DISSERTATION

A NEW POST-PROCESSING PARADIGM? IMPROVING HIGH-IMPACT WEATHER FORECASTS WITH MACHINE LEARNING

Submitted by

Gregory Reid Herman

Department of Atmospheric Science

In partial fulfillment of the requirements

For the Degree of Doctor of Philosophy

Colorado State University

Fort Collins, Colorado

Fall 2018

Doctoral Committee:

Advisor: Russ S. Schumacher

Elizabeth A. Barnes
Susan C. van den Heever
Daniel S. Cooley
Thomas M. Hamill

ABSTRACT

A NEW POST-PROCESSING PARADIGM? IMPROVING HIGH-IMPACT WEATHER FORECASTS WITH MACHINE LEARNING

High-impact weather comes in many different shapes, sizes, environments, and storm types, but all pose threats to human life, property, and the economy. Because of the significant societal hazards inflicted by these events, having skillful forecasts of the risks with sufficient lead time to make appropriate precautions is critical. In order to occur, these extreme events require a special conglomeration of unusual meteorological conditions. Consequently, effective forecasting of such events often requires different perspectives and tools than routine forecasts. A number of other factors make advance forecasts of rare, high-impact weather events particularly challenging, including the lack of sufficient resolution to adequately simulate the phenomena dynamically in a forecast model; model biases in representing storms, and which often become increasingly pronounced in extreme scenarios; and even difficulty in defining and verifying the high-impact event. This dissertation systematically addresses these recurring challenges for several types of high-impact weather: flash flooding and extreme rainfall, tornadoes, severe hail, and strong convective winds. For each listed phenomenon, research to more concretely define the current state of the science in analyzing, verifying, and forecasting the phenomenon. From there, in order to address the aforementioned persistent limitations with forecasting extreme weather events, machine learning-based post-processing models are developed to generate skillful, calibrated probabilistic forecasts for high-impact weather risk across the United States.

Flash flooding is a notoriously challenging forecast problem. But the challenge is rooted even more fundamentally with difficulties in assessing and verifying flash flooding from *observations* due to the complex combination of hydrometeorological factors affecting flash flood occurrence and intensity. The first study in this dissertation investigates the multi-faceted flash flood analysis problem from a simplified framework considering only quantitative precipitation estimates (QPEs) to assess flash flood risk. Many different QPE-to-flash flood potential frameworks and QPE sources are considered over a multi-year evaluation period and QPE exceedances are compared against flash flood observations and warnings. No conclusive “best” flash flood analysis framework is clearly identified, though

specific strengths and weaknesses of different approaches and QPE sources are identified in addition to regional differences in optimal correspondence with observations.

The next two-part study accompanies the flash flood analysis investigation by approaching *forecasting* challenges associated with extreme precipitation. In particular, more than a decade of forecasts from a convection-parameterized global ensemble, the National Oceanic and Atmospheric Administration's Second Generation Global Ensemble Forecast System Reforecast (GEFS/R) model, are used to develop machine learning (ML) models for probabilistic prediction of extreme rainfall across the conterminous United States (CONUS) at Days 2 and 3. Both random forests (RFs) and logistic regression models (LR) are developed, with separate models trained for each lead time and for eight different CONUS regions. Models use the spatiotemporal evolution of a host of different atmospheric fields as predictors in addition to select geographic and climatological predictors. The models are evaluated over four years of withheld forecasts. The models, and particularly the RFs, are found to compare very favorably with both raw GEFS/R ensemble forecasts and those from a superior global ensemble produced by the European Centre for Medium-Range Weather Forecasts (ECMWF) both in terms of forecast skill and reliability. The trained models are also inspected to discern what statistical findings are identified through ML. Many of the findings quantify anecdotal knowledge that is already recognized regarding the forecast problem, such as the relative skill of simulated precipitation in areas where extreme precipitation events are associated with large-scale processes well resolved by the GEFS/R compared with areas where extreme precipitation predominantly occurs in association with convection in the warm-season. But more subtle spatiotemporal biases are also diagnosed, including a northern displacement bias in the placement of convective systems and a southern displacement bias in placing landfalling atmospheric rivers.

