	Noah	MYJ

Table 2.2 Deterministic model and ensemble member specifications. An asperand (@) denotes deterministic models used for both 2010 and 2011. A single asterisk (*) denotes ensemble members that were part of both the 2010 and 2011 ensembles. A double asterisk (**) denotes ensemble members that had different land surface models for 2010 and 2011 but were otherwise the same for both years. A pound sign (#) denotes ensemble members that were part of the 2010 ensemble only, while an ampersand (&) denotes ensemble members that were part of the 2011 ensemble only. NAMf refers to the 12-km NAM forecast, and ARPSa refers to the Advanced Regional Prediction System three-dimensional variational data assimilation (Xue et al. 2003; Gao et al. 2004). Elements in the ICs column followed by a "+" or "-" denote SREF perturbations added or subtracted from the ICs of the arw_cn member. Ensemble member boundary layer schemes included: Mellow-Yamada-Janjić (MYJ; Mellor and Yamada 1982; Janjić 2002), Yonsei University (YSU; Noh et al. 2003), Mellor-Yamada-Nakanishi-Niino (MYNN; Nakanishi 2000, 2001; Nakanishi and Niino 2004, 2006), and quasi-normal scale elimination (QNSE; Sukoriansky et al. 2006). Ensemble member microphysics schemes included: Thompson et al. (2004), WRF single-moment 6-class (WSM6; Hong and Lim 2006), WRF double-moment 6-class (WDM6; Lim and Hong 2010), Ferrier et al. (2002), Milbrandt and Yau (2005; M-Y), and Morrison et al. (2005). All ensemble members used the Rapid Radiative Transfer Model longwave radiation scheme (RRTM; Mlawer et al. 1997) and the Goddard shortwave radiation scheme (Chou and Suarez 1994). Land surface models included the Noah (Chen and Dudhia 2001) and RUC (Smirnova et al. 1997, 2000).

4-km UH (m^2s^{-2})	Percentile	1-km UH (m^2s^{-2})
25	0.9821775 (2010)	248.8906 (2010)
	0.9811282 (2011)	280.6609 (2011)
50	0.9930652 (2010)	467.4218 (2010)
	0.9927213 (2011)	505.1666 (2011)
75	0.9967307 (2010)	670.5211 (2010)
	0.9967047 (2011)	718.7805 (2011)
100	0.9985068 (2010)	893.3596 (2010)
	0.9983788 (2011)	935.0260 (2011)
125	0.9993252 (2010)	1158.1796 (2010)
	0.9990868 (2011)	1098.2116 (2011)

Table 2.3 4-km and 1-km equivalent UH threshold values. 2010 percentile and 1-km UH values are located above the corresponding 2011 percentile and 1-km UH values.

UH Threshold (m^2s^{-2})	1-km AUC	4-km AUC	(1-km AUC) – (4-km AUC) Difference	(1-km AUC) – (4-km AUC) 2.5 Percentile, 97.5 percentile
25	0.860	0.838	0.022	-0.0326, 0.0342
50	0.782	0.775	0.007	-0.0487, 0.0516
75	0.715	0.708	0.007	-0.0630, 0.0552
100	0.655	0.651	0.004	-0.0585, 0.0559
125	0.608	0.602	0.006	-0.0537, 0.0494

Table 2.4 Results from the 2-sided resampling hypothesis test between the 1-km and 4-km deterministic forecasts. None of the (1-km AUC) – (4-km AUC) differences fall outside of the range given in the final column, indicating that none of the differences are significant at the 95% level.

Model/ Ensemble	AUC	(4-km Det. AUC) – (Model/Ens. AUC) Difference	(4-km Det. AUC) – (Model/Ens. AUC) 2.5 percentile, 97.5 percentile
4-km Ens., Sigma = 60 km	0.872	-0.0340	-0.0363, 0.0345
4-km Ens., Sigma = 90 km	0.874	-0.0360*	-0.0359, 0.0340
4-km Ens., Sigma = 120 km	0.872	-0.0340*	-0.0335, 0.0357
4-km Det.	0.838	N/A	N/A

Table 2.5 Results from the 2-sided resampling hypothesis test between the 4-km ensemble and the 4-km deterministic forecasts for a UH threshold of 25 m^2s^{-2}. An asterisk (*) denotes significance at the 95% level.

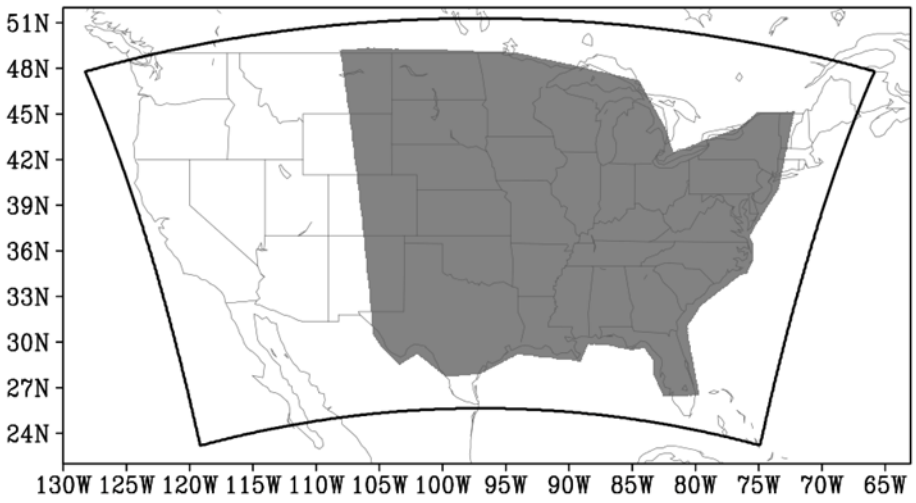

Figure 2.1 Model domain (black contour) and analysis domain (gray shading).

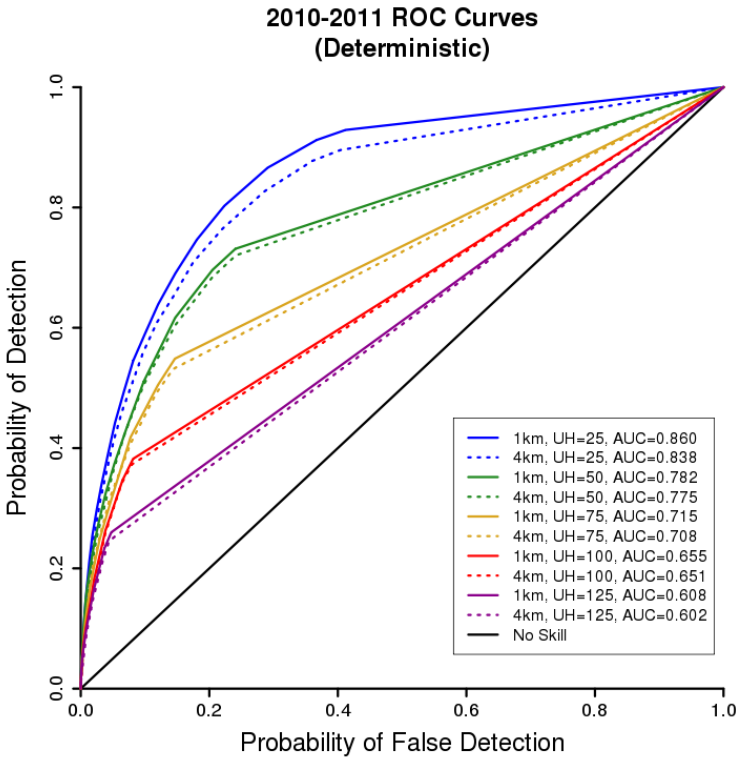

Figure 2.2 Relative operating characteristic curves for the 1-km (solid) and 4-km (dashed) deterministic models for UH threshold values corresponding to 25 m^2s^{-2} (blue), 50 m^2s^{-2} (green), 75 m^2s^{-2} (goldenrod), 100 m^2s^{-2} (red), and 125 m^2s^{-2} (violet) on the 4-km grid. Here, as in the text, the 4-km UH threshold values are used to additionally refer to the corresponding 1-km UH values to simplify the analysis. The solid black line indicates the no-skill line.

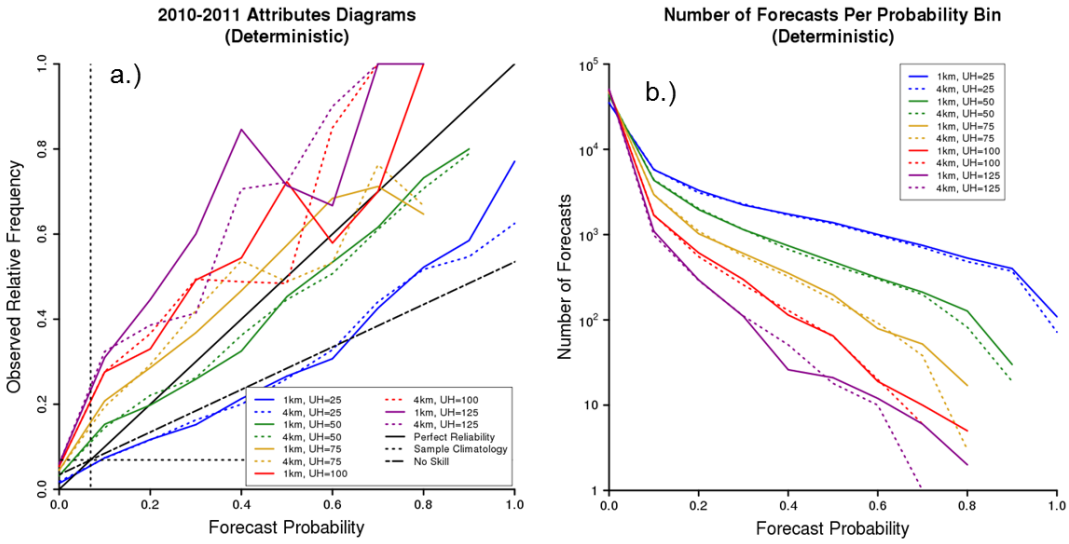

Figure 2.3 (a) Attributes diagrams for the 1-km (solid) and 4-km (dashed) deterministic models for UH threshold values corresponding to 25 m^2s^{-2} (blue), 50 m^2s^{-2} (green), 75 m^2s^{-2} (goldenrod), 100 m^2s^{-2} (red), and 125 m^2s^{-2} (violet) on the 4-km grid. Here, as in the text, the 4-km UH threshold values are used to additionally refer to the corresponding 1-km UH values to simplify the analysis. The solid black line indicates the line of perfect reliability, the long dashed line indicates the no-skill line, and the short dashed lines represent sample climatological frequency (abbreviated as sample climatology). (b) Number of forecasts per forecast probability bin for the 1-km (solid) and 4-km (dashed) deterministic models. The colors represent the same UH thresholds as in (a). Note the logarithmic y-axis.

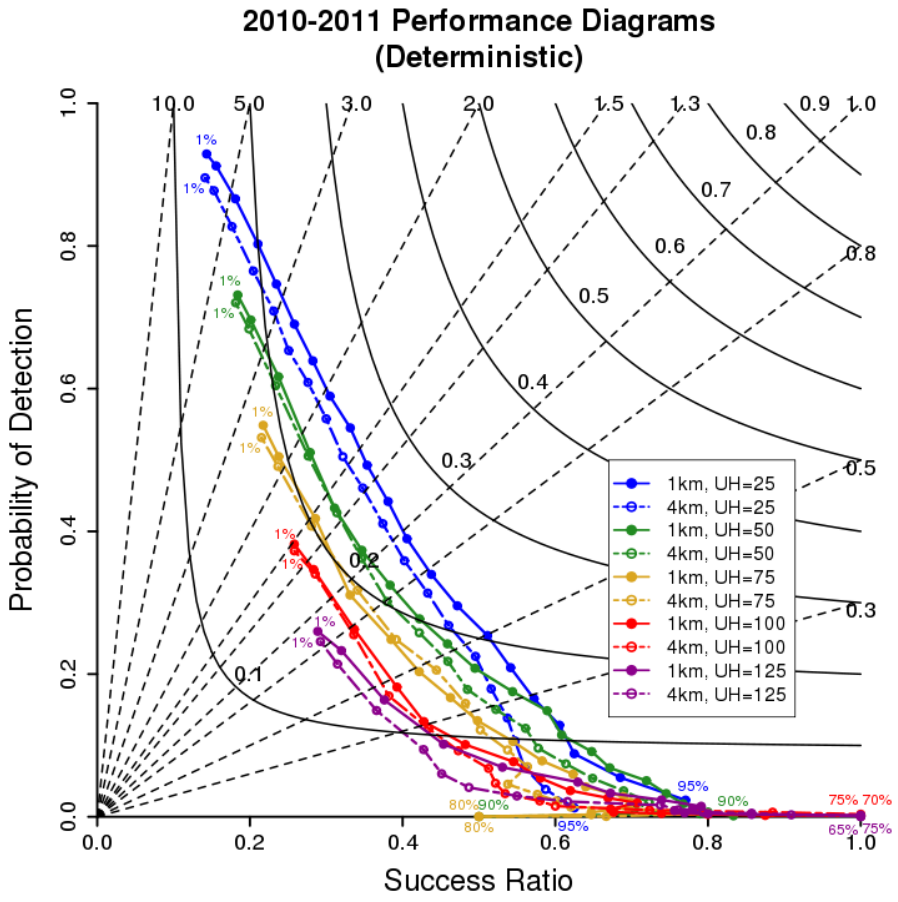

Figure 2.4 Performance diagrams for 1-km (solid lines with filled points) and 4-km (dashed lines with open points) deterministic models for UH threshold values corresponding to 25 m^2s^{-2} (blue), 50 m^2s^{-2} (green), 75 m^2s^{-2} (goldenrod), 100 m^2s^{-2} (red), and 125 m^2s^{-2} (violet) on the 4-km grid. Here, as in the text, the 4-km UH threshold values are used to additionally refer to the corresponding 1-km UH values. For the 1- and 4-km UH = 25 m^2s^{-2} forecasts, the following 21 probability levels are plotted: 1%, 2%, 5%, 10%, 15%, 20%, 25%, 30%, 35%, 40%, 45%, 50%, 55%, 60%, 65%, 70%, 75%, 80%, 85%, 90%, and 95%. A subset of these probability levels are plotted for the remaining 8 forecasts, since these forecasts never produce 95% severe probabilities. The first and last probability level is labeled for each forecast. Solid black lines indicate lines of constant CSI, while dashed black lines represent lines of constant bias.

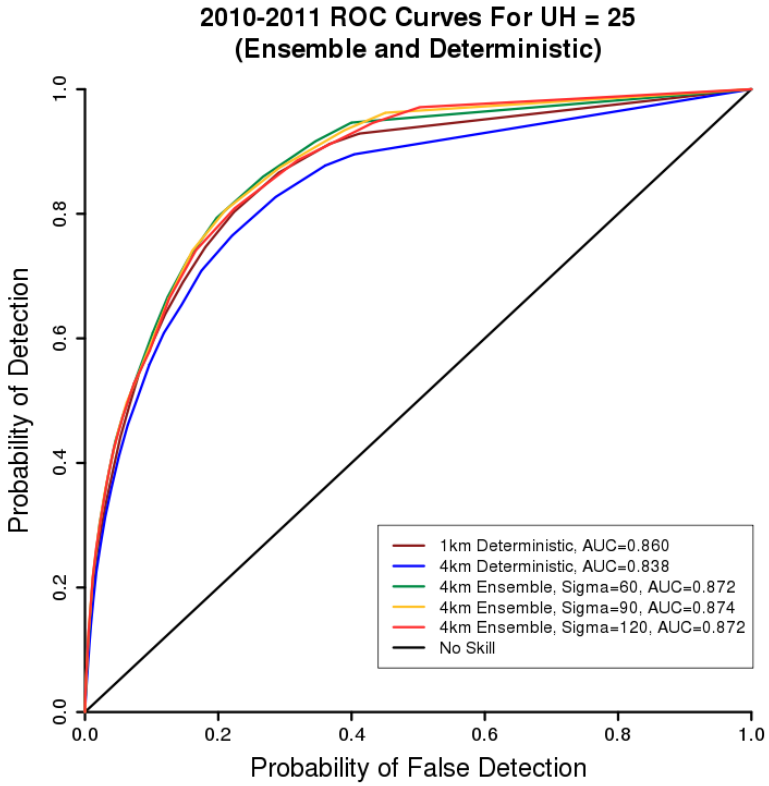

Figure 2.5 Relative operating characteristic curves for 1-km deterministic (dark red), 4-km deterministic (blue), σ = 60-km 4-km ensemble (green), σ = 90-km 4-km ensemble (goldenrod), and σ = 120-km 4-km ensemble (red) forecasts. All forecasts use the UH threshold value corresponding to 25 m^2s^{-2} on the 4-km grid. The solid black line indicates the no skill line.

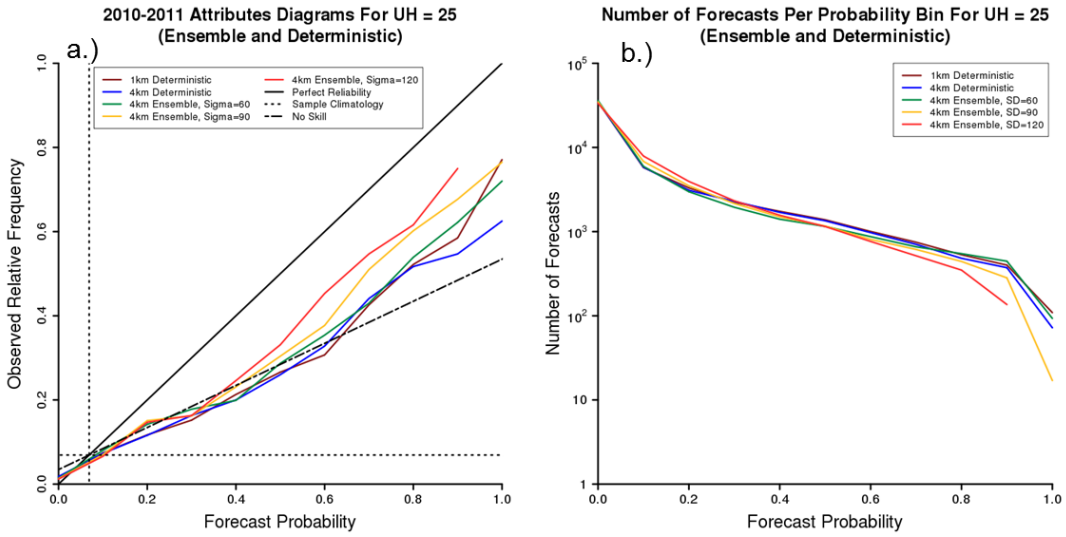

Figure 2.6 (a) Attributes diagrams for 1-km deterministic (dark red), 4-km deterministic (blue), σ = 60-km 4-km ensemble (green), σ = 90-km 4-km ensemble (goldenrod), and σ = 120-km 4-km ensemble (red) forecasts. All forecasts use the UH threshold value corresponding to 25 m^2s^{-2} on the 4-km grid. The solid black line indicates the line of perfect reliability, the long dashed line indicates the no-skill line, and the short dashed lines represent sample climatological frequency (abbreviated as sample climatology). (b) Number of forecasts per forecast probability bin for 1-km deterministic (dark red), 4-km deterministic (blue), σ = 60-km 4-km ensemble (green), σ = 90-km 4-km ensemble (goldenrod), and σ = 120-km 4-km ensemble (red) forecasts. Note the logarithmic y-axis.

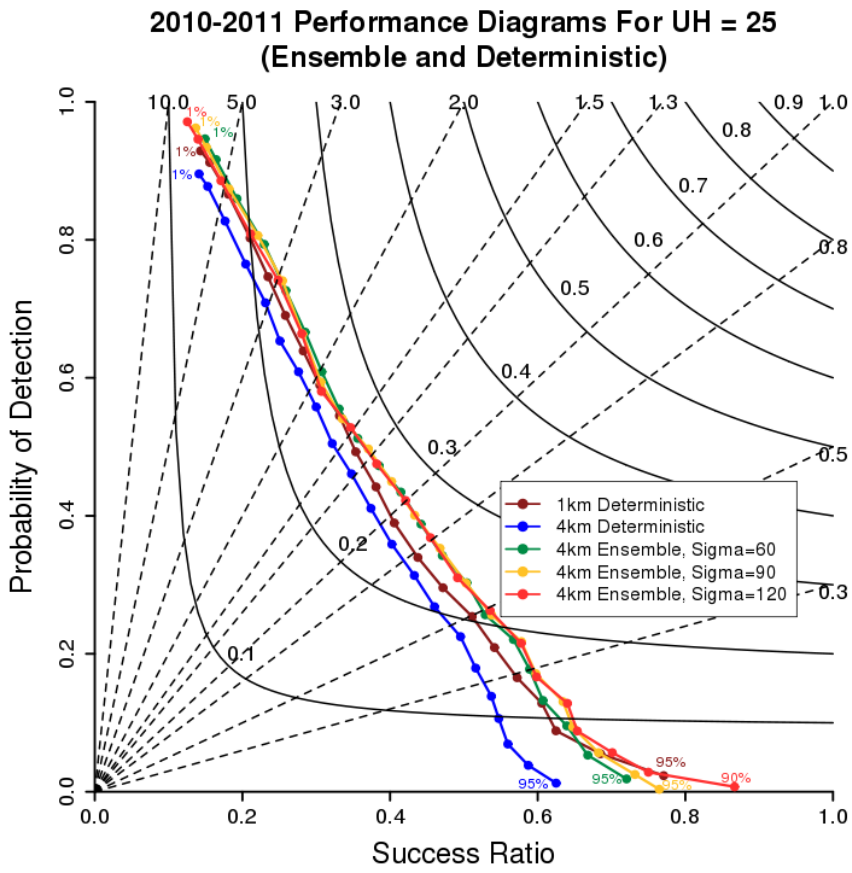

Figure 2.7 Performance diagrams for 1-km deterministic (dark red), 4-km deterministic (blue), σ = 60-km 4-km ensemble (green), σ = 90-km 4-km ensemble (goldenrod), and σ = 120-km 4-km ensemble (red) forecasts. All forecasts use the UH threshold value corresponding to 25 m^2s^{-2} on the 4-km grid. Except for the σ = 120-km 4-km ensemble forecasts, which produced no 95% or greater probabilities, each of the five forecasts have the following 21 probability levels plotted: 1%, 2%, 5%, 10%, 15%, 20%, 25%, 30%, 35%, 40%, 45%, 50%, 55%, 60%, 65%, 70%, 75%, 80%, 85%, 90%, and 95%. The first and last probability level is labeled for each forecast. Solid black lines indicate lines of constant CSI, while dashed black lines represent lines of constant bias.