The final extended study shifts weather phenomenon focus from extreme rainfall to severe weather: tornadoes, large hail, and severe convective winds. While both high-impact, the two classes of weather hazards share some commonalities and contrasts. While rainfall is directly forecast by dynamical weather models, most severe weather occurs on too small of spatial scales to be directly simulated by the same models. Consequently, unlike with extreme precipitation, when developing post-processed severe weather forecasts, there is no obvious benchmark for objectively determining whether and how much improvement the post-processing is yielding. A natural alternative, albeit much more stringent, benchmark is operational forecasts produced by human forecasters. Operational severe weather forecasts are

produced by the Storm Prediction Center (SPC), but there is limited published verification of their outlooks quantifying their probabilistic skill. In the first part of this study, an extended record SPC severe weather outlooks were evaluated to quantitatively assess the state of operational severe weather forecasting, including strengths and weaknesses. SPC convective outlooks were found to decrease in skill with increased forecast lead time, and were most skillful for severe winds, with the worst performance for tornado outlooks. Many seasonal and regional variations were also observed, with performance generally best in the North and East and worst in the South and especially West. The second part of the study follows similar methodology to the extreme precipitation models, developing RF-based probabilistic forecast models forced from the GEFS/R for Days 1–3 across CONUS, analogous to the format in which SPC produces its convective outlooks. RF properties are inspected to investigate the statistical relationships identified between GEFS/R fields and severe weather occurrence. Like with the extreme precipitation model, RF severe weather forecasts are generated and evaluated from several years of withheld validation cases. These forecasts are compared alongside SPC outlooks and also blended with them to produce a combined forecast. Overall, by statistically quantifying relationships between the synoptic-scale environment and severe weather in a manner consistent with the community's physical understanding of the forecast problems, the RF models are able to demonstrate skill over SPC outlooks at Days 2 and 3, and can be blended with SPC outlooks to enhance skill at Day 1.

Overall, multiple high-impact weather phenomena—extreme precipitation and severe weather—are investigated from verification, analysis, and forecasting standpoints. On verification and analysis, foundations have been laid both to improve existing operational products as well as better frame and contextualize future studies. ML post-processing models developed were highly successful in advancing forecast skill and reliability for these hazardous weather phenomena despite being developed from predictors of a coarse, dated dynamical model in the GEFS/R. The findings also suggest adaptability across a wide array of forecast problems, types of predictor inputs, and lead times, raising the possibility of broader applicability of these methods in operational numerical weather prediction.

ACKNOWLEDGMENTS

I would like to thank my entire graduate committee for their guidance and support throughout this long, arduous journey, and for their many helpful and constructive questions and suggestions offered along the way. I also wish to thank the Schumacher research group past and present—Sam Childs, Peter Goble Faith Groff, Stacey Hitchcock, Nathan Kelly, Erik Nielsen, John Peters, Bob Tournay, Vanessa Vincente, and Liu Zhang—for their suggestions and support over the years. Special thanks to Erik and Stacey; we've been through a lot together, from long nights in field projects, to similarly long nights working away at the department. They've been a great sounding board for everything from minutiae to large-scale research ideas. Discussions with them have been invaluable, both in improving the quality of my dissertation research, but also with providing much needed support outside of my research. Particular special thanks to Erik for also helping with the production of several figures that appear in this dissertation. Many thanks to all my friends, but especially also to Marie McGraw and Chris Slocum, who patiently listened to many of my vents over the years and also provided helpful thoughts and advice. Many thanks also to my family, and especially my parents, for their continued support of all my endeavors and pursuits, and for providing much needed encouragement throughout my graduate studies.

Many thanks to the collaborators at the Weather Prediction Center (WPC) who have helped mold much of my dissertation research and identify areas where the operational applicability of the research could be enhanced. In no particular order, this include Mark Klein, Diana Stovern, Sarah Perfater, Ben Albright, Mike Erickson, Jim Nelson, and Dave Novak. I would also like to thank the many Flash Flood and Intense Rainfall Experiment participants with whom I've interacted at WPC over the past five years; I've had many fruitful discussions at these experiments that have, either directly or indirectly, led to improved research quality. The suggestions and encouragement received by operational forecasters in different positions and locations has also been highly motivating and led to various minor changes that have hopefully improved the operational applicability of this dissertation research. Overall, these interactions with the operational community have been invaluable.