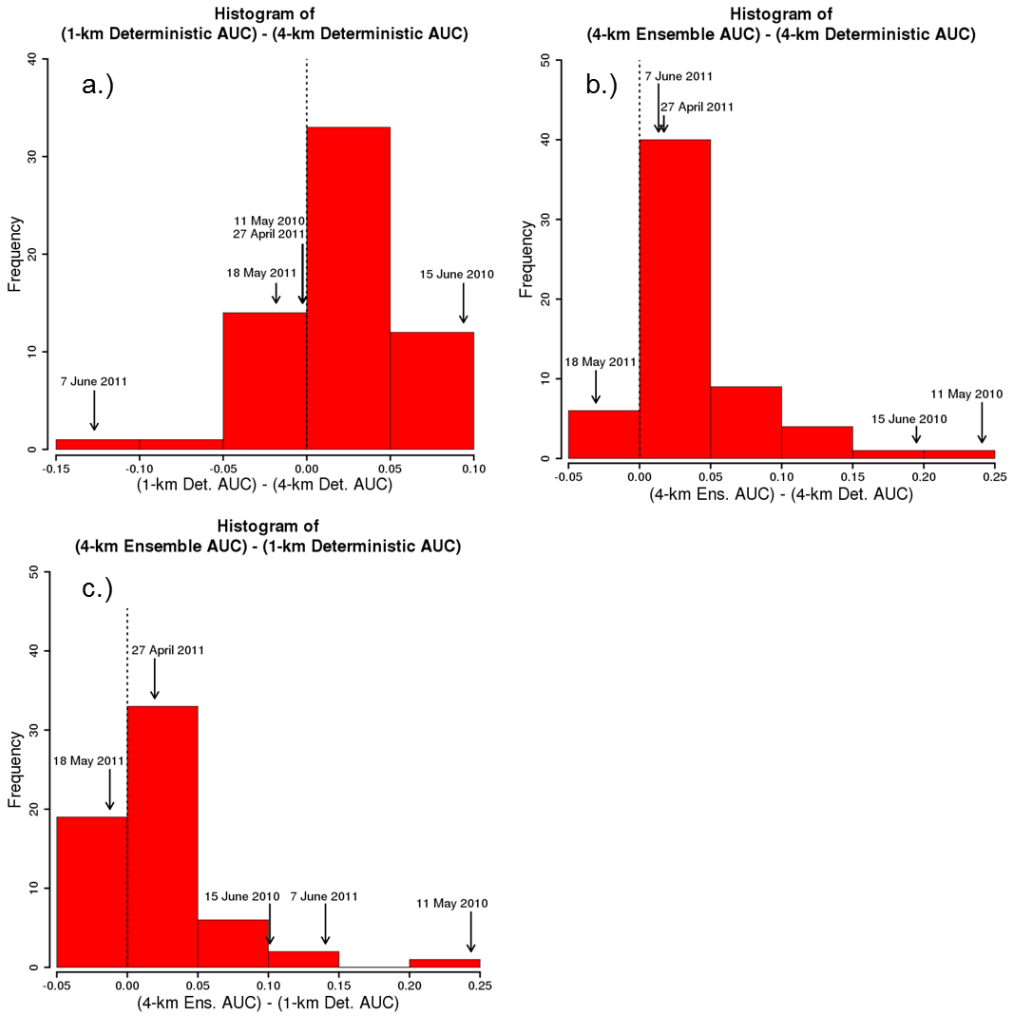

Figure 2.8 (a) Histogram showing the distribution of 1-km deterministic AUC – 4-km deterministic AUC for individual days. Positive (negative) values on the x-axis indicate days on which the 1-km deterministic forecast had a greater (lower) AUC than the 4-km deterministic forecast. The labeled arrows indicate where the five case study days fall in the distribution. All forecasts use the UH threshold corresponding to 25 m^2s^{-2} on the 4-km grid. (b) Histogram showing the distribution of 4-km ensemble (σ = 90-km) AUC – 4-km deterministic AUC for individual days. Positive (negative) values on the x-axis indicate days on which the 4-km ensemble forecast had a greater (lower) AUC than the 4-km deterministic forecast. The labeled arrows indicate where the five case study days fall in the distribution. All forecasts use the 25 m^2s^{-2} UH threshold. (c) Histogram showing the distribution of 4-km ensemble (σ = 90-km) AUC – 1-km deterministic AUC for individual days. Positive (negative) values on the x-axis indicate days on which the 4-km ensemble forecast had a greater (lower) AUC than the 1-km deterministic forecast. The labeled arrows indicate where the five case study days fall in the distribution. All forecasts use the 25 m^2s^{-2} UH threshold.

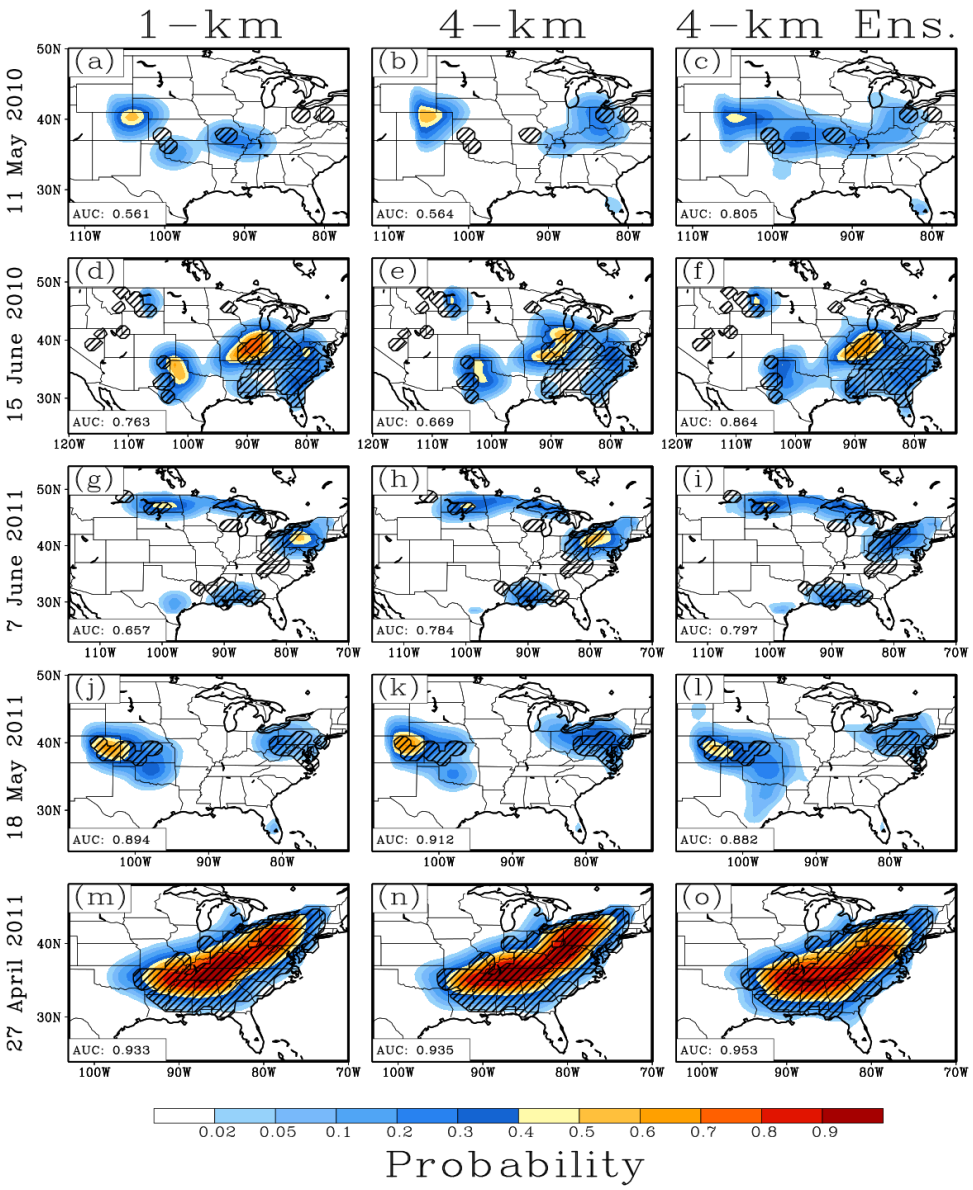

Figure 2.9 Probabilistic severe weather forecasts (shaded) for the (a) 1-km deterministic forecast, the (b) 4-km deterministic forecast, and the (c) 4-km ensemble forecast (σ = 90-km) for 11 May 2010. Black hatching denotes 80-km grid boxes that contain at least one observed storm report. All forecasts use the UH threshold value corresponding to 25 m^2s^{-2} on the 4-km grid. (d)-(f), (g)-(i), (j)-(l), and (m)-(o) same as (a)-(c) but for 15 June 2010, 7 June 2011, 18 May 2011, and 27 April 2011, respectively.

Chapter 3: Spread and Skill in Mixed- and Single-Physics Convection Allowing Ensembles at Different Spatial Scales

Eric D. Loken[1,2,4], Adam J. Clark[4], Ming Xue[2,3], Fanyou Kong[3]

[1]Cooperative Institute for Mesoscale Meteorological Studies, The University of Oklahoma, Norman, Oklahoma
[2]School of Meteorology, University of Oklahoma, Norman, Oklahoma
[3]Center for Analysis and Prediction of Storms, The University of Oklahoma, Norman, Oklahoma
[4]NOAA/OAR National Severe Storms Laboratory, Norman, Oklahoma

Chapter Introduction

In Chapter 2, next-day probabilistic severe weather forecasts derived from an 11-member, 4-km grid-spacing WRF-ARW ensemble were found to be significantly different than those derived from a 4-km WRF-ARW deterministic model in terms of aggregate AUC over 63 days from the 2010-2011 Spring Forecasting Experiments. Notably, the members of the ensemble in Chapter 2 used one common dynamic core, perturbed initial and lateral boundary conditions (ICs/LBCs), and multiple microphysics and boundary layer parameterizations. An important question is how sensitive the results obtained in Chapter 2 are to the configuration of the ensemble used. Indeed, this question merits some brief discussion, as it will help motivate the study and results presented later in this chapter.

At convection-parameterizing resolution, multi-model ensembles have generally been found to enhance forecast spread and skill relative to single-model ensembles (e.g., Eckel and Mass 2005, Stensrud et al. 2000, Wandishin et al. 2001). Multi-model ensembles may provide advantages at convection-allowing resolution as well. For example, when comparing subsets from the Community Leveraged Unified Ensemble (CLUE) during the 2016 HWT SFE, Jirak et al. (2016) found that the 7-member multi-

core Storm Scale Ensemble of Opportunity (SSEO) and a 10-member multi-core ensemble generally had the two greatest fractions skill scores (FSSs) and areas under the relative operating characteristics curve (AUCs) for probabilistic forecasts of ≥ 40 dBZ 1-km above-ground level (AGL) reflectivity. Relative to the multi-core ensemble subsets, the single-core subsets had lower AUCs and generally lower FSSs, especially during hours of peak convection in the afternoon (i.e., 1900-2300 UTC; Jirak et al. 2016). Additionally, Johnson and Wang (2012), who compared single- and multi-model ensembles during the 2009 HWT SFE, found that a multi-model convection-allowing ensemble could provide an advantage over a single-model ensemble for neighborhood forecasts at lead times greater than 24 hours. Collectively, these findings suggest that, had a multi- instead of a single-model ensemble been used in Chapter 2, the skill of the ensemble may have been increased, resulting in a larger difference between the ensemble and 4-km deterministic configuration forecasts.

The results obtained in Chapter 2 may also be sensitive to the ensemble's use of perturbed vs. unperturbed ICs/LBCs. For example, Kong et al. (2014) found that a convection-allowing ensemble with perturbed ICs/LBCs and mixed physics (i.e., multiple microphysics and boundary layer parameterizations) generated substantially greater spread compared to a convection-allowing ensemble using unperturbed ICs/LBCs and mixed physics. Similarly, Clark et al. (2010) found that, in a 4-km grid-spacing convection-allowing ensemble, perturbed ICs/LBCs generally accounted for a large portion of the forecast variance for most forecast fields studied. Given these findings, it is likely that using an ensemble with *unperturbed* ICs/LBCs in Chapter 2 would reduce ensemble spread, making the forecast from each ensemble member more

similar to that of the ensemble's control member. Since the control member of the ensemble also serves as the 4-km deterministic model configuration, it is likely that using unperturbed ICs/LBCs would reduce the difference between the ensemble and 4-km deterministic configuration forecasts in Chapter 2.

The impact of using a single- instead of a mixed-physics convection-allowing ensemble is more uncertain. While previous research at both convection-parameterizing and convection-allowing resolution has found that mixed-physics ensembles produce more forecast spread and skill for fields related to convection (e.g., precipitation and simulated low-level reflectivity) compared to single-physics ensembles (e.g., Stensrud et al. 2000; Duda et al. 2014; Johnson et al. 2017), the benefits of using a mixed- over a single-physics ensemble may depend on situational factors, including the amount of large-scale forcing for ascent (Stensrud et al. 2000) and possibly the spatial scale of the forecast. For nocturnal convection events, Johnson and Wang (2017) found relatively small—although significant—differences in forecast spread and skill between mixed- and single-physics convection-allowing ensemble configurations. Subjectively, these differences were often subtle and were manifest in a variety of ways. For example, depending on the case, the mixed-physics configurations could have more members predicting nocturnal convective initiation events, different convective structures of initiating storms, and/or greater areal coverage of convection compared to the single-physics configuration. Despite these differences, the mixed- and single-physics configurations generally forecast convection over the same regions. It is therefore unclear how the use of a mixed- or single-physics configuration would impact the results obtained in Chapter 2. In a broader context, it is unclear how mixed- and single-

physics ensemble configurations compare in terms of forecast spread and skill at a variety of spatial scales and forecast hours. This question is addressed in the following chapter for multiple forecast fields, including: 2-m temperature, 2-m dewpoint temperature, 500-mb geopotential height, and hourly and 6-hourly accumulated precipitation.

Abstract

Spread and skill of mixed- and single-physics convection-allowing ensemble forecasts are investigated at a variety of spatial scales. Forecast spread is assessed for 2-m temperature, 2-m dewpoint, 500-mb geopotential height, and hourly accumulated precipitation both before and after a bias-correction procedure is applied. Time series indicate that the mixed-physics ensemble forecasts generally have greater variance than comparable single-physics forecasts. While the differences tend to be relatively small, they are greatest at the smallest spatial scales and when the ensembles are not calibrated for bias. Interestingly, while *differences* between the mixed- and single-physics ensemble variances are smaller for the larger spatial scales, variance *ratios* suggest that the mixed-physics ensemble generates more spread relative to the single-physics ensemble at larger spatial scales. Given that, at larger spatial scales, the variance values are quite small, variance differences may be more appropriate to consider than ratios.

Forecast quality is evaluated for hourly and 6-hourly accumulated precipitation using mean square error, fractions skill score, area under the relative operating characteristic curve, and attributes diagrams. Generally, little difference in skill is found between the mixed- and single-physics forecasts. The greatest differences arise for the

largest precipitation thresholds and during the late afternoon and evening, when precipitation is maximized climatologically, suggesting the mixed-physics ensemble may provide the greatest relative benefit in situations when moderate/heavy precipitation is more likely.

Overall, given that mixed- and single-physics ensembles have similar spread and skill, developers may prefer to implement single- as opposed to mixed-physics convection-allowing ensembles in operations.

1. Introduction

Over the past decade, advances in computing power have enabled numerical weather prediction (NWP) forecasts from fine-resolution convection-allowing ensembles. As early as 2007, the Center for Analysis and Prediction of Storms (CAPS) began running an experimental 10-member, 48-hour ensemble with 4-km grid spacing over the contiguous United States (CONUS) to facilitate the prediction of severe weather during the 2007 NOAA Hazardous Weather Testbed Spring Forecasting Experiment (HWT SFE; Xue et al. 2007). This convection-allowing ensemble produced skillful and useful forecasts of composite reflectivity, accumulated precipitation, and probability of precipitation (Xue et al. 2007; Schwartz et al. 2010; Clark et al. 2009). More recent HWT SFEs have evaluated other convection-allowing ensembles, including the 7-member Storm Scale Ensemble of Opportunity (SSEO; Jirak et al. 2012), as well as various applications of convection-allowing ensembles, including their use to create probabilistic all-hazards severe weather forecast guidance (Kain et al. 2008; Sobash et al. 2011), tornado pathlength forecasts (Clark et al. 2013) and

probabilistic tornado forecasts (Gallo et al. 2016). Outside of the HWT SFEs, convection-allowing ensemble systems, such as the National Center for Atmospheric Research Ensemble Prediction System (NCAR EPS; Schwartz et al. 2015a), are being run experimentally and evaluated for use in operations.

In general, ensembles can offer benefits over deterministic models because they account for uncertainties in initial conditions (ICs) and model physics (e.g., Roebber et al. 2004; Leutbecher and Palmer 2008; Clark et al. 2009). However, convection-allowing ensembles show unique promise because they not only account for these uncertainties, but each of their members—by virtue of their fine grid spacing—is able to explicitly simulate convection, which has been shown to result in better predictions of convective mode and evolution (e.g., Kain et al. 2006; Done et al. 2004). Indeed, while it has long been known that ensemble mean forecasts tend to outperform forecasts from similarly-configured deterministic models at convection-parameterizing resolution (e.g., Epstein 1969a; Leith 1974; Clark et al. 2009), recent evidence suggests that convection-allowing ensembles tend to outperform deterministic models at convection-allowing resolution as well (e.g., Coniglio et al. 2010, Loken et al., 2017; Schwartz et al., 2017).

Despite the promise of convection-allowing ensembles, much is still unknown about their optimal configuration (e.g., Roebber et al. 2004; Romine et al. 2014; Duda et al. 2014; Johnson and Wang 2017). One problem is that the vast majority of convection-allowing ensembles are under-dispersive (i.e., observed events routinely fall outside of the forecast probability density function (PDF)), especially for precipitation fields (e.g., Clark et al. 2008, 2010; Romine et al. 2014). Many previous studies have investigated methods to increase ensemble spread at convective-parameterizing

resolutions, including perturbing initial conditions (e.g., Toth and Kalnay 1993, 1997; Molteni 1996) and using multiple models (e.g., Wandishin et al. 2001; Hou et al. 2001; Ebert 2001; Eckel and Mass 2005) and physics parameterizations (e.g., Stensrud et al. 2000; Gallus and Bresch 2006). More recent work has studied the impact of incorporating multiple planetary boundary layer (PBL) and/or microphysics schemes within convection-allowing ensembles (e.g., Schwartz et al. 2010; Duda et al. 2014; Johnson and Wang 2017), generally finding that mixed-microphysics and mixed-PBL ensembles result in improved ensemble spread and skill. For example, during the 2015 Plains Elevated Convection at Night (PECAN) experiment, Johnson and Wang (2017) found that both of two mixed-physics convection-allowing ensembles—which used a variety of microphysics and PBL schemes—produced better nocturnal precipitation and non-precipitation forecasts compared to a single-physics ensemble, which used just Thompson microphysics (Thompson et al. 2004) and the Mellor-Yamada-Nakanishi-Niino (MYNN; Nakanishi 2000, 2001; Nakanishi and Niino 2004, 2006) PBL. The mixed-physics ensembles in Johnson and Wang (2017) generally produced better subjective forecasts of nocturnal convection as well: relative to the single-physics ensemble, they reduced nocturnal mesoscale convective system (MCS) location errors, produced improved storm structures in nocturnal initiating convection, and had more members forecast observed nocturnal convective initiation. That multiple microphysics and PBL parameterizations can improve forecasts related to convection is unsurprising; previous research has found simulated thunderstorms to be quite sensitive to microphysics parameterizations (e.g., Gilmore et al. 2004; van den Heever and Cotton 2004; Snook and Xue 2008). However, it is currently unknown—especially for

convective-allowing ensembles—whether the benefits of using multiple microphysics and PBL parameterizations are apparent only at relatively small spatial scales. Given that larger spatial scales are associated with greater predictability (Lorenz 1969), it is possible that accounting for the uncertainties in modeled microphysics and PBL may matter less for larger spatial scales, where predictability is already relatively high. For example, it is possible that, while a mixed-physics ensemble improves the precise placement of forecast convective systems and produces better forecasts of storm structure, the overall forecasts (e.g., the general location of forecast precipitation-producing systems) provided by a mixed- and single-physics convection-allowing ensemble may not be drastically different at synoptic (or larger meso-) scales. It is also possible that the relative benefits (i.e., superior forecast spread and skill) of using multiple microphysics and PBL parameterizations may depend on the variable of interest (e.g., mass-related or low-level variables, Clark et al. 2010) and/or forecast hour/time of day. Given that ensembles with only one microphysics and one PBL scheme are easier for model developers to maintain, it is important to determine if and when a single-physics convection-allowing ensemble can perform nearly as well as a mixed-physics ensemble.

For this task, the present study uses data from the 2016 Community Leveraged Unified Ensemble (CLUE; Clark et al. 2016), a collection of 65 ensemble members with similar specifications and post-processing methods contributed by a variety of organizations (e.g., the National Severe Storms Laboratory (NSSL), the Center for Analysis and Prediction of Storms (CAPS), the University of North Dakota, NOAA's Earth Systems Research Laboratory/Global Systems Division (ESRL/GSD), and the

National Center for Atmospheric Research (NCAR)) during the 2016 HWT SFE. Forecast spread (i.e., variance) is analyzed for 2-m temperature, 2-m dewpoint temperature, 500-mb height, and hourly accumulated precipitation at a variety of spatial scales; forecast skill is evaluated for hourly and 6-hourly accumulated precipitation. Up to 36-hour forecasts are considered.

The remainder of this paper is organized as follows: section 2 details the methods used, section 3 presents the results, section 4 summarizes and discusses the results, and section 5 concludes the paper by considering implications for ensemble design and offering suggestions for future work.

2. Methods

(a) Dataset

The 65-member CLUE was run for 24 days during the 2016 NOAA HWT SFE, which spanned from early May to early June. Herein, 36-hour forecast data from two 2016 CLUE subsets is analyzed for 23 days of the 2016 NOAA HWT SFE (Table 3.1; note that 24 May 2016 is excluded from analysis since not all members had data available on that day). The two ensemble subsets examined include a 9-member CAPS subset with multiple microphysics and PBL schemes (henceforth referred to as the mixed-physics ensemble) and a 10-member CAPS subset with only Thompson microphysics and the Mellor-Yamada-Janjic (MYJ) PBL scheme (henceforth referred to as the single-physics ensemble). All members from both ensemble subsets operate at 3-km horizontal grid spacing over a domain covering the contiguous U.S. (Fig. 3.1); further, all members contain 1680 grid points in the east-west direction and 1152 grid points in the north-south direction, have perturbed initial and lateral boundary

conditions (ICs/LBCs), and use the Noah land surface model (Chen and Dudhia 2001) and the Advanced Research Weather Research and Forecasting dynamic core. Initialization for all members is done on weekdays using analyses from the 0000 UTC 12-km North American Model (NAM). Radar (WSR-88D) data and surface and upper-air observations are assimilated using the Advanced Regional Prediction System three-dimensional variational data assimilation and cloud analysis system (ARPS 3DVAR; Xue et al. 2003, Gao et al. 2004; Clark et al. 2016). Specifications for both ensemble subsets are summarized in Table 3.2. Notably, both the mixed- and single-physics ensembles use one common member (core01), since it is the control member of both subsets. The mixed-physics ensemble contains 9 members instead of 10 since data from core02 was unavailable throughout the analysis period.

(b) Evaluating ensemble spread

To determine ensemble spread, forecast variance is computed for four variables—2-m temperature, 2-m dewpoint temperature, 500-mb height and hourly accumulated precipitation—for forecast hours 0-36 using equation (B7) in Eckel and Mass (2005):

$$\text{Variance} = \frac{1}{\mathrm{M}} \sum_{\mathrm{m}=1}^{\mathrm{M}} \left[\frac{1}{(\mathrm{n}-1)} \sum_{\mathrm{i}=1}^{\mathrm{n}} (\mathrm{e}_{\mathrm{m,i}} - \bar{\mathrm{e}}_{\mathrm{m}})^2 \right] \qquad (1),$$

where M is the number of forecast-observation data pairs (which, here, includes the number of non-overlapping spatial windows in the domain over each of the 23 days in the analysis), n is the number of ensemble members, $\mathrm{e}_{\mathrm{m,i}}$ is the value of the *i*th ensemble member at *m*, and $\bar{e}_m$ is the ensemble mean at *m*. To assess the impact of

spatial scale, variance is calculated for square neighborhoods of varying sizes using the "upscaling" method (Ebert 2009), which assigns the mean of the finer-resolution grid boxes making up a given neighborhood to that neighborhood. In this study, 11 neighborhood sizes are analyzed. The neighborhoods contain: 1, 2, 4, 8, 16, 24, 48, 96, 144, 192, and 240 grid boxes per side. Since all ensemble members operate at 3-km horizontal grid spacing, the 11 neighborhoods measure, respectively, 3-, 6-, 12-, 24-, 48-, 72-, 144-, 288-, 432-, 576-, and 720-km on each side. Only neighborhoods falling completely within the analysis domain are included in the variance calculations, and the "upscale" averaging is done prior to computing the ensemble mean. The difference between the mixed- and single-physics ensemble variance (i.e., mixed-physics variance – single-physics variance) and the ratio of single-physics ensemble variance to mixed-physics ensemble variance (i.e., single-physics variance/mixed-physics variance) are also computed.

Because systematic biases from each ensemble member contribute to forecast spread but not to forecast uncertainty (since systematic biases are not uncertain; e.g., Eckel and Mass 2005; Clark et al. 2011; Clark et al. 2010), a probability matching technique (Ebert 2001; Clark et al. 2010) is used to eliminate systematic biases among the ensemble members. This technique assigns the probability distribution function (PDF) of one dataset to another dataset to eliminate bias. Herein, because the core01 member serves as the control member of both the mixed- and single-physics ensemble, the PDF of the core01 is assigned to each of the other ensemble members. Hence, after probability matching, all ensemble members contain the same bias (i.e., the bias of the core01 member). Unlike in Clark et al. (2010), the PDF of the observations is *not*

assigned to each ensemble member since the primary purpose of this portion of the study is to evaluate ensemble spread (as opposed to skill), and using the PDF of the core01 member—which is already appropriately gridded for analysis—is much more convenient than using an observational dataset. As done with the raw dataset, bias-corrected variance difference (i.e., bias-corrected mixed-physics variance – bias-corrected single-physics variance) and the ratio of bias-corrected single-physics ensemble variance to mixed-physics ensemble variance (i.e., bias-corrected single-physics variance/bias-corrected mixed-physics variance) are computed.

Bad data, manifest as an unphysical "spike" in variance, are noted at several forecast hours in both the raw (i.e., non-bias-corrected) and bias-corrected 2-m temperature and 2-m dewpoint temperature fields. In the raw dataset, bad variance values occur: in the 2-m temperature data at forecast hour 23 (for the mixed-physics ensemble) and forecast hour 33 (for both the mixed- and single-physics ensembles) and in the 2-m dewpoint temperature data at forecast hour 33 (for both the mixed- and single-physics ensembles). In the bias-corrected dataset, bad variance values occur: in the 2-m temperature data at forecast hour 28 (for both the mixed- and single-physics ensembles) and in the 2-m dewpoint data at forecast hour 24 (for the mixed-physics ensemble) and at forecast hour 33 (for both the mixed- and single-physics ensembles). In each case, the bad variance is replaced by the mean variance from the forecast hour immediately preceding and immediately following the bad data.

(c) Evaluating ensemble skill

While ensemble spread is analyzed for four variables (2-m temperature, 2-m

dewpoint temperature, and hourly accumulated precipitation), hourly and 6-hourly accumulated precipitation are chosen to evaluate ensemble skill; these variables are chosen due to the relative ease of verification (i.e., due to the existence of an observational dataset at the desired scales of verification). NCAR/EOL Stage IV precipitation data (Lin 2011) are treated as "truth" and used to evaluate the ensemble precipitation forecasts. The Stage IV data are produced on an approximately 4.8-km polar stereographic grid with 1121 east-west grid points and 881 north-south grid points; therefore, a neighborhood budget method is used to remap the data to a 3-km Lambert conformal grid with 1680 east-west grid points and 1152 north-south grid points to match the grid used by the forecasts. The remapped Stage IV data are used for verification and are compared against the raw precipitation forecasts from the mixed- and single-physics ensembles. Metrics used for verification include: mean square error (MSE; e.g., Eckel and Mass 2005), fractions skill score (FSS; Roberts and Lean 2008), area under the relative operating characteristics curve (AUC; e.g., Marzban 2004), and attributes diagrams (Hsu and Murphy 1986).

Mean square error (MSE), a traditional point-to-point verification metric, is computed for hourly accumulated precipitation for forecast hours 1-36 using equation (B6) in Eckel and Mass (2005):

$$\mathrm{MSE} = \left(\frac{n}{n+1}\right)\frac{1}{M}\sum_{m=1}^{M}(\bar{e}_m - o_m)^2 \qquad (2),$$

where n is the number of ensemble members, M is the number of forecast-observation pairs (which includes the number of non-overlapping spatial windows in the domain

over each day in the analysis), $\bar{e}_m$ is the ensemble mean at *m*, and o_m is the observation at *m*. Equation (2) is used for ensembles of finite size. MSE for the mixed- and single-physics ensemble forecasts is computed over the same 11 square-shaped neighborhoods (i.e., ranging from 1 to 240 grid boxes per side) used to compute the variance. As with the variance computations, upscaling (Ebert 2009) is applied to find each ensemble member's forecast value at each neighborhood. Then, the ensemble mean is calculated. Upscaling of the Stage IV observation data is used to produce the $o_m$values in equation (2). Mixed- and single-physics ensemble MSE, as well as the difference between mixed- and single-physics ensemble MSE (i.e., mixed-physics ensemble MSE – single-physics ensemble MSE), are plotted against time for forecast hours 0-36.

Given its design to be computed over a variety of neighborhoods, FSS is useful for determining forecast skill at a variety of spatial scales. Unlike some other forecast evaluation metrics (e.g., area under the relative operating characteristics curve), FSS depends on bias; more biased forecasts always produce lower FSS values at large spatial scales and usually produce lower FSS values at small spatial scales (Mittermaier and Roberts 2010). FSS can be expressed mathematically as:

$$\mathrm{FSS} = 1 - \frac{\frac{1}{M}\sum_{m=1}^{M}(F_m - O_m)^2}{\frac{1}{M}\left[\sum_{m=1}^{M} F_m^2 + \sum_{m=1}^{M} O_m^2\right]} \qquad (3),$$

where M is the number of forecast-observation pairs (which includes the number of overlapping spatial windows in the domain over each day in the analysis), F_m is the ensemble mean forecast fraction at *m*, and O_m is the observed fraction at *m*. Herein, FSS

is computed for accumulated hourly precipitation for forecast hours 1-36 using 0.10-, 0.25-, 0.50-, 0.75-, and 1.00-inch precipitation thresholds. Forecasts (observations) meeting or exceeding the threshold are considered to be "yes" forecasts (observations). Ten square neighborhoods are examined to determine how FSS varies with spatial scale; these neighborhoods consist of, respectively: 1, 2, 3, 4, 6, 8, 12, 16, 24, and 48 grid boxes per side. Again, since each grid box measures 3-km per side, the ten neighborhoods correspond to spatial scales of, respectively: 3-, 6-, 9-, 12-, 18-, 24-, 36-, 48-, 72-, and 144-km. Mixed- and single-physics ensemble FSS is plotted against spatial scale for each of the five precipitation thresholds at forecast hours 1, 12, 24, and 36. Additionally, to determine how FSS varies with time, a time series of mixed- and single-physics ensemble FSS is plotted for forecast hours 1-36 for each of the 10 spatial scales for the 0.10-inch threshold forecasts.

A skillful baseline FSS score is given by:

$$\mathrm{FSS}_{\mathrm{useful}} = 0.5 + \frac{f_0}{2} \qquad (4),$$

where f_0 represents the fractional coverage of "yes" forecasts over the entire domain (and—in this case—over all days in the analysis; Roberts and Lean 2008). Note that $\mathrm{FSS}_{\mathrm{useful}}$, as given in equation (4), is equivalent to $\mathrm{FSS}_{\mathrm{uniform}}$ in Roberts and Lean (2008). The smallest scale for which FSS = $\mathrm{FSS}_{\mathrm{useful}}$ is considered to be the smallest useful scale (i.e., the scale at which the forecast contains useful information; Roberts and Lean 2008).

Finally, area under the relative operating characteristics curve (AUC; e.g.,

Marzban 2004), which measures a forecast system's ability to discriminate between events and non-events (e.g., Mason and Graham 2002), is used to verify 6-hour accumulated precipitation forecasts. AUC values greater than or equal to 0.70 are considered skillful in an ensemble framework (Buizza et al. 1999). 6-hourly accumulated precipitation forecasts and observations are created, respectively, by summing the hourly precipitation forecasts and the Stage IV hourly observation data over 6-hour periods ending at 0600 UTC, 1200 UTC, 1800 UTC, and 0000 UTC. The same five precipitation thresholds used in the FSS analysis are used in the AUC computations to convert the quantitative precipitation (QPF) forecasts into binary forecasts. In each ensemble member, grid boxes that meet or exceed the given threshold are assigned a value of 1, while all other grid boxes are assigned a value of 0. Next, at each grid box, the ratio of ensemble members containing a 1 to the number of members containing a 0 is computed. This fraction is smoothed using a 2-dimensional kernel density function (e.g., Brooks et al. 1998, Sobash et al. 2011, Loken et al. 2017) with varying degrees of spatial smoothing. Standard deviations from 1.5-km to 72-km are tested. Note that this is essentially the same procedure used by Loken et al. (2017) to produce probabilistic severe weather forecasts using forecast updraft helicity (Chapter 2). AUC is then computed by summing contingency table elements (i.e., hits, misses, false alarms, and correct negatives; e.g., Loken et al. 2017 (Chapter 2)) over all grid boxes in the domain and over all days in the analysis. As in Loken et al. 2017, probability of detection (POD; equation 3 in Loken et al. 2017) and probability of false detection (POFD; equation 4 in Loken et al. 2017) are computed at the following levels of probability: 1, 2, 5, 10, 15, 20, 25, 30, 35, 40, 45, 50, 55, 60, 65, 70, 75, 80, 85, 90,

and 95%. Grid boxes meeting or exceeding the given probability level are considered to be "yes" forecasts, while other grid boxes are considered to be "no" forecasts at the given probability level.

Because AUC does not give information about forecast reliability (Wilks 2001), attributes diagrams (Hsu and Murphy 1986) are used to assess forecast reliability. Attributes diagrams, which plot observed relative frequency against forecast probability, are used principally to assess the impact of spatial smoothing on reliability at each of the five precipitation thresholds and at each of the four 6-hour forecast periods.

3. Results

(a) Ensemble spread

1) RAW DATA

For each of the four variables analyzed (i.e., 2-m temperature, 2-m dewpoint temperature, 500-mb height, and hourly accumulated precipitation), the smallest (largest) spatial scales generally have the greatest (lowest) variances at a given forecast hour (Fig. 3.2a-d). This finding makes sense. As the spatial scale (i.e., size of the neighborhood) increases, the variance becomes less sensitive to small, local differences between ensemble members because of the increased spatial averaging. Physically, it also makes sense that the smallest scales will have the greatest variances, since smaller eddies are more difficult to predict and are therefore associated with more uncertainty (e.g., Lorenz 1969).

Consistent with the findings of Clark et al. (2010), a diurnal-cycle is noted in the

2-m temperature, 2-m dewpoint, and hourly precipitation variance time series (Fig. 3.2a,b,d). The hourly precipitation time series (Fig. 3.2d) contains the most well-defined diurnal cycle; local maxima in variance exist around forecast hours 4 and 25, while local minima exist near forecast hours 16 and 36 for most spatial scales. Less well-pronounced diurnal cycles are seen in the 2-m temperature and 2-m dewpoint variance time series (Fig. 3.2a,b). Both variables have a local maximum around forecast hour 24 and a local minimum around forecast hour 12 for most spatial scales. For each of the three variables, the amplitude of the diurnal cycle decreases as spatial scale increases. As in Clark et al. (2010), the 500-mb height variance time series does not exhibit a diurnal cycle. 500-mb height variance generally increases with time, with the variance increasing faster for the smaller spatial scales.

Variance difference plots (Fig. 3.3a-d) indicate that the mixed-physics ensemble nearly always generates greater variance than the single-physics ensemble at a given spatial scale and forecast hour for a given variable. However, the difference between the mixed- and single-physics ensemble variance generally decreases as spatial scale increases for all four variables.

While the difference between the mixed- and single-physics variances for 500-mb height generally increases with time (for all spatial scales), the difference in variances depends more on the diurnal cycle for the other three variables. For example, for 2-m temperature, the difference in variance is locally maximized around forecast hour 10 and again around forecast hour 26, while the differenced is minimized around forecast hour 18 (Fig. 3.3a). For 2-m dewpoint, the difference is locally maximized around forecast hour 24 and locally minimized shortly thereafter, around forecast hour

26 (Fig. 3.3b). For hourly precipitation, the difference in variance has two sharp local maxima: one around forecast hour 4 and another, slightly smaller one, around forecast hour 25. Between these local maxima is a local minimum in variance around forecast hour 16 (Fig. 3.3c).

Ratios of $\frac{\text{single-physics ensemble variance}}{\text{mixed-physics ensemble variance}}$ are computed to determine how the proportion of spread generated by the mixed-physics ensemble varies with time (Fig. 3.4a-d). While the 500-mb height variance ratios remain approximately constant with time and do not differ dramatically with spatial scale (Fig. 3.4c), the variance ratios from the other fields have more noticeable variations with time and spatial scale. For example, the 2-m temperature ratios reach a local maximum at approximately forecast hour 18 (Fig. 3.4a), indicating that, proportionally, the mixed-physics ensemble contributes less variance at that time than at other forecast hours. The 2-m dewpoint and hourly precipitation ratios also vary with time, although with much less well-defined local maxima and minima (Fig. 3.4b,d). Despite these variations, it should be noted that, for all four variables, the variance ratios generally remain below 1.0 for the vast majority of spatial scales and forecast hours, signifying that the mixed-physics ensemble generally produces more spread, proportionally, relative to the single-physics ensemble.

Interestingly, for the 2-m temperature, 2-m dewpoint temperature, and hourly accumulated precipitation fields, the variance ratio is smallest—indicating that the mixed-physics generates proportionally more spread—for the largest spatial scales (Fig. 3.3a,b,d). Thus, even while the *difference* between the mixed- and single-physics variances is lowest for the largest spatial scales (Fig. 3.2a,b,d), the *proportion* of

variance created by the mixed-physics ensemble is largest—at least for these three variables.

2) BIAS-CORRECTED DATA

While the bias-corrected (Fig. 3.5a-d) and raw (Fig. 3.2a-d) time series generally have similar shapes, the raw variances tend to be slightly greater than the bias-corrected variances at a given forecast hour. This finding is expected, given that the bias-correction procedure removes some of the "artificial" spread that results from systematic biases among the ensemble members (Clark et al. 2010). The reduced spread in the bias-corrected time series is most clearly seen in the 500-mb height variances (Fig. 3.5c; fig. 3.2c).

The greatest difference between the raw and bias-corrected variance time series exists for the hourly precipitation field. In the bias-corrected hourly precipitation time series, the single-physics ensemble variance is greater than the mixed-physics ensemble variance at most hours for spatial scales from 3- to 24-km (Fig. 3.5d); this result is in direct contrast to the raw time series, in which the mixed-physics ensemble always generates greater spread relative to the single-physics ensemble for all spatial scales and at all forecast hours (Fig. 3.2d).

The bias-corrected differences between the mixed- and single-physics ensemble forecast variances are explicitly shown in Fig. 3.6a-d. Comparing the raw (Fig. 3.3d) and bias-corrected (Fig. 3.6d) variance difference time series for hourly precipitation, it is directly seen that—for the bias-corrected data—the single-physics ensemble contributes more spread than the mixed-physics ensemble for at least some forecast

hours at spatial scales from 3- to 72-km, while—for the raw data—the mixed-physics ensemble always contributes more spread than then the single-physics ensemble, regardless of forecast hour or spatial scale. Interestingly, a general comparison of the raw and bias-corrected variance difference time series indicates that the mixed-physics – single-physics variance difference is generally lower for the bias-corrected data than for the raw data. This finding makes sense given that the mixed-physics ensemble likely contains more systematic biases—and thus more artificial spread (Clark et al. 2010)—than the single-physics ensemble simply by virtue of its greater microphysics and PBL diversity. Thus, removing the systematic biases from both ensembles would be expected to reduce the variance of the mixed-physics ensemble more than the variance of the single-physics ensemble, leading to a decrease in the variance differences.

While the variance differences are generally lower after the bias-correction procedure is applied, the bias-corrected variance ratios (Fig. 3.7a-d) are generally similar to the corresponding raw variance ratios (Fig. 3.4a-d), at least for the 2-m temperature, 2-m dewpoint, and 500-mb height variables. The similar ratios in the raw and bias-corrected datasets perhaps occur because, while the bias-corrected data generally have lower differences between the mixed- and single-physics variances, they also have lower overall variances, resulting in similar ratios. The raw (Fig. 3.4d) and bias-corrected (Fig. 3.7d) hourly precipitation variance ratios, however, are noticeably different since the single-physics ensemble generates more variance than the mixed-physics ensemble in the bias-corrected dataset but not in the raw dataset.

(b) Ensemble Skill

1) MSE

At a given forecast hour, MSE is greater for smaller spatial scales (Fig. 3.8a); this result is expected given that MSE is a point-to-point verification metric, and smaller spatial scales reduce the probability of a perfect correspondence between the forecasts and observations, leading to larger objective errors (but not necessarily larger subjective errors; Ebert 2008, 2009). For most spatial scales, the time series of MSE is bimodal: local maxima exist around forecast hours 3 and 24, while local minima exist around forecast hours 15 and 36. For most spatial scales (i.e., for those less than and equal to 144 km), the difference between mixed- and single-physics MSE is maximized around forecast hour 3 and then drops to near 0 by forecast hour 7 (Fig. 3.8b). Interestingly, the MSE difference is seldom negative for most spatial scales and forecast hours, suggesting that the mixed-physics ensemble forecasts have slightly more errors than the single-physics ensemble forecasts. Nonetheless, aside from the first 6 forecast hours (and even inside the first 6 forecast hours at larger spatial scales), the difference between the mixed- and single-physics MSE is quite small.

2) FSS

At spatial scales from 3- to 144-km and for forecast hours ranging from 1 to 36, the 0.10-inch precipitation threshold forecasts yield the greatest FSS, while the FSS progressively decreases (at a given spatial scale) as the precipitation threshold is increased to 0.25-, 0.50-, 0.75-, and 1.00-inch (Fig. 3.9a-d). For forecast hours 1, 12, 24, and 36, the mixed- and single-physics ensembles produce similar FSS values for a given precipitation threshold. Interestingly, the single-physics ensemble tends to give slightly

greater FSS values than the mixed-physics ensemble (for the vast majority of precipitation thresholds) at forecast hours 1 and 12, while the mixed-physics ensemble tends to give slightly greater FSS values than the single-physics ensemble at forecast hours 24 and 36. Nonetheless, the differences in mixed- and single-physics ensemble FSS (for a given spatial scale) are small at all forecast hours and all spatial scales examined.

As expected, at all four forecast hours, FSS increases as spatial scale increases. However, it should be noted that, by forecast hour 24, all of the hourly precipitation forecasts—with the sole exception of the mixed-physics, 0.10-inch threshold forecast (which just barely exceeds FSS_{useful} for the 144-km spatial scale)—fall below the threshold for useful skill (i.e., FSS_{useful}) for all spatial scales examined. By forecast hour 36, all forecasts are well below FSS_{useful} for all spatial scales analyzed.

Since the 0.10-inch threshold is found to produce the greatest FSS values at a given spatial scale and forecast hour, a time series of FSS is constructed using that threshold (Fig. 3.10). The time series shows that, at all spatial scales, FSS decreases rapidly within the first several hours of the forecast and then continues to decrease more gradually. Both the 3-km mixed- and single-physics forecasts drop below FSS_{useful} after forecast hour 2, while the 144-km mixed- and single-physics forecasts do not permanently dip below FSS_{useful} until around forecast hour 30. Importantly, for a given spatial scale and forecast hour, the mixed- and single-physics ensembles produce similar FSS values. That is, with everything held constant (i.e., constant precipitation threshold, forecast hour, etc.), the mixed- and single-physics ensembles are skillful out to approximately the same forecast hour and down to approximately the same spatial

scale.

3) AUC FROM 6-HOUR PROBABILISTIC PRECIPITATION FORECASTS

In general, AUC is maximized for the 6-hour period ending at 0600 UTC (henceforth denoted as F06) and minimized for the 6-hour period ending at 0000 UTC (henceforth denoted as F00; Fig. 3.11a-d). For a given threshold and forecast period, the mixed-physics ensemble generally gives greater AUC values than the single-physics ensemble, although the mixed- and single-physics AUC values are typically quite similar. With less spatial smoothing, the 0.10-inch precipitation threshold forecasts are generally superior (Fig. 3.11a); however, as more spatial smoothing is applied, the AUC values of all forecasts examined become increasingly similar. More spatial smoothing also increases the AUC of all forecasts.

The impact of varying the standard deviation of the Gaussian kernel (henceforth referred to as the spatial smoothing parameter) is assessed explicitly in Fig. 3.12a-d. For all four forecast periods examined, AUC increases relatively rapidly as the spatial smoothing parameter is increased from 1.5- to 12-km and then increases more gradually as the spatial smoothing parameter is further increased to 72-km. The larger precipitation threshold forecasts appear to benefit more from additional spatial smoothing relative to the smaller precipitation threshold forecasts.