Many thanks also to the many manuscript reviewers, mostly anonymous, and editors who have collectively offered many constructive comments and criticisms that have resulted in many changes to produce a more thorough, accurate, complete, and better presented research product. Thanks also

to all of those who helped generate or provide me with access to datasets used for my dissertation research. Particular thanks to Tom Hamill and Gary Bates for developing and maintaining the reforecast model used extensively in the dissertation, and Adam Clark for providing access to an additional model archive used for much of the later stages of my dissertation research. Many thanks also to others, such as David John Gagne, who have provided much-needed guidance in this relatively young area of research.

I would like to acknowledge my funding sources throughout my graduate studies: CSU PRSE Fellowship; NOAA Award NA16OAR4590238; NSF Grant ACI-1450089; NOAA Award NA14OAR4320125, Amendment 26; UCAR/COMET award Z16-23465; NASA Grant NNX15AD11G; the Professor Susan C. van den Heever Monfort Professorship 53633; and NSF grant AGS-1409686. I also wish to thank the National Science Foundation, and the National Center for Atmospheric Research Computational and Information Systems Laboratory for providing computational resources used extensively for this research. Thanks also to CSU Atmospheric Science IT, Ammon Redman and Matt Bishop, for their assistance with adapting computational resources for the unique needs of my research and dealing with my numerous technical hiccups along the way.

Bridge has been a saving grace for me outside of my academic life throughout my graduate studies. I have had the great privilege of representing the USA three times in bridge world championships as a graduate student. I wish to thank all my bridge friends and mentors over the years for providing a helpful distraction. I want to especially thank Burke Snowden, Dawn Foltz, Michael Rosenberg, Debbie Rosenberg, and Kate McCallum for their support both at and away from the table.

And of course, last but not least, special thanks to Russ. I couldn't have done this without him, and I couldn't have asked for a better advisor.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

CHAPTER 1

INTRODUCTION

A serene sunrise. A dreary drizzle. An obfuscating fog. A violent cyclone. Weather has been a subject of awe, fascination, and bewilderment for millennia and beyond. Weather assumes many dispositions, is ever-evolving, and is never quite the same twice. It affects almost every aspect of society in some capacity, from the minutiae of daily life—what to wear, scheduling and activity planning, personal mood and productivity—to the economy, from agriculture to energy to health to transportation. However, it is not the quiescent conditions but rather the intense and unusual weather that brings the *greatest* societal impacts. High-impact weather can manifest in almost any location, and at any time of year, and in a multitude of different forms including: winter storms—snowstorms, ice storms, blizzards; flooding, including river floods and flash floods; severe weather—tornadoes, large hail, and strong winds; and windstorms, from tropical cyclones to downslope events. In 2017 alone, flash flooding was responsible for over 100 fatalities and almost $60B in the United States, while severe weather was responsible for 81 fatalities and over $2.5B in additional damages (NWS 2018).

Because of the immense societal impacts imposed by inclement weather, there have been efforts to forecast weather since before the times of Aristotle (Taub 2004). Early efforts were rather primitive, with forecasts made based on astronomy and pattern recognition from observing the skies. However, forecasting methods have become increasingly sophisticated with technological innovations. The invention of the telegraph in the mid-19th century allowed weather reports to be transmitted much more quickly than was done previously. These reports, which could now travel faster than the weather itself, allowed for weather forecasts to be legitimately based on real-time data for the first time, and to be made with some skill farther in advance. Realizing the potential utility in this information, concurrent with the expansion of the telegraph network came more concerted efforts to organize, decipher, and communicate weather reports and information (Willis and Hooke 2006). For example, beginning in 1871 Cleveland Abbe headed what is now the National Weather Service (NWS) (Willis and Hooke 2006). Alongside leading major efforts to collect and disseminate weather observations for improved forecasting, Abbe also argued for more effective numerical weather prediction (NWP) by coupling those observations with known governing physics of the atmosphere to derive predictions mathematically (Abbe 1901). Through all of these advancements, one thread—pattern recognition—has held constant throughout all weather forecasting efforts.