4) ATTRIBUTES DIAGRAMS

Varying the spatial smoothing parameter directly influences forecast reliability. With the smallest amount of spatial smoothing examined (i.e., when the spatial

smoothing parameter is set to 1.5 km), all five precipitation threshold forecasts tend to under-forecast at lower forecast probabilities but over-forecast at higher probabilities; this result applies to both the mixed- and single-physics ensembles (Fig. 3.13 a-e). As the precipitation threshold is increased, a larger proportion of the probabilities are under-forecast and fewer probabilities are over-forecast. A similar effect occurs as the spatial smoothing parameter is increased for a given precipitation threshold: all probabilities tend in the under-forecasting direction. That is, as the spatial smoothing parameter is increased, probabilities that were already under-forecast become more under-forecast, probabilities that had near-perfect reliability become slightly under-forecast, and probabilities that were over-forecast become less over-forecast (or even slightly under-forecast). This result makes sense given that greater spatial smoothing tends to increase the number of extremely low-probability (i.e., < 0.20) forecasts and decrease the number of the higher probability forecasts (Fig. 3.14a-e). Because the observed relative frequency does not change as the spatial smoothing parameter of the forecast is varied, fewer forecasts in a given probability bin results in a progression toward under-forecasting.

For all values of the spatial smoothing parameter, the reliability curves from the mixed- and single-physics ensembles are very similar. For the vast majority of spatial smoothing parameter values and for the vast majority of forecast probabilities, the single-physics forecasts tend slightly in the direction of under-forecasting relative to the mixed-physics forecasts, but the difference between the mixed- and single-physics reliability curves is very slight.

For both the mixed- and single-physics ensembles, no value of the spatial

smoothing parameter examined optimizes reliability at all forecast probabilities. However, for each precipitation threshold, the mean distance from the line of perfect reliability seems to be minimized with the lower values of the spatial smoothing parameter. This is an interesting result because of the finding that more spatial smoothing results in greater AUC values. Hence, there appears to be a tradeoff between forecast reliability and discrimination ability (i.e., AUC) as the forecasts are increasingly smoothed: less-smoothed forecasts appear to have better reliability but a worse ability to discriminate between the occurrence and non-occurrence of a precipitation event greater than or equal to a given threshold.

Forecast reliability is similar for all four forecast periods (Fig. 3.15 a-d). Nonetheless, the F00 forecasts appear to be slightly more reliable than the other forecast periods, since—relative to the corresponding forecasts at F06, F12, and F18—the most-smoothed forecasts at F00 tend to be closer to the line of perfect reliability at the highest forecast probabilities and the less-smoothed forecasts tend to be slightly closer to the line of perfect reliability at the low forecast probabilities. These results make sense given that, compared to the other forecast periods, the F00 period contains slightly more forecasts in the lowest probability bins (which suffer from under-forecasting bias) and slightly less forecasts in the higher probability bins (which suffer from over-forecasting bias; Fig. 3.16a-d). However, it should be noted that, while the F00 period enjoys slightly greater reliability relative to the other forecast periods, it has lower AUC values.

As found during the analysis across the five precipitation thresholds, mixed- and single-physics ensemble forecasts are found to have similar reliability curves for a given

spatial smoothing parameter and a given forecast period.

4. Summary and discussion

This study investigated how the spread and skill of mixed- and single-physics convection-allowing ensemble forecasts varied with forecast hour and spatial scale. Ensemble forecast spread was examined for four variables—2-m temperature, 2-m dewpoint temperature, 500-mb height, and hourly accumulated precipitation—using both raw and bias-corrected variance time series for forecast hours 0-36. Meanwhile, ensemble skill was evaluated for hourly and 6-hourly accumulated precipitation forecasts. These forecasts were created—and assessed—in a variety of ways. First, ensemble mean hourly quantitative precipitation forecasts were evaluated at a variety of spatial scales using traditional MSE (adjusted for an ensemble of finite size). Next, binary (i.e., yes/no) hourly precipitation forecasts were created using 0.10-, 0.25-, 0.50-, 0.75-, and 1.00-inch thresholds; these were evaluated for forecast hours 1-36 at a variety of spatial scales using FSS. Finally, probabilistic 6-hourly precipitation forecasts were created at each of the above five thresholds by spatially smoothing raw ensemble probabilities (i.e., the fraction of ensemble members meeting or exceeding the threshold) at each grid point; varying values of the spatial smoothing parameter (from 1.5- to 72-km) were tested. Discrimination ability was measured using AUC, while reliability was assessed using attributes diagrams.

When the raw ensemble data were examined, the mixed-physics ensemble was found to have greater variance than the single-physics ensemble for all four variables studied at nearly all forecast hours (from 0-36) and spatial scales (from 3- to 720-km).

However, the differences in variance were generally greatest at the smallest spatial scales and decreased as spatial scale increased. One explanation for this finding is that, as the spatial scale of the analysis is increased, precipitation systems occupy a smaller fraction of each analysis neighborhood. This is significant because the two ensembles' different representation of microphysics uncertainty only impacts each ensemble's forecast where convection exists; therefore, less fractional coverage of convection within each neighborhood implies less difference between the two ensemble forecasts. Another explanation is that localized differences in the two ensembles' forecast fields (for any of the four variables) tend to get averaged out as larger neighborhoods are considered.

Interestingly, while the variance *differences* suggested that the mixed-physics and single-physics ensemble spread became increasing similar at larger spatial scales, the variance *ratios* suggested that, proportionally, the mixed-physics ensemble provided greater spread at the larger spatial scales compared to the smaller spatial scales, at least for the 2-m temperature, 2-m dewpoint temperature, and hourly accumulated precipitation fields (the 500-mb height variance ratios were generally quite similar at all spatial scales and forecast hours). This result was surprising. It indicated that, for the 2-m temperature, 2-m dewpoint, and hourly precipitation fields, the mixed-physics ensemble variance decreased less than the single-physics ensemble variance as spatial scale increased. Nevertheless, at large spatial scales, where the variance ratio was the lowest, the variance of both ensembles was quite small. This finding suggests that perhaps more weight should be given to the variance differences as opposed to the variance ratios when comparing the mixed- and single-physics ensemble variances at

the larger spatial scales.

To remove the impact of systematic biases on the ensemble variance, a bias-correction procedure based on probability matching was applied (Ebert 2001; Clark et al. 2010); the PDF of each ensemble member was replaced with the PDF of the core01 member, since this member was present in both the mixed- and single-physics ensembles. As in Clark et al. (2010), the bias-corrected variances were generally lower than the corresponding raw variances, which makes sense given that probability matching reduces the "artificial" ensemble spread from systematic biases (Clark et al. 2010; Eckel and Mass 2005). The bias-corrected differences between the mixed- and single-physics ensemble variance were also lower than the corresponding raw variance differences, probably because the mixed-physics ensemble contained more systematic biases than the single-physics ensemble and therefore experienced a greater reduction in variance after calibration. The smaller variance differences after calibration suggest that bias-correction may reduce some of the spread benefits provided by the mixed-physics ensemble relative to the single-physics ensemble. With that said, the proportion of variance generated by the mixed- compared to the single-physics ensemble generally remained similar both before and after bias-correction. One notable exception was for the hourly accumulated precipitation field, which had a noticeable shift toward higher variance ratios (i.e., more relative variance generated by the single-physics ensemble) at all spatial scales after the bias-correction procedure was applied. In fact, the variance ratio even exceeded 1 (indicating the single-physics ensemble generated more spread relative to the mixed-physics ensemble) at many forecast hours and spatial scales. That the precipitation variance and variance ratios were noticeably different before and after

the bias-correction procedure was applied suggests that a large portion of the forecast precipitation variance in each ensemble (and at all spatial scales) can be attributed to the *magnitude* of the precipitation forecast and not merely the *placement* of precipitation systems.

Ensemble skill metrics generally indicated that the mixed- and single-physics ensembles had similar skill at most forecast hours and spatial scales examined. Aside from the first few forecast hours, the two ensembles' MSE values were very similar at all neighborhoods. Similarly, FSS values from the binary precipitation forecasts at each of the five thresholds were generally similar for the mixed- and single-physics ensembles. The greatest difference between the mixed- and single-physics FSSs occurred around forecast hour 24 (i.e., around 0000UTC), when precipitation variance was found to be maximized, both in this study and in Clark et al. (2010). Indeed, previous studies (e.g., Dai et al. 1999; Wallace 1975; Easterling and Robinson 1985) have shown that summertime precipitation is generally maximized during the late afternoon in the southeastern and western U.S. and during the late evening and early morning in the Great Plains. Therefore, there may be greater benefit to using the mixed-physics ensemble during the late afternoon and evening periods, when accounting for uncertainties in model microphysics and PBL parameterizations may be more important due to the greater climatological probability of precipitation. Nevertheless, even during the evening hours, the FSS values depended much more on the precipitation threshold used to create the forecast than the use of mixed- or single-physics parameterizations. Moreover, the mixed- and single-physics FSSs were generally similar at all spatial scales examined.

To account for uncertainties in time, the same 5 thresholds were used to create 6-hourly probabilistic precipitation forecasts. For all four forecast periods and at all five precipitation thresholds, the mixed-physics forecasts generally had slightly greater AUC. The greatest differences between the mixed- and single-physics AUC occurred for the larger precipitation thresholds and the late afternoon and early evening forecast periods. Again, these findings indicate that the mixed-physics ensemble may offer the most benefit to the single-physics ensemble in situations where heavy—or at least moderate—rainfall is more probable, which makes sense, as the microphysics parameterizations within each ensemble explicitly influence the structure and evolution of simulated convection. Interestingly, the degree of spatial smoothing did not have much influence on the relative skill of the mixed- and single-physics ensemble forecasts, which perhaps implies that the two ensemble forecasts differed more on the magnitude/character (i.e., convective vs. stratiform nature) of the forecast precipitation than its location. This finding is interesting because it supports Johnson and Wang (2017)'s finding that their mixed-physics ensembles more accurately predicted storm structure relative to their single-physics ensemble, but it also contradicts their observation that the two mixed-physics ensembles tended to reduce errors in forecast precipitation location relative to the single-physics ensemble. Perhaps differences in forecast precipitation location between the two ensembles are slight and matter less when spatial smoothing is applied to create the ensemble forecasts.

Both the mixed- and single-physics ensembles had similar forecast reliability curves at a given value of the spatial smoothing parameter; however, at nearly all values of spatial smoothing and all forecast probabilities, the single-physics ensemble tended

slightly more toward under-forecasting relative to the mixed-physics ensemble. One possible explanation for this finding is that the single-physics ensemble perhaps had less systematic bias—and therefore less spread—than the mixed-physics ensemble, which made it slightly more difficult for the single-physics ensemble to meet or exceed the given precipitation thresholds. Interestingly, for both the mixed- and single-physics ensembles, a tradeoff was noted between forecast discrimination ability and reliability: forecasts with greater (smaller) AUC values had worse (improved) reliability. Given this finding, forecast developers should consider the needs of users to determine an optimal degree of spatial smoothing for the ensemble forecasts.

5. Conclusion: Implications for convection-allowing ensemble design and future work

Overall, for the vast majority of forecast hours for all four variables studied, the mixed-physics ensemble seems to provide slightly greater ensemble spread relative to the single-physics ensemble, especially at smaller spatial scales and especially if the ensemble is not calibrated for bias. This result is consistent with previous work that has found multiple microphysics and PBL parameterizations can be an important way to generate spread in convection-allowing ensembles (e.g., Johnson et al. 2017; Clark et al. 2010). However, as the spatial scale of interest is increased, and as systematic bias is taken into account, the mixed- and single-physics ensemble variances generally become more similar.

The mixed-physics ensemble also appears to produce *slightly* more skillful precipitation forecasts than the single-physics ensemble, especially for larger

precipitation thresholds, and especially for late afternoon and evening periods, when precipitation tends to be maximized climatologically over much of the contiguous U.S. during late spring and early summer. It is possible that, especially for periods when heavy or moderate rainfall is likely, the mixed-physics ensemble may provide better guidance of convective structure relative to the single-physics ensemble.

Nevertheless, the differences between the mixed- and single-physics ensembles' spread and skill are generally small, especially when systematic biases are taken into account (i.e., the ensemble is well-calibrated) and at larger spatial scales. It must be noted that, while a mixed-physics ensemble provides slightly greater spread and skill relative to a single-physics ensemble, a single-physics ensemble is easier for model developers to maintain. Therefore, especially if a well-calibrated single-physics convection-allowing model could be developed, the small forecast advantages of using a mixed-physics ensemble may not outweigh the maintenance advantages of using a single-physics ensemble in operations.

With that said, this study has a number of important shortcomings that should be addressed before a final recommendation to model developers can be made. Most notably, this study did not determine whether any of the differences in spread or skill between the mixed- and single-physics ensemble forecasts were statistically significant. Additionally, while a variety of statistic measures of ensemble spread and skill were analyzed, the raw forecast fields from each ensemble were not examined; therefore, subjective differences between the two ensemble forecasts remain unassessed. Future work should, at minimum, address these two limitations before any decisions are made about the configuration of convection-allowing ensembles in an operational setting.

Mixed- and single-physics ensemble forecasts should also ideally be compared for more variables over more seasons.

Even after the above limitations are addressed, many avenues exist for future work. For example, while the mixed-physics ensemble in this study consisted of both multiple microphysics and multiple PBL parameterizations, it would be interesting to assess the individual impact of multiple microphysics and PBL parameterizations on ensemble spread and skill. Doing so would provide greater insight into precisely when and why the mixed- and single-physics ensembles perform similarly in terms of spread and skill. Additionally, given that the present study only used the raw data when computing ensemble forecast skill, future work may wish to evaluate the mixed- and single-physics ensemble forecasts after calibrating both ensembles for bias. Such an analysis would help determine how much of the mixed-physics ensemble's superior precipitation forecasting skill can be attributed to the presence of systematic ensemble biases.

Acknowledgements

This work was made possible by a Presidential Early Career Award for Scientists and Engineers (PECASE). Additional support was provided by NOAA/Office of Oceanic and Atmospheric Research under NOAA-University of Oklahoma Cooperative Agreement NA11OAR4320072, U.S. Department of Commerce. Stage IV precipitation data was provided by NCAR/EOL under the sponsorship of the National Science Foundation. https://data.eol.ucar.edu/

Month	Day
May	02-06; 09-13; 16-20; 23; 25-27; 30-31
June	02-03

Table 3.1 Dates from the 2016 NOAA HWT SFE included in the dataset (23 dates; note that 24 May 2016 is not used in the analysis since not all ensemble members had available data on that day).

Ens. Member	IC	BC	Microphysics	PBL
core01[a,b]	NAMa+3DVAR	NAMf	Thompson	MYJ
core03[a]	core01+arw-p1_pert	arw-p1	P3	YSU
core04[a]	core01+arw-n1_pert	arw-n1	MY	MYNN
core05[a]	core01+arw-p2_pert	arw-p2	Morrison	MYJ
core06[a]	core01+arw-n2_pert	arw-n2	P3	YSU
core07[a]	core01+nmmb-p1_pert	nmmb-p1	MY	MYNN
core08[a]	core01+nmmb-n1_pert	nmmb-n1	Morrison	YSU
core09[a]	core01+nmmb-p2_pert	nmmb-p2	P3	MYJ
core10[a]	core01+nmmb-n2_pert	nmmb-n2	Thompson	MYNN
s-phys-rad02[b]	core01+arw-p1_pert	arw-p1	Thompson	MYJ
s-phys-rad03[b]	core01+arw-n1_pert	arw-n1	Thompson	MYJ
s-phys-rad04[b]	core01+arw-p2_pert	arw-p2	Thompson	MYJ
s-phys-rad05[b]	core01+arw-n2_pert	arw-n2	Thompson	MYJ
s-phys-rad06[b]	core01+arw-p3_pert	arw-p3	Thompson	MYJ
s-phys-rad07[b]	core01+nmmb-p1_pert	nmmb-p1	Thompson	MYJ
s-phys-rad08[b]	core01+nmmb-n1_pert	nmmb-n1	Thompson	MYJ
s-phys-rad09[b]	core01+nmmb-p2_pert	nmmb-p2	Thompson	MYJ
s-phys-rad10[b]	core01+nmmb-n2_pert	nmmb-n2	Thompson	MYJ

Table 3.2 Mixed- and single-physics ensemble member specifications (adapted from Clark et al. 2016). A superscript "a" denotes use in the mixed-physics ensemble, while a superscript "b" denotes use in the single-physics ensemble. NAMa and NAMf denote the 12-km NAM analysis and forecast, respectively. 3DVAR refers to the Advanced Regional Prediction System three-dimensional variational data assimilation and cloud analysis (Xue et al. 2003; Gao et al. 2004). Elements in the IC column ending with "pert" are perturbations from a 16-km 3-h Short-Range Ensemble Forecast (SREF; Du et al. 2014) member. Elements in the BC column after the first row refer to SREF member forecasts. Ensemble microphysics schemes include: Thompson (Thompson et al. 2004), Predicted Particle Properties (P3; Morrison and Milbrandt 2015), Milbrandt and Yau (MY; Milbrandt and Yau 2005), and Morrison (Morrison et al. 2005). Ensemble boundary layer schemes include: Mellow-Yamada-Janjić (MYJ; Mellor and Yamada 1982; Janjić 2002), Yonsei University (YSU; Noh et al. 2003), and Mellor-Yamada-Nakanishi-Niino (MYNN; Nakanishi 2000, 2001; Nakanishi and Niino 2004, 2006).

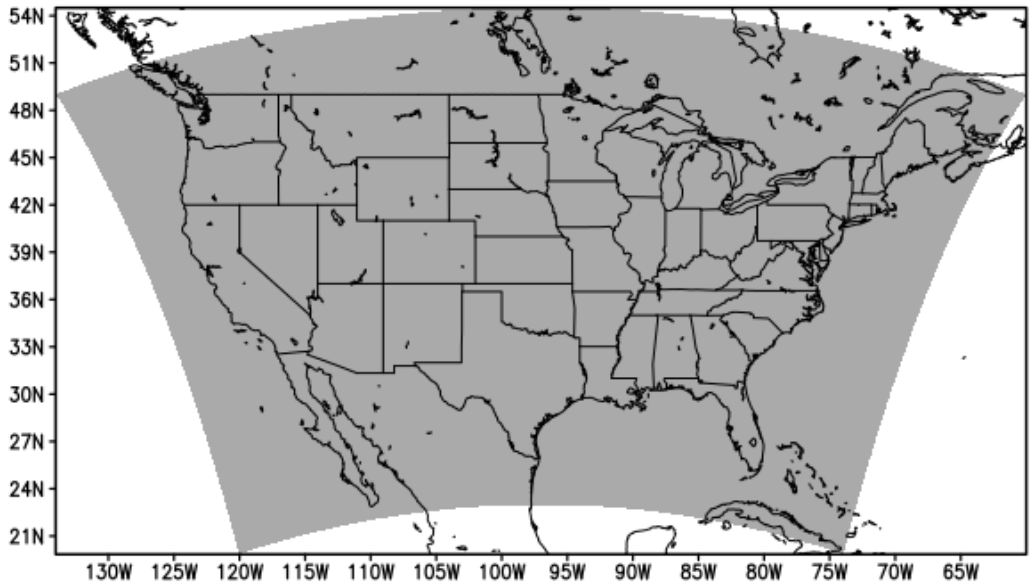

Figure 3.1 Analysis domain of the 2016 Community Leveraged Unified Ensemble (CLUE; gray shading).

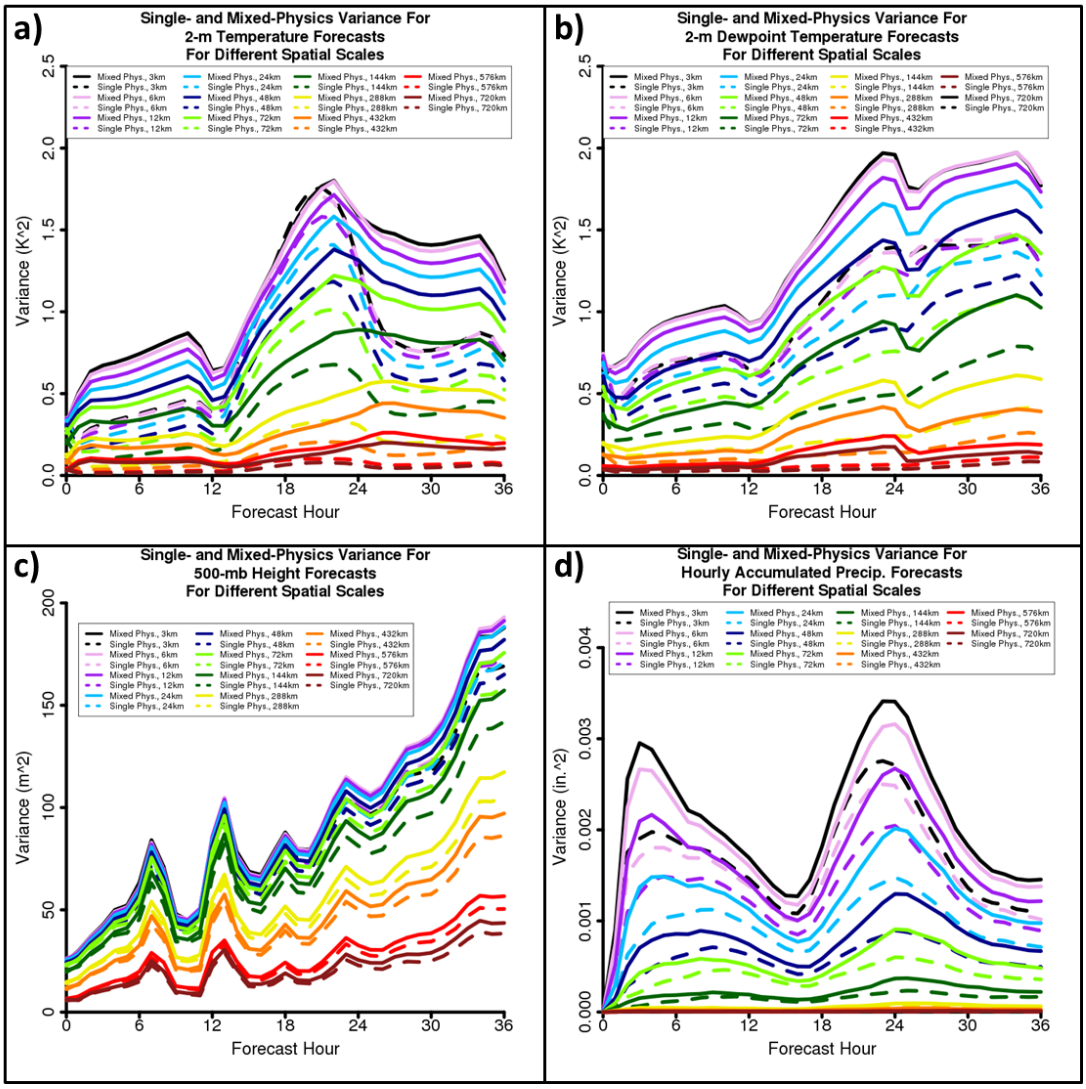

Figure 3.2 Raw variance time series for mixed- (solid) and single-physics (dashed) ensemble forecasts of (a) 2-m temperature, (b) 2-m dewpoint temperature, (c) 500-mb height, and (d) hourly accumulated precipitation. For each forecast variable, eleven spatial scales are shown: 3- (black), 6- (pink), 12- (purple), 24- (light blue), 48- (dark blue), 72- (light green), 144- (dark green), 288- (yellow), 432- (orange), 576- (red), and 720-km (dark red).