In tandem with these observation-based weather forecasting improvements came significant advancements to our physical understanding of the atmosphere and geophysical fluid dynamics more broadly. It was noted, both by Abbe and others, that knowing the laws governing the evolution of the atmosphere in conjunction with its present state would enable computation of future atmospheric states. Notably, Lewis Fry Richardson made among the first documented attempts at weather prediction from direct computation. During World War I, discretizing in space and time, Richardson attempted to use governing properties of geophysical fluid dynamics to predict pressure tendency six hours into the future. With significant manual labor, he eventually completed this forecast, predicting a 145 hPa pressure rise at his forecast location (Lynch 2008). Despite what might perhaps be characterized as a wildly unsuccessful attempt at atmospheric prediction, Richardson, Abbe, and others set the framework and foundation for a new paradigm for forecasting weather, now termed numerical weather prediction (NWP).

With innovations in computing technology over the subsequent decades came efforts to develop explicit models to simulate the atmosphere. Since the mid-20th century when the first automated numerical simulation of the atmosphere was successfully completed (Lynch 2008), there have been rapid, steady increases in computer power, and alongside these computing advances, increasingly sophisticated, skillful, and detailed dynamical weather models have been developed. More of the physics has been represented, model resolution has increased, and numerics have been improved to use resources more effectively and minimize numerical error. However, even with ever increasing sophistication, these models will never consistently produce close to perfect forecasts.

There are a multitude of sources of uncertainty in a dynamical model simulation and in turn result in forecast error. Analysis error—having an incorrect or incomplete representation of the *current* atmospheric state—is one major source of uncertainty. Given the infinitesimal scale on which the world and thus atmosphere operate, it is a practical impossibility to devise a perfect representation of the system state at any time past or present. Observations are not collected on this scale outside a laboratory setting, and all observations on any scale have some degree of error and uncertainty. Even with perfectly sampled representations of every particle, current dynamical models do not simulate on this scale; instead, simplifications are made to allow for efficient computation. Consequently, additional analysis error is inevitable in the process of translating real-world observations into the corresponding representation in a model. Initial conditions are one major source of uncertainty, but boundary conditions can yield erred representations of the atmospheric state as well. In limited area models,

the error associated with lateral boundary conditions can introduce error to the model solution even in the absence of analysis error. Additionally, in *all* models, top and bottom boundaries exist, and the representation of these boundaries can introduce additional error. On the bottom level, resolution of topography, soil type and properties in a discretized model all represent possible sources of uncertainty. Similarly, improper representation of the upper boundary can introduce additional uncertainty (e.g. Kalnay 2003).

Even in the absence of any analysis error, an imperfect model will still produce small errors in the model's representation of the future atmosphere, and continued non-linear error growth due to chaos will continue to increase the departure of the model solution from reality (e.g. Lorenz 1963). Models are imperfect in a number of ways, starting with model numerics. Floating point operations on modern computers have finite precision, resulting in truncation error both from a storage of numerical values and on operations between numbers. This exact source of error first led to the discovery of non-linear error growth in modeling atmospheric flows (Lorenz 1963). Further, many equations governing the atmosphere involve operations in continuous space, such as integrals and derivatives, which require simplifications to compute numerically in discrete space (Durran 2010). Relatedly, the resolution of numerical models also prevents accurate representation and simulation of many small-scale physical processes. Lack of proper and complete physical understanding of all atmospheric processes further compounds model error and uncertainty (Kalnay 2003). For all these reasons, even as they continue to improve, numerical weather models have always had and will continue to have error modes, biases, and other deficiencies.

Nevertheless, there is skillful predictive information from these dynamical models. Broadly, the objective in weather forecasting is and has always been to use all information available—climatologies, historical records, current observations, and model forecasts—to make the "best" weather forecast possible. Humans have been doing this since long before NWP, whether they conceptualize their actions this way or not. This process of digesting all of these pieces of information and producing a final forecast, referred to as post-processing[1], is in many senses *the* primary role of the human forecaster. However, humans are inherently limited in the rate in which they can ingest and assimilate data, and in the number of computations able to be performed within a prescribed time. In a process known as statistical post-processing (SPP), modern computers offer the potential to assist with the post-processing

[1]The process of assimilating current observations and running a dynamical model being the 'initial' processing.

process by ingesting more data and making more quantitative corrections than is possible by a human forecaster in the same amount of time.

There is a long history of successful application of SPP to dynamical model output (e.g. Klein et al. 1959; Glahn and Lowry 1972) and using statistics to forecast based on observations (e.g. Hope and Neumann 1970; Neumann 1972; Jarvinen and Neumann 1979). Model Output Statistics (MOS; e.g. Glahn and Lowry 1972), is a simple, effective multivariate linear regression technique relating a set of dynamical model predictors to sensible weather predictands such as minimum and maximum temperature, wind speeds, and precipitation probability. This basic technique has long demonstrated skill over both the underlying models and even human forecasters (e.g. Jacks et al. 1990; Vislocky and Fritsch 1997; Hamill et al. 2004; Baars and Mass 2005), but is inherently limited by the linear assumptions underlying the method. SPP techniques have also been successfully applied to precipitation forecasts, from early linear approaches (e.g. Bermowitz 1975; Antolik 2000) to more contemporary techniques that can exploit more complex variable relationships, including neural networks (e.g. Hall et al. 1999), reforecast analogs (e.g. Hamill and Whitaker 2006; Hamill et al. 2015), logistic regression (e.g. Applequist et al. 2002), random forests (e.g. Gagne et al. 2014; Ahijevych et al. 2016), and other parametric techniques (e.g Scheuerer and Hamill 2015). For other meteorological applications, other machine learning algorithms, such as support vector machines (e.g Zeng and Qiao 2011; Herman and Schumacher 2016b) and boosting (e.g. Herman and Schumacher 2016b; Hong et al. 2016) have been applied. Related techniques have also been applied to forecasting related high-impact phenomena, such as severe hail (Brimelow et al. 2006; Gagne et al. 2015) and tornadoes (Alvarez 2014).

Most existing SPP methodology to date has focused on minimizing forecast error in a bulk sense. However, as noted above, societal impacts of weather are *not* evenly distributed among all days; instead, the highest impacts are felt from rare, intense events that by definition deviate significantly from the climatology of the location. Though governed by the same fundamental physics, these events are necessarily less tested and familiar, and small model errors under quiescent conditions are often exacerbated. Consequently, the tools used for skillful forecasting of routine conditions are not always effective for the rare, high-impact scenarios; instead, tailored tools designed to assess the situation from a different perspective are often required.

In developing these tools, this dissertation seeks to raise the bar for real-time high-impact weather forecasts, and in particular for extreme precipitation and severe weather, through a series of systematic investigations outlined as follows. For each broad class of weather phenomena, the dissertation begins

with an investigation to better identify and contextualize the state of the forecast problem. From there, follow-on research implements machine learning-based post-processing to develop probabilistic forecast models in an attempt to explore the statistics and dynamics of each forecast problem and produce more skillful and reliable forecast guidance compared with existing products. It is hoped that by assessing more data and imposing fewer assumptions on the relationships within these data, these new tools will add value over existing tools available in forecast operations. Chapter 2 provides a comprehensive and systematic investigation comparing different ways of objectively assessing flash flooding from the perspective of QPE exceedances. This study is intended to identify strengths and weaknesses of using different frameworks and data sources to forecast and analyze extreme rainfall and flash flooding. It also serves as a springboard to Chapter 3, in which the development and evaluation of a machine learning-based probabilistic extreme precipitation forecast model for Days 2 and 3 across CONUS is described. Chapter 4 continues this work with more detailed analysis of the machine learning models and what they reveal about the dynamics and statistics of the extreme precipitation forecast problem for different CONUS regions. Chapters 5 and 6 investigate the extrapolability of this methodology to other high-impact weather phenomena by applying similar tools and algorithms towards the probabilistic forecasting of severe weather at Days 1–3 across CONUS. Chapter 5 evaluates the current state of operational severe weather forecasts by verifying an extended record of probabilistic Day 1–3 convective outlooks issued by the Storm Prediction Center (SPC). Chapter 6 follows on the work of Chapters 3–5 by developing machine learning-based models for issuing probabilistic forecasts of tornadoes, severe hail, and severe convective winds across CONUS in an analogous way to SPC in their convective outlooks. ML-based outlooks are evaluated alongside and combined with existing SPC outlooks to directly assess the utility of the ML-guidance in the operational pipeline. Chapters 3, 4, and 6 all develop forecast models based on dynamical model predictors from a somewhat antiquated, low-resolution, convection-parameterized global model known as the Global Ensemble Forecast System Reforecast Version 2 (GEFS/R), described in much more detail in Chapter 3. Chapter 7 synthesizes the findings of the dissertation, offers some overarching conclusions, and suggests future research and development directions for statistical forecasting and numerical weather prediction more broadly.