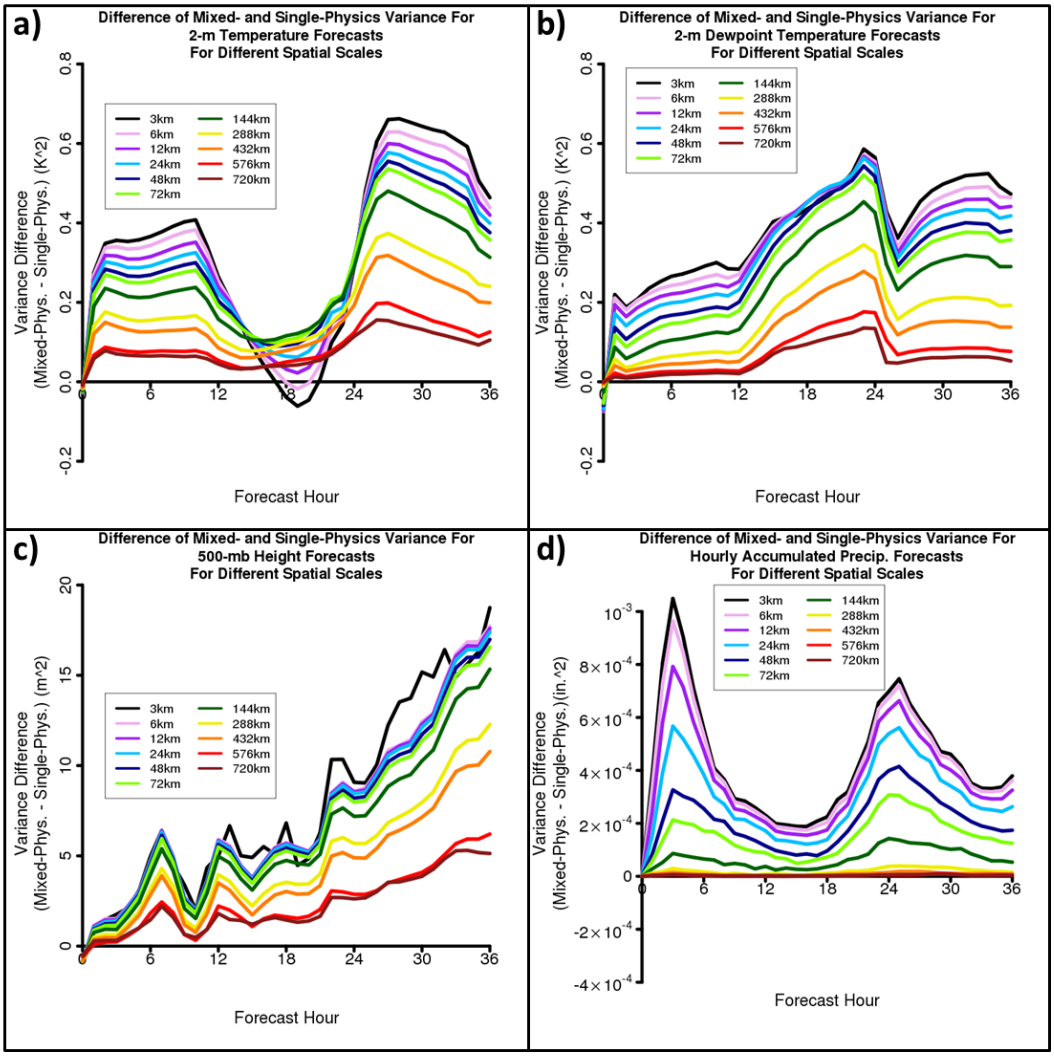

Figure 3.3 Time series of raw variance differences (mixed-physics variance – single physics variance) for forecasts of (a) 2-m temperature, (b) 2-m dewpoint temperature, (c) 500-mb height, and (d) hourly accumulated precipitation. For each forecast variable, eleven spatial scales are shown: 3- (black), 6- (pink), 12- (purple), 24- (light blue), 48- (dark blue), 72- (light green), 144- (dark green), 288- (yellow), 432- (orange), 576- (red), and 720-km (dark red).

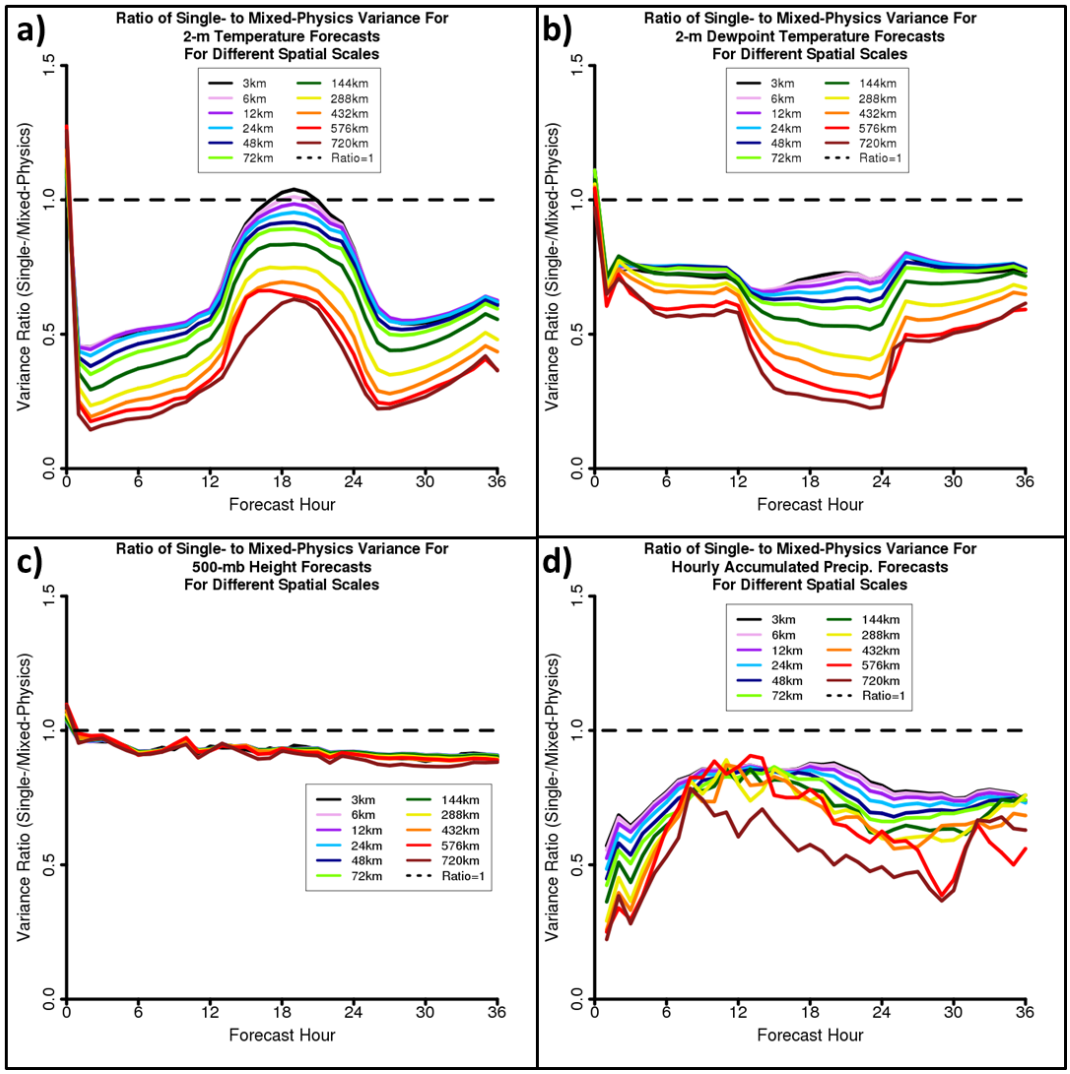

Figure 3.4 Time series of raw variance ratios (single-physics variance/mixed-physics variance) for forecasts of (a) 2-m temperature, (b) 2-m dewpoint temperature, (c) 500-mb height, and (d) hourly accumulated precipitation. For each forecast variable, eleven spatial scales are shown: 3- (black), 6- (pink), 12- (purple), 24- (light blue), 48- (dark blue), 72- (light green), 144- (dark green), 288- (yellow), 432- (orange), 576- (red), and 720-km (dark red). The black dashed line denotes where the single- to mixed-physics variance ratio is equal to 1.

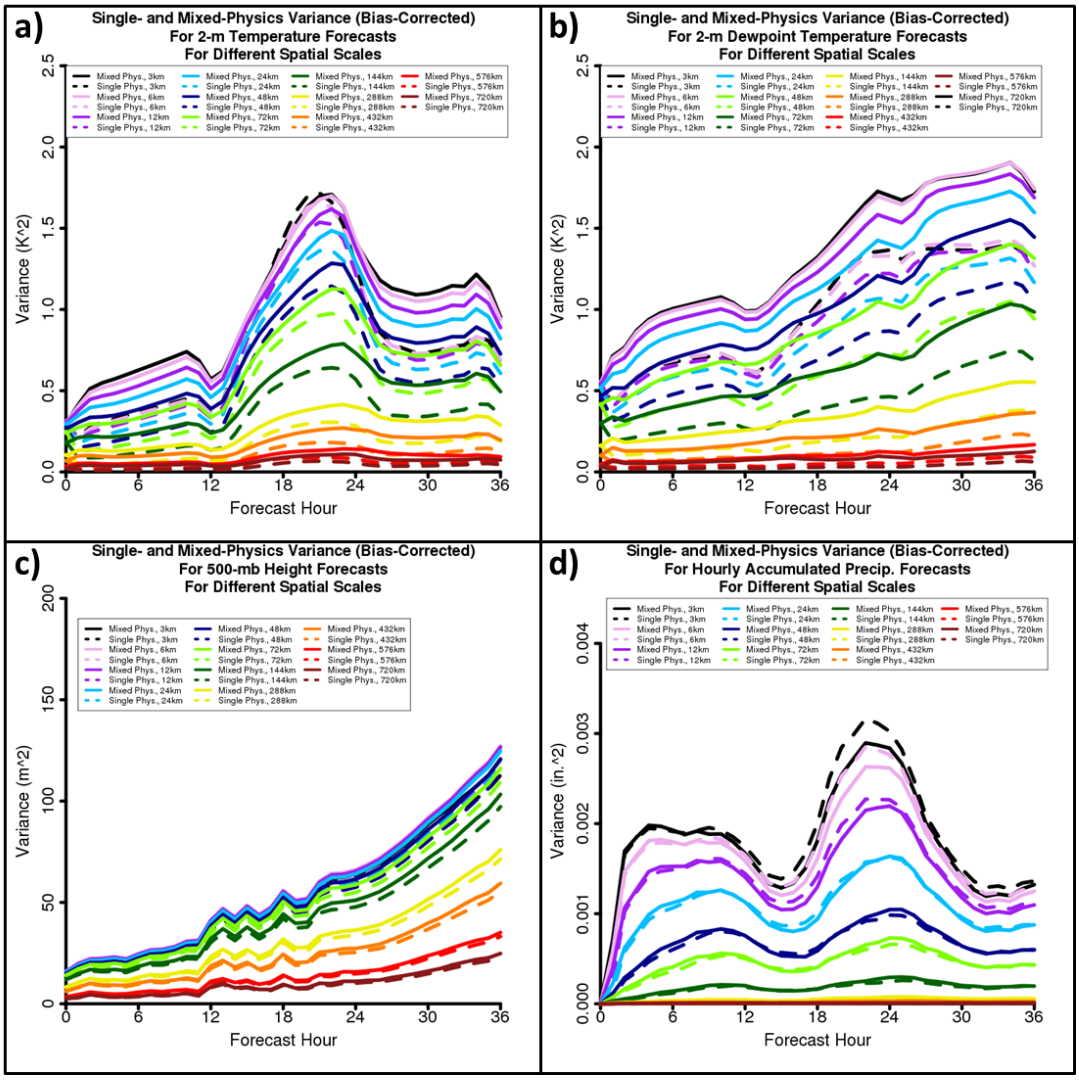

Figure 3.5 Bias-corrected variance time series for mixed- (solid) and single-physics (dashed) ensemble forecasts of (a) 2-m temperature, (b) 2-m dewpoint temperature, (c) 500-mb height, and (d) hourly accumulated precipitation. For each forecast variable, eleven spatial scales are shown: 3- (black), 6- (pink), 12- (purple), 24- (light blue), 48- (dark blue), 72- (light green), 144- (dark green), 288- (yellow), 432- (orange), 576- (red), and 720-km (dark red).

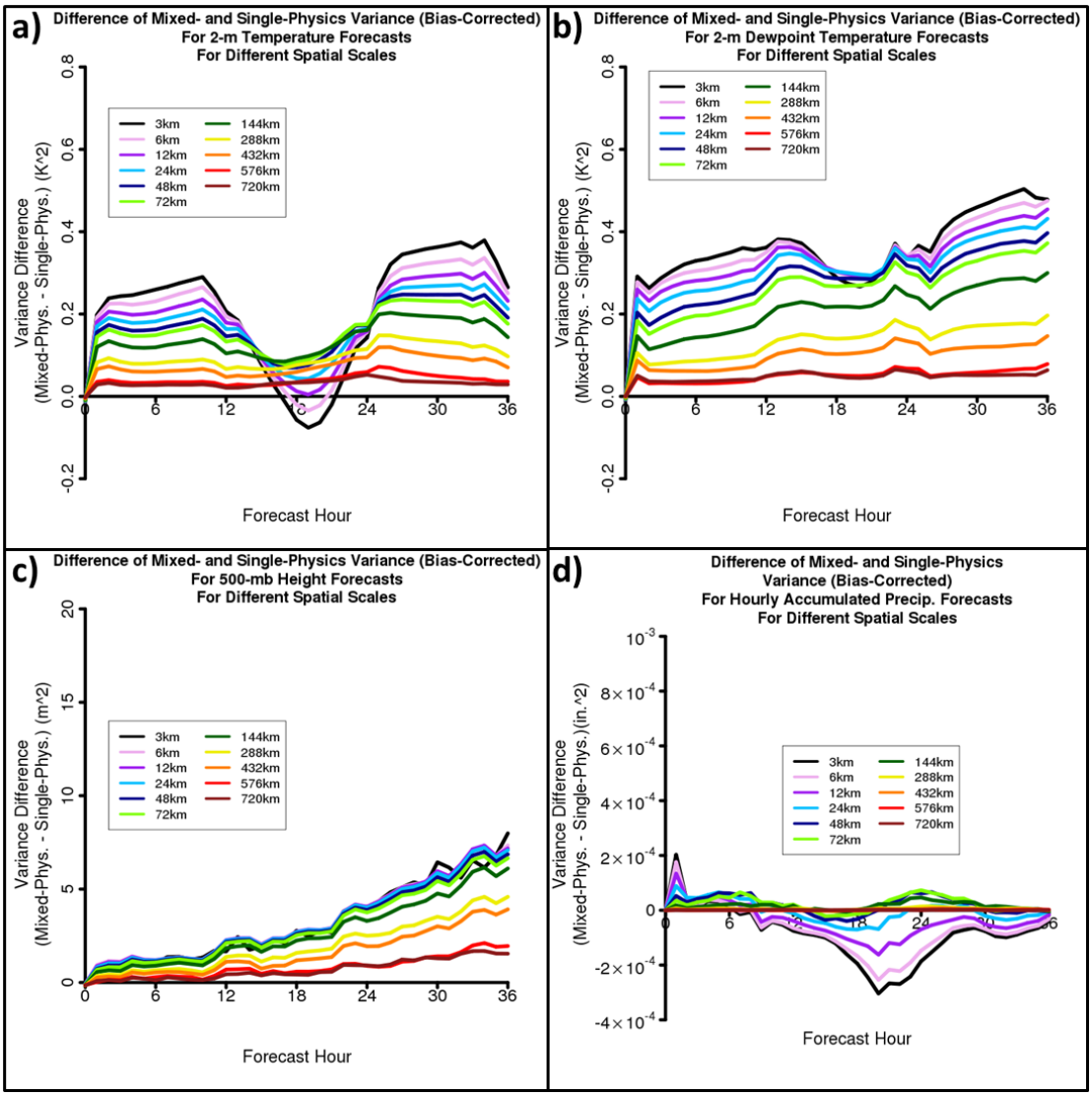

Figure 3.6 Time series of bias-corrected variance differences (mixed-physics variance – single physics variance) for forecasts of (a) 2-m temperature, (b) 2-m dewpoint temperature, (c) 500-mb height, and (d) hourly accumulated precipitation. For each forecast variable, eleven spatial scales are shown: 3- (black), 6- (pink), 12- (purple), 24- (light blue), 48- (dark blue), 72- (light green), 144- (dark green), 288- (yellow), 432- (orange), 576- (red), and 720-km (dark red).

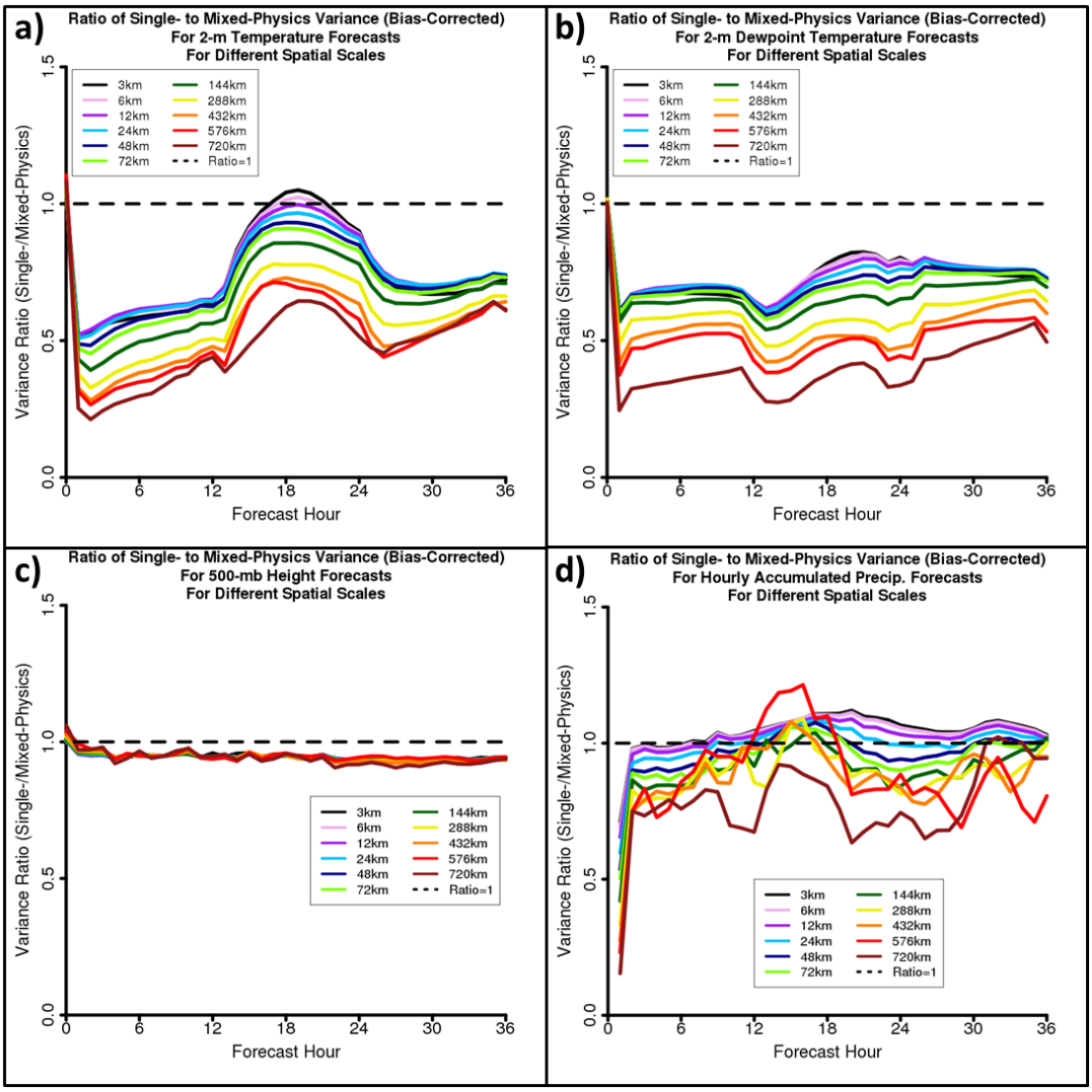

Figure 3.7 Time series of bias-corrected variance ratios (single-physics variance/mixed-physics variance) for forecasts of (a) 2-m temperature, (b) 2-m dewpoint temperature, (c) 500-mb height, and (d) hourly accumulated precipitation. For each forecast variable, eleven spatial scales are shown: 3- (black), 6- (pink), 12- (purple), 24- (light blue), 48- (dark blue), 72- (light green), 144- (dark green), 288- (yellow), 432- (orange), 576- (red), and 720-km (dark red). The black dashed line denotes where the single- to mixed-physics variance ratio is equal to 1.