Dissertation research is presented in a topically-consistent progression, and is not always chronologically consistent with the order in which research was conducted. As such, there may be instances where experiment setup and design choices may appear inferior to possible alternatives based on research findings presented earlier in the dissertation. Such decisions were often made before those

study outcomes had been realized. Chapters 2 (Herman and Schumacher 2018b), 3 (Herman and Schumacher 2018c), 4 (Herman and Schumacher 2018a), 5 (Herman et al. 2018), and 6 (Herman 2018) correspond to peer-reviewed publications in American Meteorological Society (AMS) journals *Weather and Forecasting*, *Monthly Weather Review*, and *Journal of Hydrometeorology*. The content of these chapters largely aligns with the material presented in the corresponding manuscripts. However, the formatting has been modified to produce a cohesive dissertation, and additional supplementary research has been provided in select places to provide research detail beyond that which is already provided in those manuscripts.

CHAPTER 2

FLASH FLOOD VERIFICATION: PONDERING PRECIPITATION PROXIES

2.1 INTRODUCTION

Flash flooding is both a highly complex and immensely important forecast problem, being one of the leading causes of weather-related fatalities over the past several decades in addition to causing billions of dollars in economic damages in the annual mean (e.g. NWS 2017c). Part of the complexity compared with other weather hazards derives from the addition of hydrologic considerations alongside the purely meteorological ones. Antecedent soil conditions and the current levels of rivers and streams have a considerable influence on the proportion of rainfall that becomes surface runoff (e.g. Wood 1976; Castillo et al. 2003; Brocca et al. 2008). Land type and land use can also play a critical role (e.g. Ogden et al. 2000; Hapuarachchi et al. 2011), spanning the gamut from extremely absorbent sands to pavement, which can effectively saturate with very little rainfall. Urban effects such as pavement curvature and storm drain networks can also affect whether a flash flood is observed (e.g. Smith et al. 2005; Meierdiercks et al. 2010; Wolff 2013). Particularly in areas of complex terrain, the hydrologic response may also be highly sensitive to the precise spatiotemporal distribution of the precipitation; slight spatial displacements or differences in storm intensity may change whether a flash flood is observed. (e.g. Yatheendradas et al. 2008; Versini et al. 2010). Beyond the challenges from the hydrologic perspective, meteorologically, a complex combination of ingredients must come together to generate and sustain rainfall rates sufficient to produce flash flooding (e.g. Doswell et al. 1996; Davis 2001; Schumacher 2017). Flash-flood producing precipitation, which predominantly originates from warm-season moist convection over most of the contiguous United States (CONUS; e.g. Schumacher and Johnson 2005, 2006; Stevenson and Schumacher 2014; Herman and Schumacher 2016a), is consequently one of the most challenging and poorly forecasted sensible weather elements in contemporary numerical weather prediction (NWP; e.g. Fritsch and Carbone 2004; Novak et al. 2014).

Further exacerbating the flash flood forecast problem is the considerable difficulty in verifying flash flood events (e.g. Welles et al. 2007; Gourley et al. 2012; Barthold et al. 2015), an essential component to forecasting any phenomenon. There is no observation source with sufficient accuracy and density to determine whether a flash flood has occurred at every location across CONUS (e.g. Gourley et al. 2012, 2013; Barthold et al. 2015). Stream gauge measurements are useful, but they inherently cannot capture urban and other types of flash floods and are much too sparse even on streams and rivers to provide

adequate spatial resolution (e.g. Gourley et al. 2013). Flash flood reports (FFRs) from human observations are subject to population bias, with report databases often missing transient floods in very rural areas or at night (e.g. Pielke et al. 2002), and also to varying reporting and report encoding practices in different regions of the United States (e.g. Ashley and Ashley 2008; Calianno et al. 2013). Flash flood warnings (FFWs) have similar inconsistencies associated with differing warning philosophies across weather forecast offices (WFOs; e.g. Barthold et al. 2015; Marjerison et al. 2016; Schroeder et al. 2016), different proclivities to warn rural areas (e.g. Marjerison et al. 2016), and the fact that they correspond to anticipated—rather than observed—impacts.

Nevertheless, because of the societal threat posed by excessive rainfall and flash flooding, there is immense utility in having accurate flood and flash flood analyses and forecasts. Given the sensitivities and complications associated with calculating the hydrologic response to precipitation and the importance and urgency of disseminating updated flash flood information, it is often attractive in operational flash flood analysis and very near-term forecasting to simplify the problem down to a matter of only QPE. In this simplified framework, the question becomes: is the precipitation a given location has received or is receiving over some duration, as estimated by the QPE, sufficient to induce a flash flood? This essentially amounts to a binary exceedance question of whether the QPE over time T is in excess of some unknown threshold Θ_T above which flash flooding will occur and below which it will not. Even in this simplified framework, there are many challenges, which can be classified into two broad areas: 1) the discrepancy between true precipitation and QPE, and 2) the determination of T and Θ_T. On the former class of complications, current QPEs struggle with accurately quantifying extreme precipitation amounts (e.g. AghaKouchak et al. 2011; Hou et al. 2014; Zhang et al. 2016). Gauge observations have insufficient spatial resolution and density, while radar observation accuracy suffers from coarse and range-dependent vertical resolution, as well as having only indirect measurements of precipitation rate. Resultantly, QPE products are inherently too coarse to adequately capture local maxima corresponding to flash flooding. Even the highest-resolution products have substantial deficiencies (e.g. Nelson et al. 2016), which are also examined in this study.

Optimal threshold and interval determination is a complex, multi-layered challenge as well. One approach that attempts to do just this is the Flash Flood Guidance (FFG) product issued routinely by NWS River Forecast Centers (RFCs; Sweeney 1992). Based on the antecedent conditions and characteristics of the basin, dynamic estimates of Θ_T are issued on a subdaily basis for T = 1, 3, and 6 hours. However, these are not a panacea; because CONUS is so hydrometeorologically diverse and there is no

agreed single best methodology to compute these thresholds, different RFCs apply different methodologies to calculate FFG thresholds (e.g. Sweeney 1992; Ntelekos et al. 2006; Schmidt et al. 2007; Villarini et al. 2010; Clark et al. 2014), which can often produce highly different estimates and large nonphysical discontinuities across RFC boundaries (e.g. Clark et al. 2014; Barthold et al. 2015). Other approaches simplify the Θ_T estimation question and avoid these nonphysical political discrepancies by considering QPE exceedances of static thresholds, themselves derived across CONUS in a consistent manner. In particular, a fixed threshold (e.g. 2 in. hr.$^{-1}$) can be used as a proxy for flash flooding, as has been used in numerous previous studies (e.g. Brooks and Stensrud 2000; Hitchens et al. 2013; Novak et al. 2014). Exceedances of thresholds defined relative to the local precipitation climatology, such as average recurrence intervals (ARIs) can serve as Θ_T estimates as well. An ARI defines a fixed frequency relative to the hydrometeorological climatology of the region; in particular, it corresponds to the expected duration, given the local climatology, between exceedances of a given threshold. For example, the 1-year ARI for 24-hour precipitation accumulations describes the accumulation amount for which one would expect the mean duration between exceedances of said amount to be one year. Past research has shown that a fixed-frequency ARI-based framework can have better correspondence with heavy precipitation impacts than the use of any fixed threshold across the hydrometeorologically diverse regions of CONUS (e.g. Reed et al. 2007).