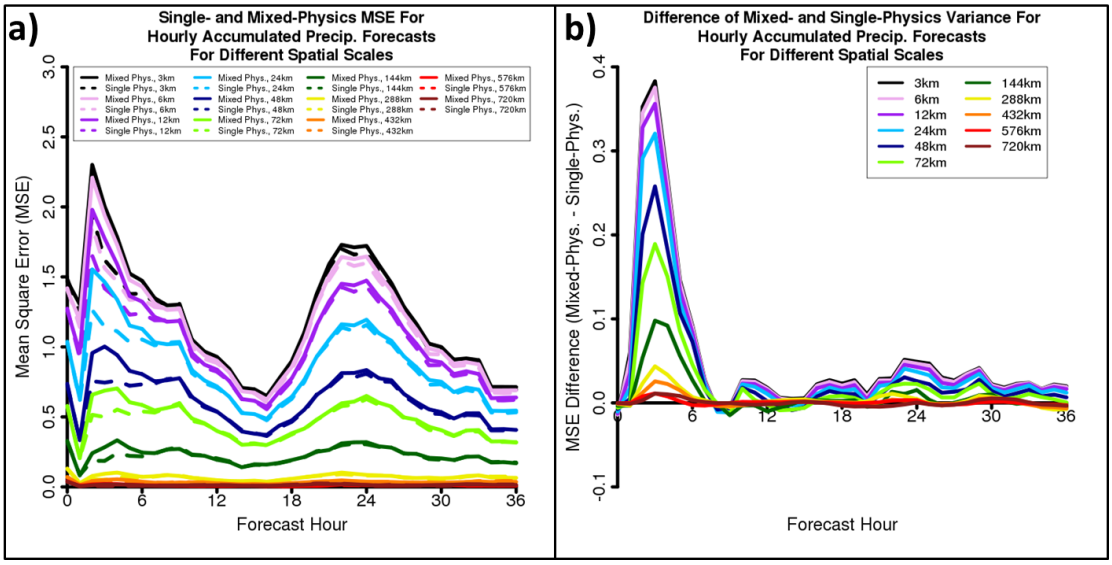

Figure 3.8 (a) Mixed- (solid) and single-physics (dashed) ensemble mean square error (MSE) time series for 3- (black), 6- (pink), 12- (purple), 24- (light blue), 48- (dark blue), 72- (light green), 144- (dark green), 288- (yellow), 432- (orange), 576- (red), and 720-km (dark red) spatial scale, and (b) MSE difference (mixed-physics MSE – single-physics MSE) time series for the same neighborhoods as in (a).

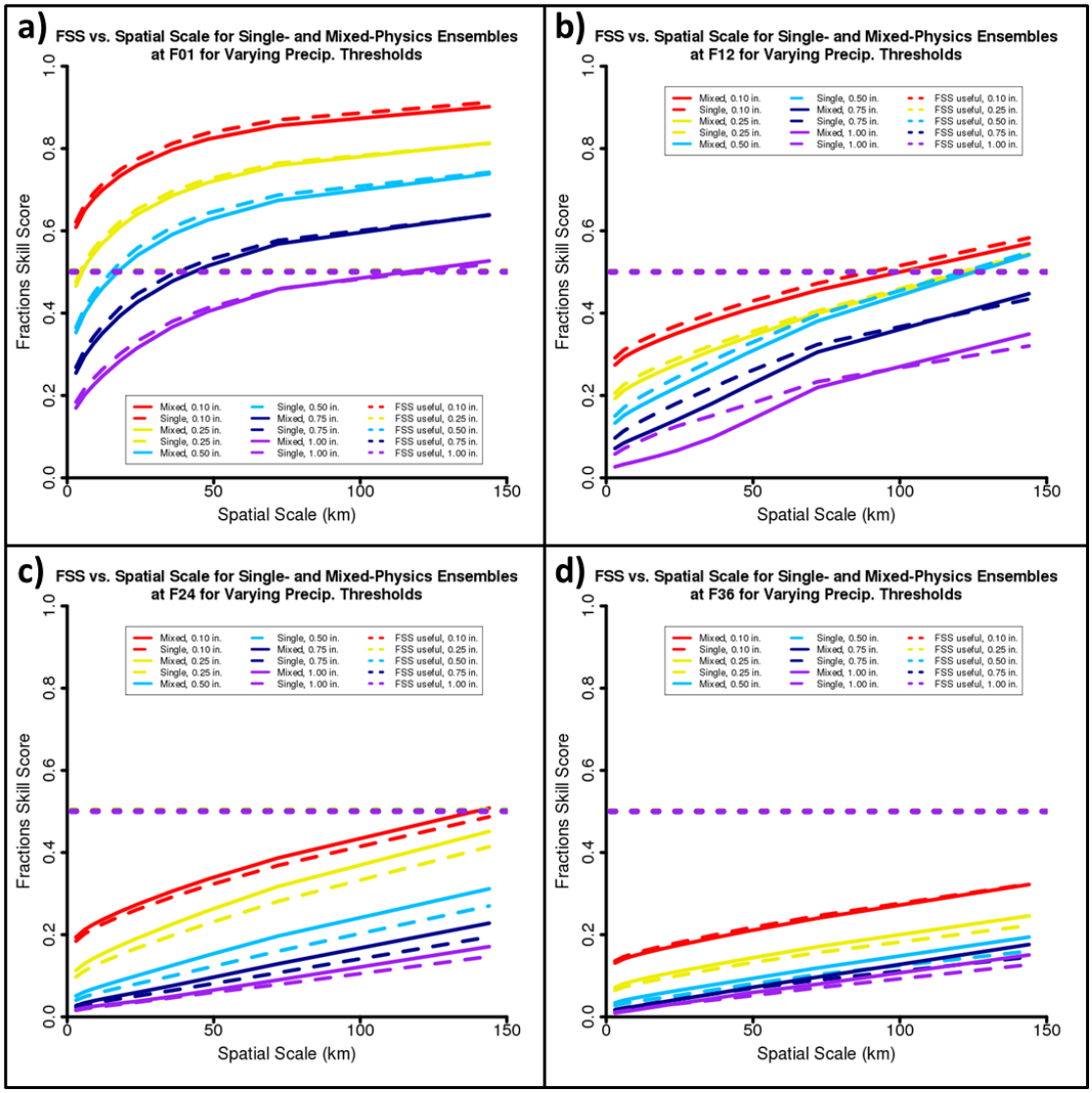

Figure 3.9 Mixed- (solid) and single-physics (long dashes) ensemble fractions skill score as a function of spatial scale at (a) forecast hour 1, (b) forecast hour 12, (c) forecast hour 24, and (d) forecast hour 36. In each case, 0.10- (red), 0.25- (yellow), 0.50- (light blue), 0.75- (dark blue), and 1.00-inch (purple) precipitation threshold forecasts are shown. FSS_{useful} values (short dashes) are also displayed for each precipitation threshold.

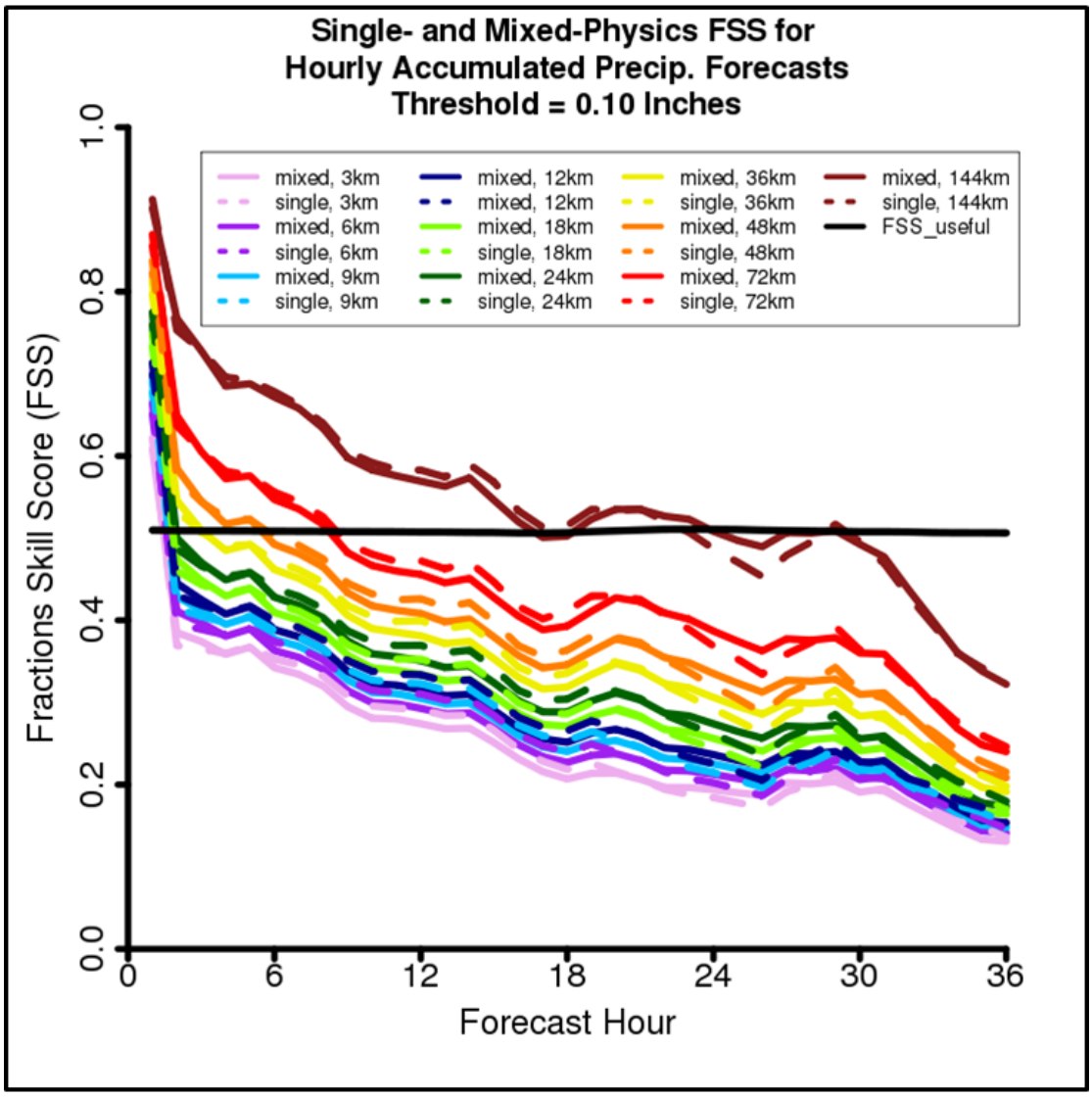

Figure 3.10 Mixed- (solid) and single-physics (dashed) ensemble fractions skill score time series for 3- (pink), 6- (purple), 9- (light blue), 12- (dark blue), 18- (light green), 24- (dark green), 36- (yellow), 48- (orange), 72- (red), and 144-km (dark red) spatial scales. FSS_{useful} (solid black) is also indicated.

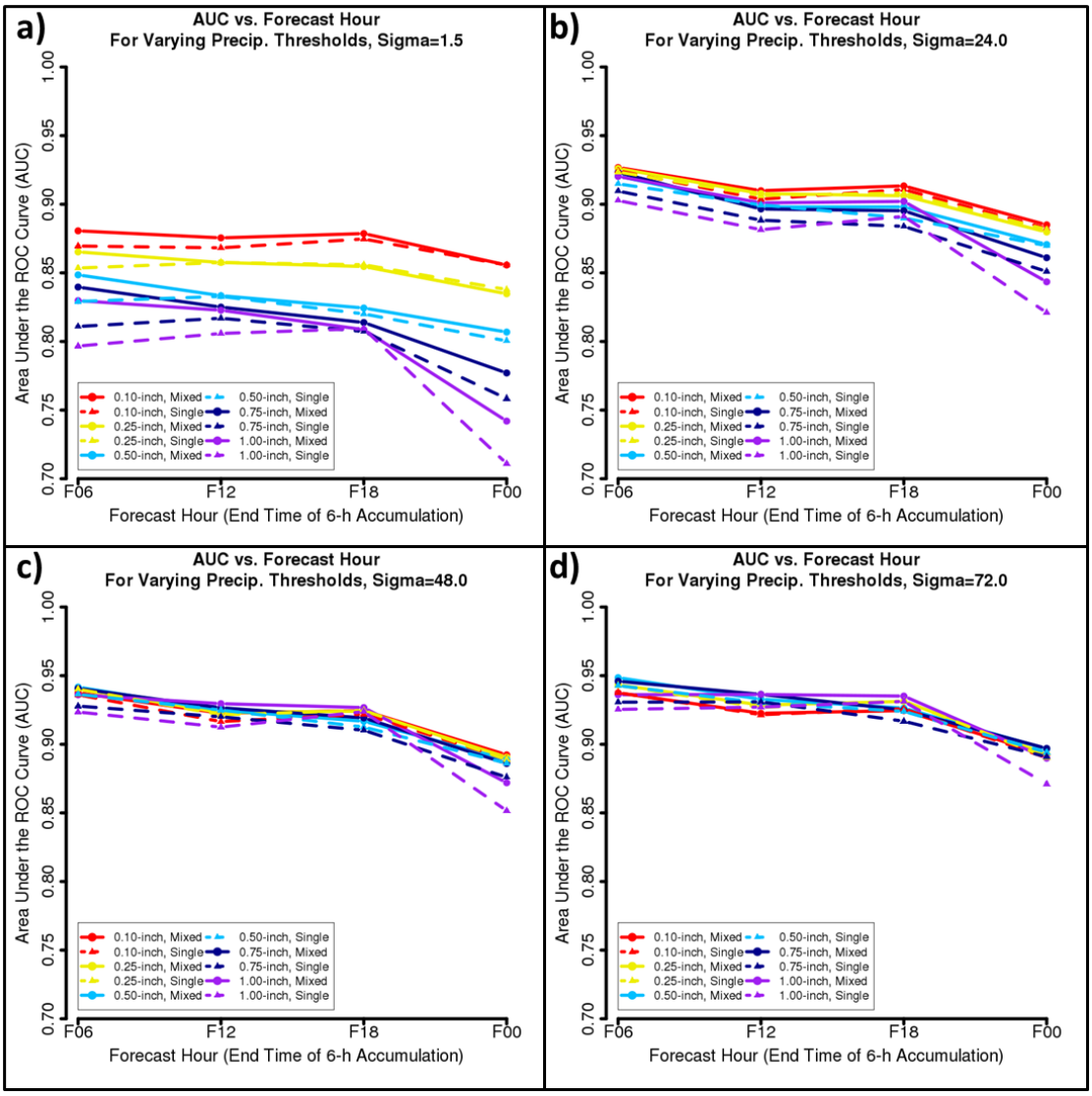

Figure 3.11 Area under the relative operating characteristics curve (AUC) for mixed- (solid with filled circles) and single-physics (dashed with filled triangles) 0.10- (red), 0.25- (yellow), 0.50- (light blue), 0.75- (dark blue), and 1.00-inch (purple) 6-hour accumulated precipitation threshold forecasts using a spatial smoothing parameter of (a) 1.5 km, (b) 24.0 km, (c) 48.0 km, and (d) 72.0 km. AUC values are shown for the 6-hour periods ending at 0600 UTC (F06), 1200 UTC (F12), 1800 UTC (F18), and 0000 UTC (F00).

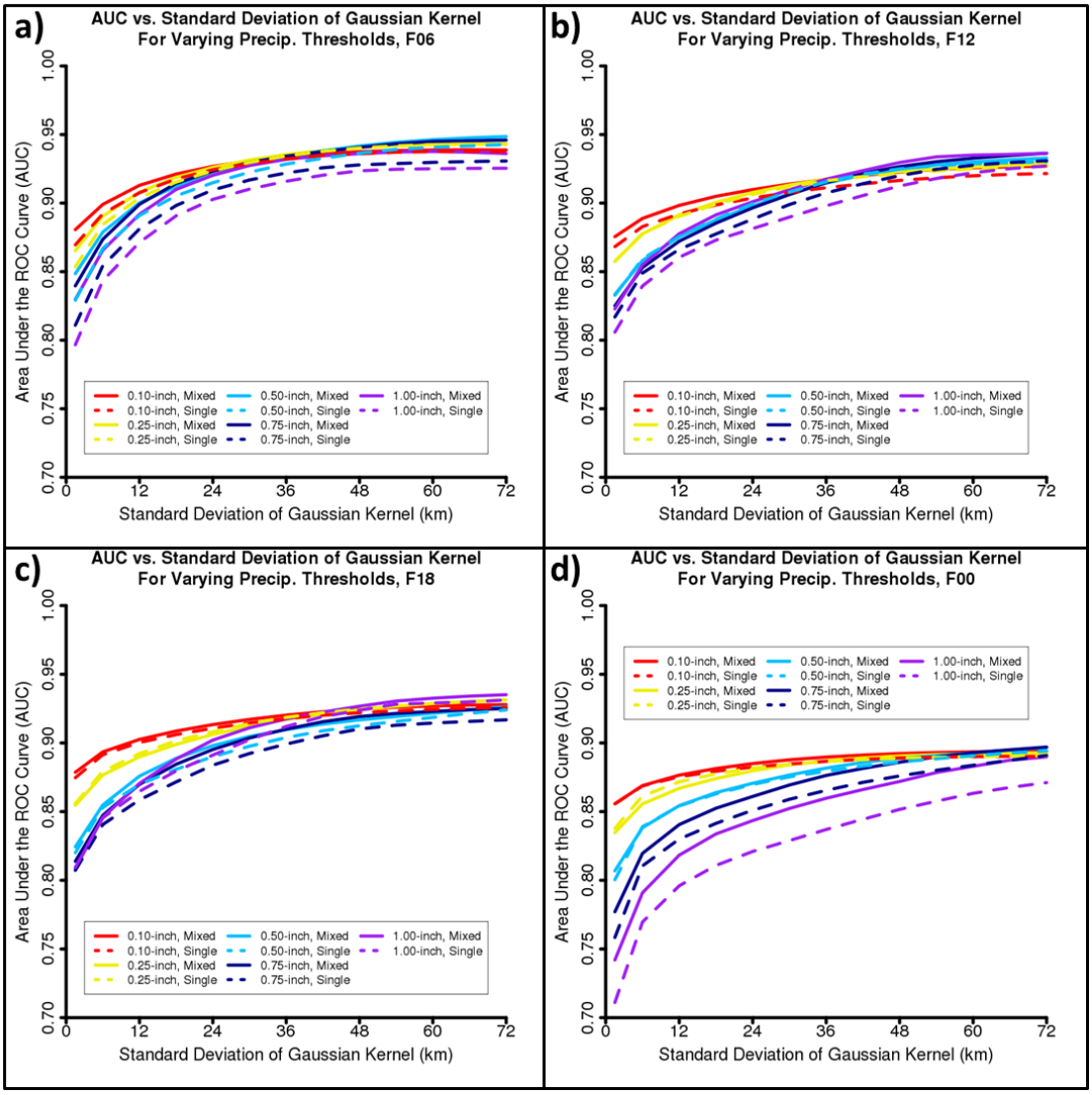

Figure 3.12 Area under the relative operating characteristics curve (AUC) for mixed- (solid) and single-physics (dashed) 0.10- (red), 0.25- (yellow), 0.50- (light blue), 0.75- (dark blue), and 1.00-inch (purple) 6-hour accumulated precipitation threshold forecasts as a function of the spatial smoothing parameter for the 6-hour period ending at: (a) 0600 UTC, (b) 1200 UTC, (c) 1800 UTC, and (d) 0000 UTC.

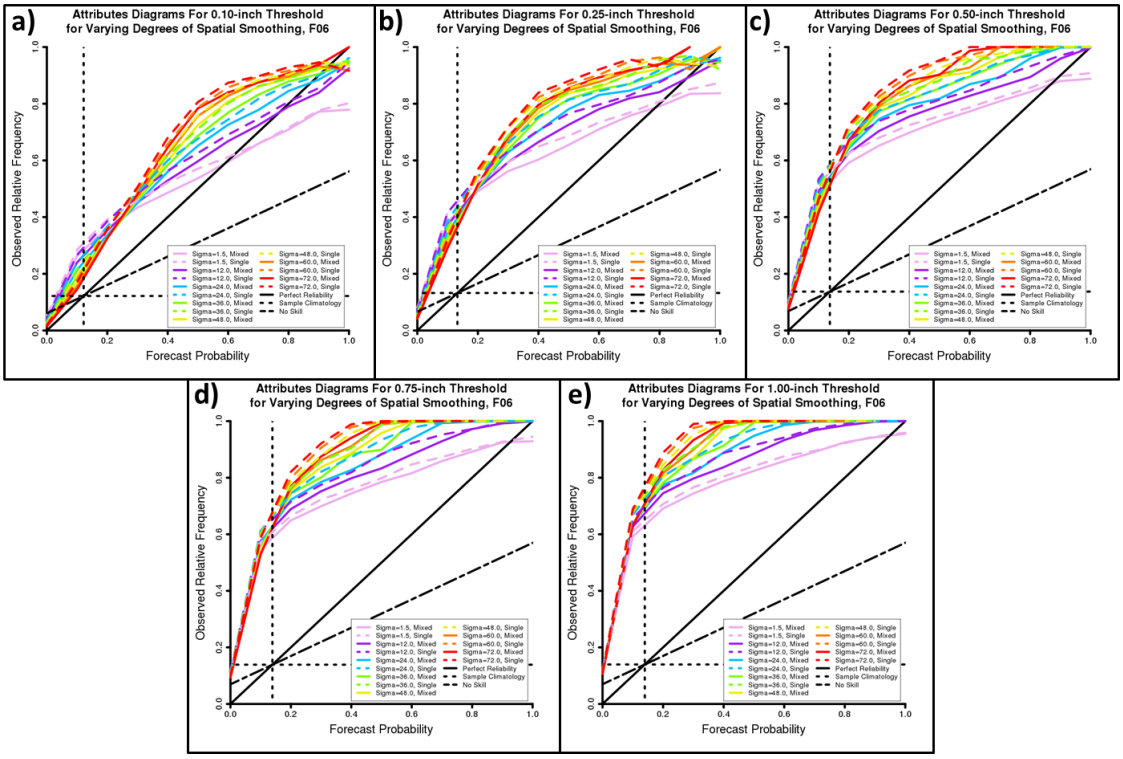

Figure 3.13 Attributes diagrams for mixed- (solid) and single-physics (dashed) ensemble 6-hour precipitation forecasts ending at 0600 UTC using a threshold of (a) 0.10 inches, (b) 0.25 inches, (c) 0.50 inches, (d) 0.75 inches, and (e) 1.00 inch. In each case, forecasts produced using a spatial smoothing parameter of 1.5- (pink), 12- (purple), 24- (light blue), 36- (light green), 48- (yellow), 60- (orange), and 72-km (red) are shown. Additionally, the line of perfect reliability (black solid), no skill (black long dashed), and lines of sample relative climatological frequency (black short dashed) are depicted in each panel.

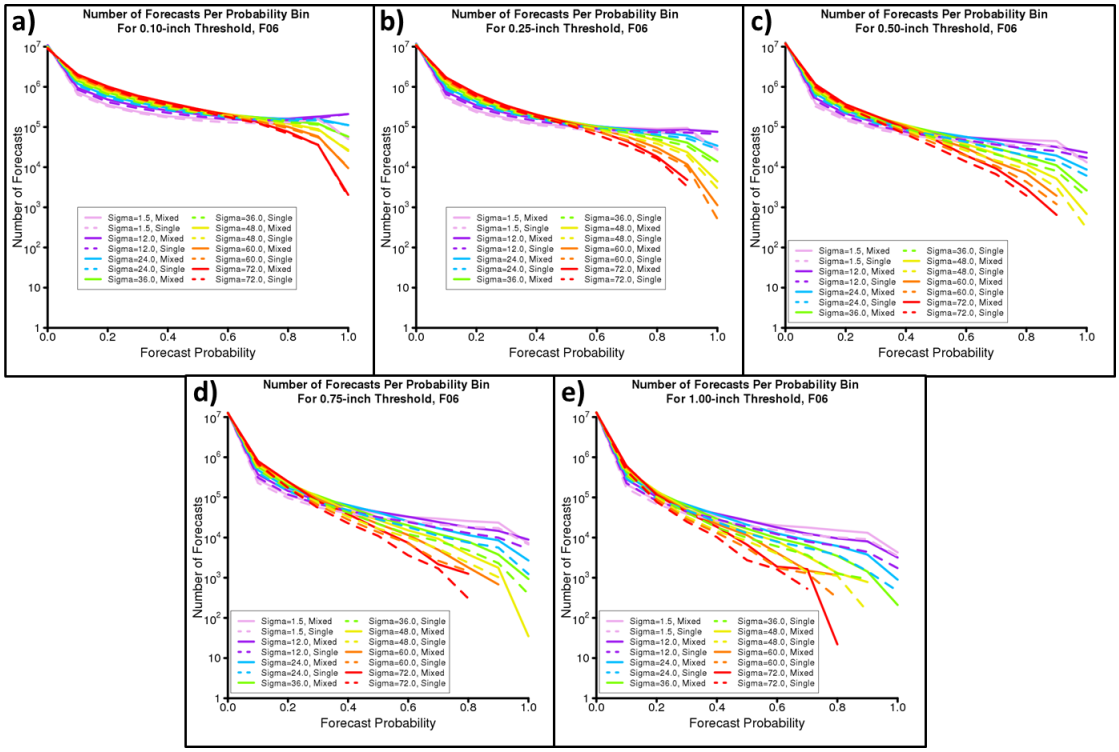

Figure 3.14 Number of forecasts per probability bin for mixed- (solid) and single-physics (dashed) ensemble 6-hour precipitation forecasts ending at 0600 UTC using a threshold of (a) 0.10 inches, (b) 0.25 inches, (c) 0.50 inches, (d) 0.75 inches, and (e) 1.00 inch. In each case, forecasts produced using a spatial smoothing parameter of 1.5- (pink), 12- (purple), 24- (light blue), 36- (light green), 48- (yellow), 60- (orange), and 72-km (red) are shown. Note the logarithmic y-axis.

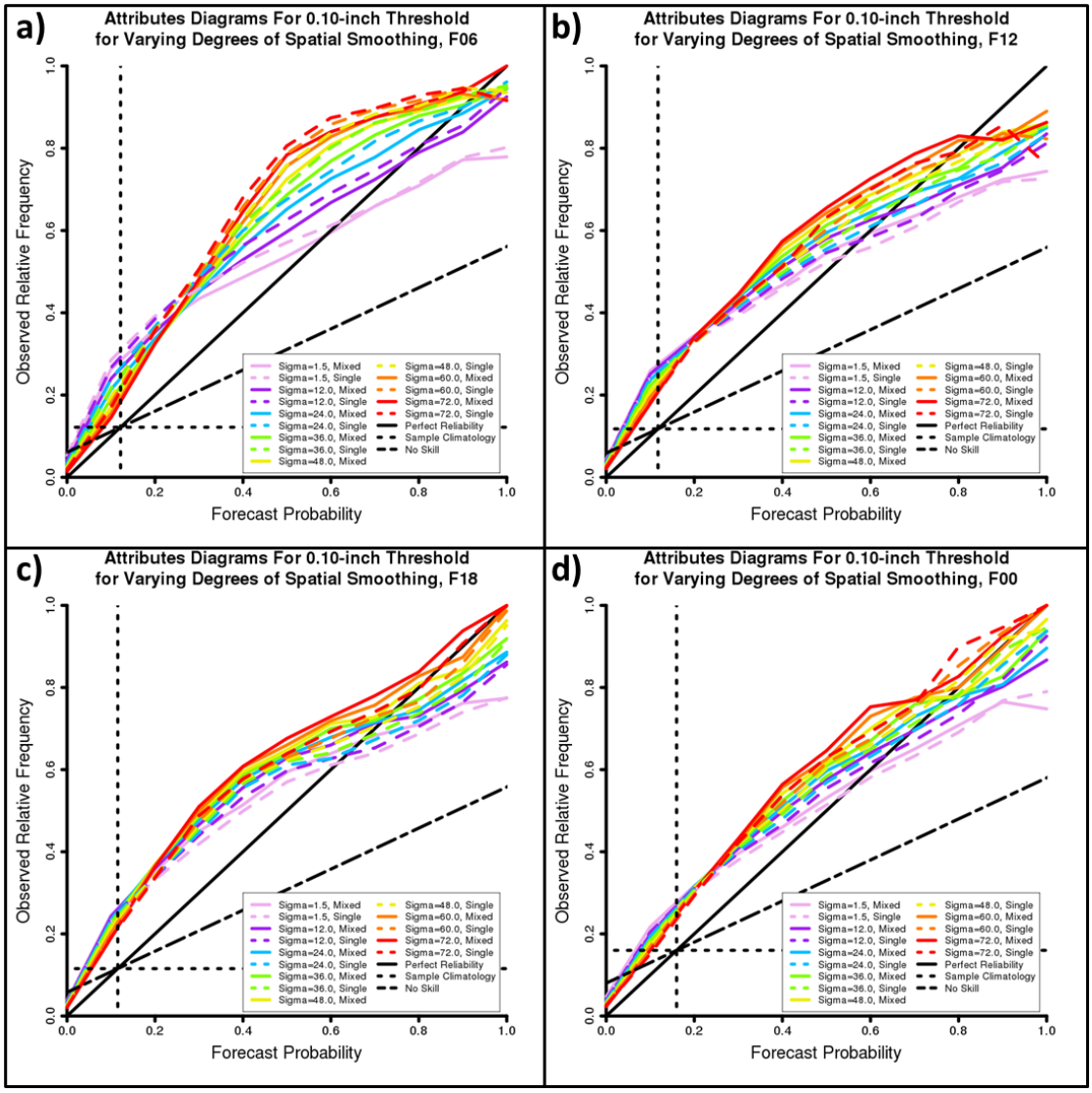

Figure 3.15 Attributes diagrams for mixed- (solid) and single-physics (dashed) ensemble 0.10-inch-threshold 6-hour accumulated precipitation forecasts for periods ending at (a) 0600 UTC, (b) 1200 UTC, (c) 1800 UTC, and (d) 0000 UTC. In each case, forecasts produced using a spatial smoothing parameter of 1.5- (pink), 12- (purple), 24- (light blue), 36- (light green), 48- (yellow), 60- (orange), and 72-km (red) are shown. Additionally, the line of perfect reliability (black solid), no skill (black long dashed), and lines of sample relative climatological frequency (black short dashed) are depicted in each panel.

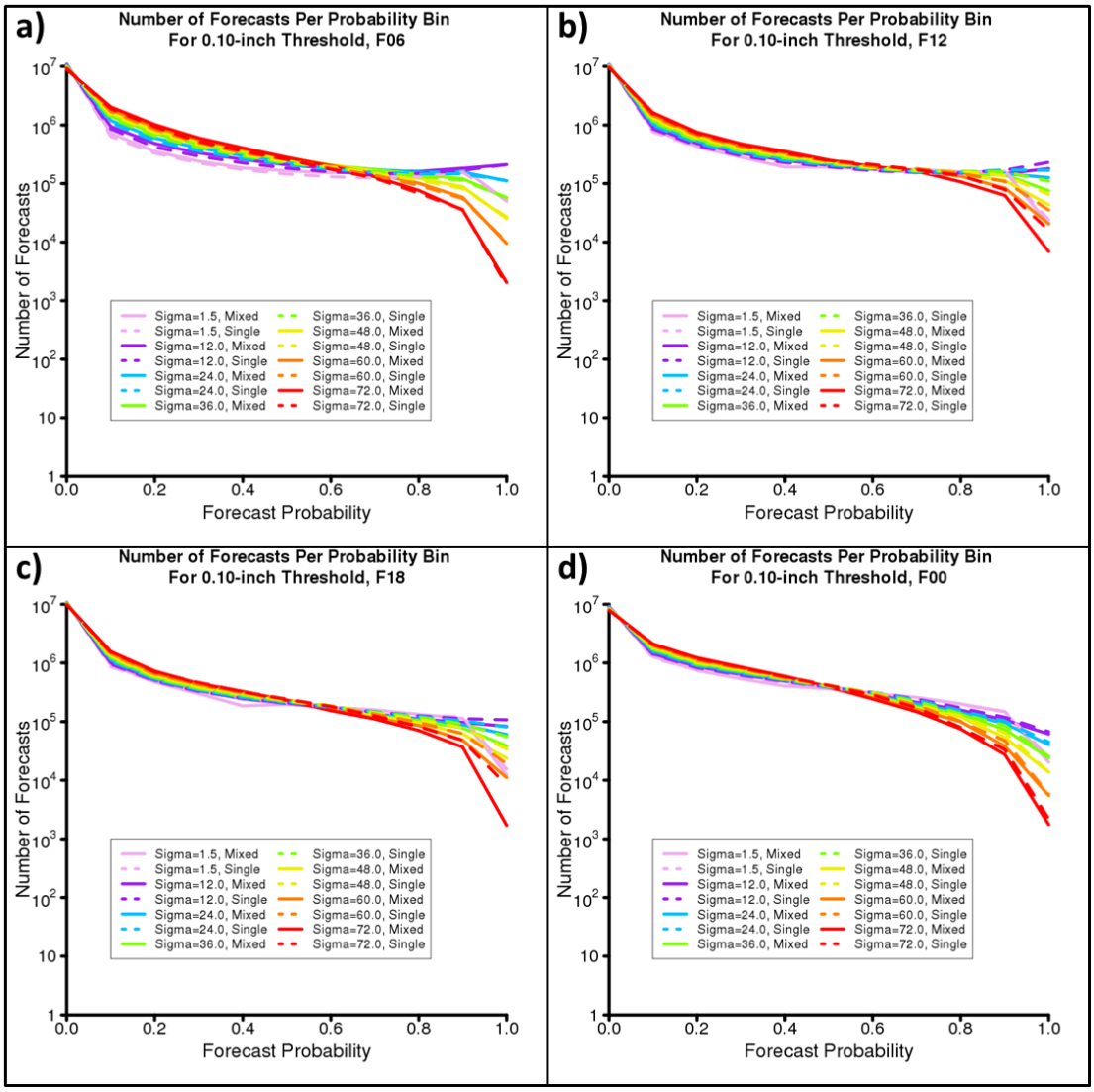

Figure 3.16 Number of forecasts per probability bin for mixed- (solid) and single-physics (dashed) ensemble 0.10-inch-threshold 6-hour accumulated precipitation forecasts for periods ending at (a) 0600 UTC, (b) 1200 UTC, (c) 1800 UTC, and (d) 0000 UTC. In each case, forecasts produced using a spatial smoothing parameter of 1.5- (pink), 12- (purple), 24- (light blue), 36- (light green), 48- (yellow), 60- (orange), and 72-km (red) are shown. Note the logarithmic y-axis.

Chapter 4: General Conclusion

1. General discussion

Over the past 30 years, convection-allowing models have evolved from a theoretical idea (e.g., Lilly et al. 1990) to an operational reality (e.g., the High Resolution Rapid Refresh (HRRR) model; Benjamin et al. 2016). Numerous studies have demonstrated the benefits that CAMs provide to forecasters, relative to convection-parameterizing models; these include a greater ability to predict storm mode, initiation, and evolution (e.g., Kain et al. 2006; Done et al. 2004; Weisman et al. 2008; Schwartz et al. 2009). More recently, convection-allowing ensembles have also shown promise for forecasting fields related to convection, such as precipitation (e.g., Schwartz et al. 2010; Clark et al. 2009; Schwartz et al. 2017). However, many questions regarding the optimal configuration of CAMs and convection-allowing ensembles have remained. For example, previous studies (e.g., Kain et al. 2008; Schwartz et al. 2009; Roberts and Lean 2008; Potvin and Flora 2015) have disagreed on how much additional forecast quality and value can be gained by reducing the horizontal grid spacing of a deterministic CAM beyond 4-km. Additionally, it has remained unclear how to choose a set of convection-allowing ensemble members to reduce under-dispersion and maximize ensemble skill (e.g., Roebber et al. 2004; Romine et al. 2014; Duda et al. 2014; Johnson and Wang 2017). Solutions to these problems have the potential to provide more efficient use of computing resources as well as increased NWP forecast skill. These benefits, in turn, have motivated research initiatives, such as the annual NOAA Hazardous Weather Testbed Spring Forecasting Experiment (HWT SFE), to investigate optimal design and configuration of CAMs and convection-allowing models. Indeed, the goal of this thesis was to use data from the 2010, 2011, and 2016 HWT

SFEs to determine how to best configure convection-allowing models and ensembles for up to next-day severe weather prediction, given current computing resources. Two research components were designed to meet this goal. The first component compared next-day probabilistic severe weather forecasts from three WRF-ARW model configurations during the 2010 and 2011 HWT SFEs: a 4-km grid-spacing deterministic CAM; an equivalently-configured 1-km deterministic CAM; and an 11-member, 4-km grid-spacing convection-allowing ensemble. The second research component used data from the 2016 Community Leveraged Unified Ensemble (CLUE) during the 2016 HWT SFE to compare the spread and skill of a 9-member mixed-physics ensemble and a 10-member single-physics ensemble.

The research question (Q1) and hypothesis (H1) associated with the first research component are, as proposed in Chapter 1:

Q1: For next-day, all-hazards severe weather forecasts derived from simulated UH, which of the following two approaches results in forecasts with higher quality and value: reducing the horizontal grid spacing of a deterministic CAM from 4 km to 1 km, or adding members to create a 4-km, 11-member CAM ensemble?

H1: While both the 1-km deterministic CAM and the 11-member, 4-km ensemble will provide greater forecast quality relative to the 4-km deterministic CAM, more quality and value will be gained by creating the 4-km ensemble than by reducing the horizontal grid spacing from 4 km to 1 km.

Given the results of, for example, Roberts and Lean (2008), VandenBerg et al. (2014), and Potvin and Flora (2015), it was expected that the deterministic 1-km grid-spacing forecasts would provide some additional forecast quality relative to the deterministic 4-km grid-spacing forecasts. However, based on the results of Kain et al. (2008) and Schwartz et al. (2009), who found no gains in quality or value for low-level reflectivity and hourly precipitation forecasts from reducing CAM grid-spacing from 4- to 2-km, it was expected that the 1-km grid-spacing forecasts would provide less additional forecast quality than the 4-km grid-spacing ensemble forecasts. The ensemble forecasts were hypothesized to give the greatest additional forecast value and quality relative to the 4-km deterministic forecasts; this hypothesis was based on previous studies (e.g., Stensrud et al. 1999; Wandishin et al. 2001; Grimit and Mass 2002) that found ensemble mean forecasts could exceed the skill of higher-resolution deterministic forecasts at convective-parameterizing resolution.

Over the aggregate 63 cases examined, the 4-km ensemble forecasts had the greatest area under the relative operating characteristics curve (AUC) of the three sets of forecasts. The difference in AUC between the 4-km ensemble forecasts and the 4-km deterministic forecasts (using UH = 25 m^2s^{-2} on the 4-km grid as the threshold) was statistically significant at the 95% level. The ensemble forecasts also had slightly better reliability relative to either deterministic forecast. Notably, the 1-km deterministic forecasts had greater—but not significantly greater AUC—relative to the 4-km deterministic forecasts. H1 was therefore generally supported, with the caveat that the 1- and 4-km deterministic forecasts had AUC values that were not significantly different. Additionally, analysis of individual cases suggested that the 4- and 1-km

deterministic forecasts routinely offered comparable forecast value, while the 4-km ensemble forecasts had the ability to provide greater value relative to either deterministic forecast.

The two research questions (Q2, Q3) and hypotheses (H2, H3) associated with the second research component are, as proposed in Chapter 1:

Q2: For each of the four variables mentioned above (i.e., hourly accumulated precipitation, 2-m temperature, 2-m dewpoint temperature, and 500-mb height), will the spread (i.e., variance) of the mixed-physics ensemble forecasts be greater than that of the single-physics ensemble forecasts at any/all spatial scales?

Q3: Will the mixed-physics ensemble produce more skillful hourly precipitation forecasts relative to the single-physics ensemble at any/all spatial scales?

H2: In general, the variance of the mixed-physics ensemble forecasts will be greater than the variance of the single-physics ensemble forecasts for the low-level variables (i.e., 2-m temperature and 2-m dewpoint temperature) and hourly accumulated precipitation but not for 500-mb height. However, as the spatial scale increases, the variance of the mixed- and single-physics forecasts will become increasingly similar for all four variables.

H3: Because of its greater member diversity, the mixed-physics ensemble will produce more skillful 6-hourly precipitation forecasts than the single-physics ensemble at

smaller spatial scales. Additionally, the mixed-physics ensemble will demonstrate skill at a smaller scale relative to the single-physics ensemble. As the spatial scale increases, the skill of the mixed- and single-physics ensemble forecasts will be increasingly similar.

In general, for 2-m temperature, 2-m dewpoint temperature, and hourly accumulated precipitation, the mixed-physics ensemble was expected to generate greater variance than the single-physics ensemble; this hypothesis was based on previous research that suggested multiple microphysics and PBL parameterizations could enhance convection-allowing ensemble spread for a variety of variables (e.g., Clark et al. 2010, Johnson and Wang 2017). However, the multiple microphysics and PBL parameterizations were not expected to dramatically increase the spread of the 500-mb geopotential height field, since the mixed-physics ensemble only differed from the single-physics ensemble in terms of microphysics parametrizations—which only impact the forecast in the vicinity of precipitation systems—and PBL schemes, which only impact the simulated lowest levels. Additionally, Clark et al. (2010) noted that mass-related fields had a smaller proportion of their spread generated due to mixed-physics compared to low-level fields. For all four variables, it was expected that the mixed- and single-physics ensembles would have more similar variances as the spatial scale of the forecast was increased, since it was expected that localized differences between the two ensembles—resulting from differences in their PBL and microphysics representations—would be averaged out at the larger scales.

While greater ensemble spread is not always associated with greater ensemble skill (e.g., Eckel and Mass 2005), in this case the mixed-physics ensemble's better

ability to account for the uncertainty in microphysics was expected to lead to its production of more skillful precipitation forecasts relative to the single-physics ensemble. This expectation was partially based on findings that simulated thunderstorms are very sensitive to microphysics parameterizations (e.g., Gilmore et al. 2004; van den Heever and Cotton 2004; Snook and Xue 2008). Based on the importance of microphysics parameterizations in simulating storm structure and evolution, it was thought that the mixed-physics ensemble—which had superior microphysics representation—would demonstrate skill down to smaller spatial scales than the single-physics ensemble. Again, the mixed- and single-physics ensembles were expected to have similar skill at larger spatial scales, as small errors in precipitation location became less important.

For the 23 cases examined, it was found that the mixed-physics ensemble generally produced greater variance than the single-physics ensemble for all four variables examined. This was nearly always true for the raw (i.e., uncalibrated) ensembles and was generally true for the bias-corrected ensembles. The biggest exception to this finding was for bias-corrected hourly precipitation; the single-physics ensemble had greater variance over multiple forecast hours for spatial scales up to 72 km. For both the raw and bias-corrected data, the differences between the mixed- and single-physics ensemble variances tended to be reduced as the spatial scale was increased. Hence, H2 was mostly supported. However, even for the 500-mb height field, the mixed-physics ensemble was found to generate more variance than the single-physics ensemble across all forecast hours.

Precipitation skill was generally similar between the forecasts. Fractions skill

score (FSS) tended to be similar at all spatial scales and at all forecast hours. While the single-physics ensemble had slightly greater FSSs at most forecast hours, the mixed-physics ensemble had greater—but not substantially greater—FSSs around forecast hour 24. For the 6-hourly accumulated precipitation forecasts, the mixed-physics ensemble nearly always had greater AUC values than the single-physics ensemble. These differences in AUC tended to be greatest for forecast periods in the late afternoon and evening and for the largest precipitation thresholds, suggesting that the mixed-physics ensemble had the greatest relative skill for heavy/moderate rainfall and during periods when rainfall is climatologically maximized. However, the differences in AUC between the mixed- and single-physics ensembles were generally small. Moreover, the differences were not largely affected as the spatial scale was increased. Therefore, H3 was mostly unsupported.

2. Recommendations for future research

One limitation of the findings presented in this thesis is that both research components were conducted using data restricted to the spring season (i.e., late April to mid-June). Other seasons have different dominant flow regimes and precipitation and severe weather climatologies (e.g., Kelly et al. 1985; Markham 1970), which may impact the results obtained in each research component. A variety of preliminary investigations seem to bear this out. For example, Row and Correia (2014) found that forecast UH-derived surrogate severe weather forecasts from the SSEO performed worse in August than during the springtime. Meanwhile, Hitchens et al. (2016) found that day-1 convective outlooks generated from forecast UH performed best in the spring

and summer and worse in the fall and winter. Therefore, it is recommended that future studies use an expanded dataset that includes the entire calendar year to determine if the results obtained herein generalize to the summer, fall, and winter seasons.

Additionally, while the two research component presented herein provided numerous objective measures of forecast skill, it is recommended that future research also include a component to measure subjective skill and value to forecasters. Although the subjective value of a forecast can be partially inferred through the analysis of individual forecast fields for each model configuration on a given day (or time), it is possible that the forecasts are used differently than anticipated. For example, it is possible that two forecasts, which produce different forecast fields with different objective skill scores, ultimately provide the same guidance to an operational forecaster. In such a case, despite the objective skill metrics, it may make sense to implement the forecast configuration that requires the lowest amount of computing power to produce.

References

Bauer, P., A. Thorpe, and G. Brunet, 2015: The quiet revolution of numerical weather prediction. *Nature*, **525**, 47-55, doi: 10.1038/nature14956.

Benjamin, S., 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.

Bergthorsson, P., B. R. Doos, S. Fryklund, O. Haug, and R. Lindquist, 1955: Routine forecasting with the barotropic model. *Tellus*, **7**, 272-274, doi: 10.1111/j.2153-3490.1955.tb01162.x.

Berner, J., S. Ha, J. P. Hacker, A. Fournier, and C. Snyder, 2011: Model uncertainty in a mesoscale ensemble prediction system: Stochastic versus multiphysics representations. *Mon. Wea. Rev.*, **139**, 1972–1995, doi: 10.1175/2010MWR3595.1.

Black, T. L., 1994: The new NMC mesoscale Eta Model: Description and forecast examples. *Wea. Forecasting*, **9**, 265–278, doi: 10.1175/1520-0434(1994)009<0265:TNNMEM>2.0.CO;2.

Bolin, B., 1955: Numerical forecasting with the barotropic model. *Tellus*, **7**, 1, 27-49, doi: 10.1111/j.2153-3490.1955.tb01139.x.

Brooks, H. E., C. A. Doswell III, and R. A. Maddox, 1992: On the use of mesoscale and cloud-scale models in operational forecasting. *Wea. Forecasting*, **7**, 120–132.

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

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.

Bushby, F. H. and M. S. Timpson, 1967: A 10-level atmospheric model and frontal rain. *Quart. J. Roy. Meteor. Soc.*, **93**, 562-564, doi: 10.1002/qj.49709339825.

Bushby, F. H., 1986: A history of numerical weather prediction. *Journal of the Meteorological Society of Japan*, **64A**, 1-10.

Charney, J. G., R. Fjortoft, and J. von Neumann, 1950: Numerical integration of the barotropic vorticity equation. *Tellus*, **2**, 237-254, doi: 10.1111/j.2153-3490.1950.tb00336.x.

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.

Chou, M.-D., and M. J. Suarez, 1994: An efficient thermal infrared radiation parameterization for use in general circulation models. NASA Tech. Memo. 104606, 85 pp.

Clark, A. J., W. A. Gallus Jr., and T.-C. Chen, 2008: Contributions of mixed physics versus perturbed initial/lateral boundary conditions to ensemble-based precipitation forecast skill. *Mon. Wea. Rev.*, **136**, 2140–2156, doi: 10.1175/2007MWR2029.1.

Clark, A. J., W. A. Gallus, 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, doi: 10.1175/2009WAF2222222.1.

Clark, A. J., W. A. Gallus Jr., M. Xue, and F. Kong, 2010: Growth of spread in convection-allowing and convection-parameterizing ensembles. *Wea. Forecasting*, **25**, 594–612, doi: 10.1175/2009WAF2222318.1.

Clark, A.J., and Coauthors, 2011: Probabilistic precipitation forecast skill as a function of ensemble size and spatial scale in a convection-allowing ensemble. *Mon. Wea. Rev.*, **139**, 1410–1418, doi: 10.1175/2010MWR3624.1.

Clark, A. J., and Coauthors, 2012a: A supplement to an overview of the 2010 hazardous weather testbed experimental forecast program spring experiment. *Bull. Amer. Meteor. Soc.*, **93**, ES1-ES5, doi: 10.1175_bams-d-11-00040.

Clark, A. J., and Coauthors, 2012b: An overview of the 2010 Hazardous Weather Testbed Experimental Forecast Program Spring Experiment. *Bull. Amer. Meteor. Soc.*, **93**, 55–74, doi: 10.1175/BAMS-D-11-00040.1.

Clark, A. J., J. S. Kain, P. T. Marsh, J. Correia, M. Xue,
and F. Kong, 2012c: Forecasting tornado pathlengths using a three-dimensional object identification algorithm applied to convection-allowing forecasts. *Wea. Forecasting*, **27**, 1090–1113, doi: 10.1175/WAF-D-11-00147.1.

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.

Clark, A., and Coauthors, 2016: Spring forecasting experiment 2016 conducted by the experimental forecast program of the NOAA/Hazardous weather testbed: Program overview and operations plan. Available online at: https://hwt.nssl.noaa.gov/Spring_2016/HWT_SFE2016_operations_plan_final.p

df. Accessed 23 Jun 2017.

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, doi: 10.1175/2009WAF2222258.1.

Dai, A., F. Giorgi, and K. E. Trenberth, 1999: Observed and model-simulated diurnal cycles of precipitation over the contiguous United States. *Journal of Geophysical Research: Atmospheres*, **104**, 6377-6402.

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. Lett.*, **5**, 110–117, doi: 10.1002/asl.72.

Du, J., J. McQueen, G. DiMego, Z. Toth, D. Jovic, B. Zhou, and H.-Y. Chuang, 2006: New dimension of NCEP Short-Range Ensemble Forecasting (SREF) system: Inclusion of WRF members. *WMO Expert Team Meeting on Ensemble Prediction Systems*, Exeter, United Kingdom, WMO. [Available online at http://www.emc.ncep.noaa.gov/mmb/SREF/WMO06_full.pdf].

Du, J., and Coauthors, 2014: NCEP regional ensemble update: Current systems and planned storm-scale ensembles. Preprints, *26th Conf. on Wea. Forecasting*, Atlanta, GA, Amer. Meteor. Soc., J.1.4.

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, doi: 10.1175/MWR-D-13-00297.1.

Easterling, D. R., and P. J. Robinson, 1985: The diurnal variation of thunderstorm activity in the United States. *J. Climate Appl. Meteor.*, **24**, 1048–1058, doi: 10.1175/1520-0450(1985)024<1048:TDVOTA>2.0.CO;2.

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, doi: 10.1175/1520-0493(2001)129<2461:AOAPMS>2.0.CO;2.

Ebert, E. E., 2008: Fuzzy verification of high-resolution gridded forecasts: A review and proposed framework. *Meteorological Applications*, **15**, 51-64.

Ebert, E. E., 2009: Neighborhood verification: A strategy for rewarding close forecasts. *Wea. Forecasting*, **24**, 1498–1510.

Eckel, F. A., and C. F. Mass, 2005: Aspects of effective mesoscale, short-range ensemble forecasting. *Wea. Forecasting*, **20**, 328–350, doi: 10.1175/WAF843.1.

Epstein, E. S., 1969a: The role of initial uncertainties in prediction. *J. Appl. Meteor.*, **8,** 190–198.

Epstein, E. S., 1969b: Stochastic dynamic prediction. *Tellus*, **21,** 739–759.

Ferrier, B. S., Y. Jin, Y. Lin, T. Black, E. Rogers, and G. DiMego, 2002: Implementation of a new grid-scale cloud and rainfall scheme in the NCEP Eta Model. Preprints, *19th Conf. on Weather Analysis and Forecasting/15th Conf. on Numerical Weather Prediction*, San Antonio, TX, Amer. Meteor. Soc., 10.1. [Available online at http:// ams.confex.com/ams/SLS_WAF_NWP/techprogram /paper_47241.htm.]

Gallo, B. T., A. J. Clark, and S. R. Dembek, 2016: Forecasting tornadoes using convection-permitting ensembles. *Wea. Forecasting*, **31**, 273–295, doi: 10.1175/WAF-D-15-0134.1.

Gallus, W. A., Jr., and J. F. Bresch, 2006: Comparison of impacts of WRF dynamic core, physics package, and initial conditions on warm season rainfall forecasts. *Mon. Wea. Rev.*, **134**, 2632–2641.

Gao, J., M. Xue, K. Brewster, and K. K. Droegemeier, 2004: A three-dimensional variational data analysis method with recursive filter for Doppler radars. *J. Atmos. Oceanic Technol.*, **21**, 457–469.

Gilmore, M. S., J. M. Straka, and E. N. Rasmussen, 2004: Precipitation uncertainty due to variations in precipitation particle parameters within a simple microphysics scheme. *Mon. Wea. Rev.*, **132**, 2610–2627, doi: 10.1175/MWR2810.1.

Grimit, E. P., and C. F. Mass, 2002: Initial results of a mesoscale short-range ensemble forecasting system over the Pacific Northwest. *Wea. Forecasting,* **17,** 192–205.

Hamill, T. M., 1999: Hypothesis tests for evaluating numerical precipitation forecasts. *Wea. Forecasting,* **14,** 155–167.

Hitchens, N. M., R. A. Sobash, and A. J. Clark, 2016: A multi-year evaluation of NSSL-WRF surrogate severe thunderstorm forecasts. *28th Conference on Severe Local Storms*, Portland, OR, Amer. Meteor. Soc., P. 111. [Available online at https://ams.confex.com/ams/28SLS/webprogram/Paper300988.html].

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.

Hou, D., E. Kalnay, and K. K. Droegemeier, 2001: Objective verification of the SAMEX ’98 ensemble forecasts. *Mon. Wea. Rev.*, **129**, 73–91, doi: 10.1175/1520-0493(2001)129<0073:OVOTSE>2.0.CO;2.

Hsu W.-R., and A. H. Murphy, 1982: The attributes diagram: A geometrical framework for assessing the quality of probability forecasts. *Int. J. Forecasting*. **2**, 285-293.

Janjić, Z. I., 2002: Nonsingular implementation of the MellorYamada level 2.5 scheme in the NCEP Meso Model. NCEP Office Note 437, 61 pp.

Jirak, I. L., S. J. Weiss, and C. J. Melick, 2012: The SPC storm-scale ensemble of opportunity: Overview and results from the 2012 Hazardous Weather Testbed Spring Forecasting Experiment. *Proc. 26th Conf. Severe Local Storms*, Nashville, TN, Amer. Meteor. Soc., 137. [Available online at https://ams.confex.com/ams/26SLS/webprogram/Paper211729.html].

Jirak, I. L., C. J. Melick, and S. J. Weiss, 2016: Comparison of the SPC storm-scale ensemble of opportunity to other convection-allowing ensembles for severe weather forecasting. Preprints, *28th Conf. Severe Local Storms*, Portland, OR, Amer. Meteor. Soc., 102. [Available online at https://ams.confex.com/ams/28SLS/webprogram/Paper300910.html].

Johnson, A., and X. Wang, 2012: Verification and calibration of neighborhood and object-based probabilistic precipitation forecasts from a multimodel convection-allowing ensemble. *Mon. Wea. Rev.*, **140**, 3054–3077, doi: https://doi.org/10.1175/MWR-D-11-00356.1.

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, doi: 10.1175/MWR-D-13-00027.1.

Johnson, A., and Xuguang Wang, 2017: Design and implementation of a GSI-based convection-allowing ensemble data assimilation and forecast system for the PECAN field experiment. Part I: Optimal configurations for nocturnal convection prediction using retrospective cases. *Wea. Forecasting*, **32**, 289-315.

Kain, J. S., S. J. Weiss, J. J. Levit, M. E. Baldwin, and D. R. Bright, 2006: Examination of convection-allowing configurations of the WRF Model for the prediction of severe convective weather: The SPC/NSSL Spring Program 2004. *Wea. Forecasting*, **21**, 167–181, doi: 10.1175/WAF906.1.

Kain, J. S., and Coauthors, 2008: Some practical considerations regarding horizontal resolution in the first generation of operational convection-allowing NWP. *Wea. Forecasting*, **23**, 931–952, doi: 10.1175/WAF2007106.1.

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**:5, 1536-1542.

Kelly, D. L., J. T. Schaefer, and C. A. Doswell III, 1985: Climatology of nontornadic severe thunderstorm events in the United States. *Mon. Wea. Rev.*, **113**, 1997–2014, doi: 10.1175/1520-0493(1985)113<1997:CONSTE>2.0.CO;2.

Kong, F., and Coauthors, 2014: CAPS storm-scale ensemble forecasting system: Impact of IC and LBC perturbations. Preprints, *26th WAF/22nd NWP Conf*, Atlanta, GA, Amer. Meteor. Soc., Paper 119.

Leith, C. E., 1974: Theoretical skill of Monte Carlo forecasts. *Mon. Wea. Rev.*, **102,** 409–418.

Leutbecher, M., and T. N. Palmer, 2008: Ensemble forecasting. *Journal of Computational Physics*, **227**, 3515-3539.

Lilly, D. K., 1990: Numerical prediction of thunderstorms—Has its time come? *Quart. J. Roy. Meteor. Soc.*, **116**, 779–798, doi: 10.1002/qj.49711649402.

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.

Lin, Y. 2011. GCIP/EOP Surface: Precipitation NCEP/EMC 4KM Gridded Data (GRIB) Stage IV Data. Version 1.0. UCAR/NCAR - Earth Observing Laboratory. https://data.eol.ucar.edu/dataset/21.093. Accessed 23 Jun 2017.

Loken, E., A. 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*. doi: 10.1175/WAF-D-16-0200.1, in press.

Lorenz, E. N., 1963: Deterministic nonperiodic flow. *J. Atmos. Sci.*, **20**, 130–141, doi: 10.1175/1520-0469(1963)020<0130:DNF>2.0.CO;2.

Lorenz, E. N., 1965: A study of the predictability of a 28-variable atmospheric model. *Tellus*, **17**, 321-333.

Lorenz, E. N., 1969: The predictability of a flow which possesses many scales of motion. *Tellus*, **21**, 289–307.

Lynch, P., 2008: The origins of computer weather predictions and climate modeling. *Journal of Computational Physics*, **227**, 3431-3444.

Markham, C. G., 1970: Seasonality of precipitation in the United States. *Annals of the Association of American Geographers*, **60**, 593-597.

Marzban, C., 2004: The ROC curve and the area under it as performance measures.

Wea. Forecasting, **19**, 1106–1114.

Mason, S. J., 1982: A model for assessment of weather forecasts. *Aust. Met. Mag.*, **30**, 291-303.

Mason, S. J., and N. E. Graham, 2002: Areas beneath the relative operating characteristics (ROC) and relative operating levels (ROL) curves: Statistical significance and interpretation. *Quart. J. Roy. Meteor. Soc.*, **128**, 2145–2166.

Mass, C. F., D. Ovens, K. Westrick, and B. A. Colle, 2002: Does increasing horizontal resolution produce more skillful forecasts? *Bull. Amer. Meteor. Soc.*, **83**, 407–430.

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

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.

Mittermaier, M., and N. Roberts, 2010: Intercomparison of spatial forecast verification methods: Identifying skillful spatial scales using the fractions skill score. *Wea. Forecasting*, **25**, 343–354, doi: 10.1175/2009WAF2222260.1.

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.

Molteni, F., R. Buizza, T. N. Palmer, and T. Petroliagis, 1996: The ECMWF Ensemble Prediction System: Methodology and validation. *Quart. J. Roy. Meteor. Soc.*, **122**, 73–119, doi: 10.1002/qj.49712252905.

Morrison, H., J. A. Curry, and V. I. Khvorostyanov, 2005: A new double-moment microphysics parameterization for application in cloud and climate models. Part I: Description. *J. Atmos. Sci.*, **62**, 1665–1677.

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, doi: 10.1175/1520-0434(1993)008<0281:WIAGFA>2.0.CO;2.

Nakanishi, M., 2000: Large-eddy simulation of radiation fog. *Bound.-Layer Meteor.*, **94**, 461–493.

Nakanishi, M., 2001: Improvement of the Mellor-Yamada turbulence closure model based on large-eddy simulation data. *Bound.-Layer Meteor.*, **99**, 349–378.

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

Nakanishi, M., and Niino, H., 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.

Noh, Y., W. G. Cheon, S.-Y. Hong, and S. Raasch, 2003: Improvement of the K-profile model for the planetary boundary layer based on large eddy simulation data. *Bound.-Layer Meteor.*, **107**, 401–427.

Petersen, R. A., and J. D. Stackpole, 1989: Overview of the NMC production suite. *Wea. Forecasting*, **4**, 313-322.

Poincare, H., 1914: *Science and Method*. Thomas Nelson and Sons.

Potvin, C., and M. Flora, 2015: Sensitivity of idealized supercell simulations to horizontal grid spacing: Implications for warn-on-forecast. *Mon. Wea. Rev.*, **143**, 2998–3024, doi: 10.1175/MWR-D-14-00416.1.

Roberts, N. M., and H. W. Lean, 2008: Scale-selective verification of rainfall accumulations from high-resolution forecasts of convective events. *Mon. Wea. Rev.*, **136**, 78–97, doi: 10.1175/2007MWR2123.1.

Roebber, P. J., D. M. Schultz, B. A. Colle, and D. J. Stensrud, 2004: Toward improved prediction: High-resolution and ensemble modeling systems in operations. *Wea. Forecasting*, **19**, 936–949, doi: 10.1175/1520-0434(2004)019<0936:TIPHAE>2.0.CO;2.

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

Romine, G. S., C. S. Schwartz, J. Berner, K. R. Fossell, C. Snyder, J. L. Anderson, and M. L. Weisman, 2014: Representing forecast error in a convection-permitting ensemble system. *Mon. Wea. Rev.*, **142**, 4519–4541, doi: 10.1175/MWR-D-14-00100.1.

Row, M., J., and J. Correia Jr., 2014: Verification of proxy severe weather reports from updraft helicity. *26th Conference on Weather Analysis and Forecasting/22nd*

Conference on Numerical Weather Prediction, Atlanta, GA, Amer. Meteor. Soc., J11.3. [Available online at https://ams.confex.com/ams/94Annual/webprogram/Paper234785.html].

Schwartz, C. S., and Coauthors, 2009: Next-day convection-allowing WRF model guidance: A second look at 2-km versus 4-km grid spacing. *Mon. Wea. Rev.*, **137**, 3351–3372, doi: 10.1175/2009MWR2924.1.

Schwartz, C. S., and Coauthors, 2010: Toward improved convection-allowing ensembles: Model physics sensitivities and optimizing probabilistic guidance with small ensemble membership. *Wea. Forecasting*, **25**, 263–280, doi: 10.1175/2009WAF2222267.1.

Schwartz, C. S., Z. Liu, 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, doi: 10.1175/WAF-D-13-00145.1.

Schwartz, C. S., G. S. Romine, R. A. Sobash, K. R. Fossell, and M. L. Weisman, 2015a: NCAR's experimental real-time convection-allowing ensemble prediction system. *Wea. Forecasting*, **30**, 1645–1654, doi: 10.1175/WAF-D-15-0103.1.

Schwartz, C. S., G. S. Romine, M. L. Weisman, R. A. Sobash, K. R. Fossell, K. W. Manning, and S. B. Trier, 2015b: 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., G. Romine, K. Fossell, R. Sobash, and M. Weisman, 2017: Toward 1-km ensemble forecasts over large domains. Mon. Wea. Rev. doi: 10.1175/MWR-D-16-0410.1, in press.

Skamarock, W. C., and Coauthors, 2008: A description of the Advanced Research WRF version 3. NCAR Tech. Note NCAR/TN-475+STR, 113 pp. [Available online at www.mmm.ucar.edu/wrf/users/docs /arw_v3.pdf].

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.

Smirnova, T. G., J. M. Brown, S. G. Benjamin, and D. Kim, 2000: Parameterization of cold-season processes in the MAPS land-surface scheme. *J. Geophys. Res*., **105** (D3), 4077–4086.

Snook, N., and M. Xue, 2008: Effects of microphysical drop size distribution on tornadogenesis in supercell thunderstorms. *Geophys. Res. Lett.*,**35,** L24803, doi: 10.1029/2008GL035866.

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., G. S. Romine, C. S. Schwartz, D. J. Gagne, and M. L. Weisman, 2016a: Explicit forecasts of low-level rotation from convection-allowing models for next-day tornado prediction. *Wea. Forecasting*, **31**, 1591-1614.

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

Stensrud, D. J., H. E. Brooks, J. Du, M. S. Tracton, and E. Rogers, 1999: Using ensembles for short-range forecasting. *Mon. Wea. Rev*, **127,** 433–446.

Stensrud, D. J., J.-W. Bao, and T. T. Warner, 2000: Using initial condition and model physics perturbations in short-range ensemble simulations of mesoscale convective systems. *Mon. Wea. Rev.*, **128**, 2077–2107, doi: 10.1175/1520-0493(2000)128<2077:UICAMP>2.0.CO;2.

Stensrud, D. J., and Coauthors, 2009: Convective-scale warn-on-forecast system: A vision for 2020. *Bull. Amer. Meteor. Soc.*, **90**, 1487–1499, doi: 10.1175/2009BAMS2795.1.

Sukoriansky, S., B. Galperin, and V. Perov, 2006: A quasinormal scale elimination model of turbulence and its application to stably stratified flows. *Nonlinear Processes Geophys*., **13**, 9–22.

Thompson, P. D., 1957: Uncertainty of initial state as a factor in the predictability of large scale atmospheric flow patterns. *Tellus* **9**, 275-295.

Thompson, G., R. M. Rasmussen, and K. Manning, 2004: Explicit forecasts of winter precipitation using an improved bulk microphysics scheme. Part I: Description and sensitivity analysis. *Mon. Wea. Rev.*, **132**, 519–542, doi: 10.1175/1520-0493(2004)132<0519:EFOWPU>2.0.CO;2.

Toth, Z., and E. Kalnay, 1993: Ensemble forecasting at NMC: The generation of perturbations. *Bull. Amer. Meteor. Soc.*, **74**, 2317–2330, doi: 10.1175/15200477(1993)074<2317:EFANTG>2.0.CO;2.

Toth, Z., and E. Kalnay, 1997: Ensemble forecasting at NCEP and the breeding method. *Mon. Wea. Rev.*, **125**, 3297–3319, doi: 10.1175/1520-0493(1997)125<3297:EFANAT>2.0.CO;2.

Tracton, M. S., and E. Kalnay, 1993: Operational ensemble prediction at the National Meteorological Center: Practical aspects. *Wea. Forecasting*, **8,** 378–398.

van den Heever, S. C., and W. R. Cotton, 2004: The impact of hail size on simulated supercell storms. *J. Atmos. Sci.*, **61**, 1596–1609, doi: 10.1175/1520-0469(2004)061<1596:TIOHSO>2.0.CO;2.

VandenBerg, M. A., M. C. Coniglio, and A. J. Clark, 2014: Comparison of next-day convection-allowing forecasts of storm motion on 1- and 4-km grids. *Wea. Forecasting*, **29**, 878–893, doi: 10.1175/WAF-D-14-00011.1.

Wallace, J. M., 1975: Diurnal variations in precipitation and thunderstorm frequency over the conterminous United States. *Mon. Wea. Rev.*, **103**, 406–419, doi: 10.1175/1520-0493(1975)103<0406:DVIPAT>2.0.CO;2.

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.

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

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, doi: 10.1175/2007WAF2007005.1.

Weisman, M. L., and Coauthors, 2015: The Mesoscale Predictability Experiment (MPEX). *Bull. Amer. Meteor. Soc.*, **96**, 2127–2149, doi: 10.1175/BAMS-D-13-00281.1.

Weygandt, S. S., and N. L. Seaman, 1994: Quantification of predictive skill for mesoscale and synoptic-scale meteorological features as a function of horizontal grid resolution. *Mon. Wea. Rev*, **122,**57–71.

Wilks, D. S., 1995: *Statistical Methods in the Atmospheric Sciences*. Academic Press, 467 pp.

Wilks, D. S., 2001: A skill score based on economic value for probability forecasts. *Meteor. Appl*, **8,** 209–219.

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

Xue, M., and Coauthors, 2007: CAPS real-time storm-scale ensemble and high-resolution forecasts as part of the NOAA Hazardous Weather Testbed 2007 Spring Experiment. Preprints, *22nd Conf. on Weather Analysis and*

Forecasting/18th Conf. on Numerical Weather Prediction, Salt Lake City, UT, Amer. Meteor. Soc., 3B. [Available online at http://ams.confex.com/ams/pdfpapers/124587.pdf].

Xue, M., F. Kong, K. A. Brewster, K. W. Thomas, J. Gao, Y. Wang, and K. K. Droegemeier, 2013: Prediction of convective storms at convection-resolving 1 km resolution over continental United States with radar data assimilation: An example case of 26 May 2008 and precipitation forecasts from spring 2009. *Adv. Meteor.*, **2013**, 259052, doi: 10.1155/2013/259052.