There are two primary objectives for this chapter. First, it seeks to evaluate the characteristics, deficiencies, and differences for existing QPE products and other tools and frameworks used in flash flood forecasting and analysis. Second, comparative evaluation of correspondence between QPE threshold exceedances and flash flood observations is performed to ascertain the merits of different QPE sources and the most effective ways to use QPE information for flash flood analysis and forecasting on both regional and national scales. Improved understanding of these properties can lead to more effective use of existing information in the short term, and identify revisions that may be adopted to existing products and algorithms to remove these undesirable properties in the longer term, resulting in more useful products for flash flood forecasting and analysis across a range of time and spatial scales. This study investigates issues surrounding these two important classes of challenges by first examining the climatological characteristics of heavy precipitation in several popular QPE sources. The issue of threshold quantification and application in flash flood analysis and forecasting is investigated with extensive comparison between different QPE threshold exceedances and flash flood observations. Chapter 2.2 describes the numerous datasets used in the study, and Chapter 2.3 describes the analysis methods

employed therefrom. Chapter 2.4 presents characteristics of the various QPE threshold exceedances and other sources employed in this study, and Chapter 2.5 assesses the correspondence between QPE exceedances and flash flood observations on regional and national scales. Chapter 2.6 summarizes the findings, describes the most important implications, and provides suggestions for future work.

2.2 Datasets

Flash flood reports (FFRs) in this study come from NWS local storm reports (LSRs) so encoded as flash floods. Archived LSRs are obtained from Iowa State University's Iowa Environmental Mesonet (IEM) Geographic Information System (GIS) archive (available online at https://mesonet.agron.iastate.edu/request/gis/). NWS flash flood warnings (FFWs) were also obtained from the IEM GIS archive. FFWs have been storm-based rather than county-based since 2008 (e.g. Waters et al. 2005; Ferree et al. 2006). Both warning type and report encoding are conducted for a given county warning area (CWA) by a governing WFO. Alternative report encoding options include "flood" and "heavy rain", while alternate weather warning and advisory options include flood warnings, flood advisories, and areal and small stream flood advisories. Practices on warning type, report encoding, and proclivity to issue warnings at all vary based on local WFO philosophy and practices (e.g. Barthold et al. 2015; Nielsen et al. 2015). Both FFRs and FFWs are available with temporal resolution to the minute.

There are several different gridded QPE sources currently in use in operational analysis and forecasting. Three leading sources are the National Centers for Environmental Prediction (NCEP) Stage IV Precipitation Analysis product (ST4; Lin and Mitchell 2005), the Climatology-Calibrated Precipitation Analysis (CCPA; Hou et al. 2014), and the Multi-Radar Multi-Sensor QPE product (MRMS; Zhang et al. 2016). ST4 provides QPEs across CONUS on a ~4 km grid for 1-, 6-, and 24-hour accumulations centered about 1200 UTC–1200 UTC meteorological days. It uses both rain gauge observations and radar-derived rainfall estimates to generate an analysis, and is further quality controlled via RFCs, particularly for 6- and 24-hour QPEs, to remove stray radar artifacts and other spurious anomalies (Lin and Mitchell 2005). ST4 products are generated by each RFC, and each center applies somewhat different treatments in generating the products. Most importantly, 1-hr QPE is not in general provided by the Northwest RFC, and has not been routinely generated by the California Nevada RFC since early 2016. When provided, 1-hr QPEs in this region are generally a simple disaggregation of 6-hr QPE into 1-hr

intervals. CCPA is derived from two QPE sources, the ST4 QPE and Climate Prediction Center's unified global daily gauge analysis. In particular, due to more rigorous and uniform quality control, the CPC-based QPE product is thought to have more accurate estimates than ST4, but has lower spatial and temporal resolution, at 1/8° and 24-hours, respectively. A linear regression technique is applied to upscaled and aggregated 6-hour ST4 QPE to correct its distribution towards the more robust CPC-based estimates, and then downscaled back to the native resolution to derive more accurate estimates while maintaining the spatiotemporal resolution of ST4. However, due to the limitations of linear regression, extremes not in the original ST4 cannot be introduced in the calibration process employed in CCPA, and extremes in ST4 are inherently regressed to some extent towards more typical values in the local precipitation climatology (Hou et al. 2014). Both ST4 and CCPA have up to a full day of latency in generation and publication of the QPE products. Lastly, MRMS, which became operational in September 2014, employs approximately 180 operational radars to create CONUS-wide radar mosaics every two minutes on a 1-km grid. In conjunction with gauge, satellite, and other environmental data, these radar mosaics are used to create CONUS-wide QPE at very high spatial and temporal resolution (Zhang et al. 2016). An initial radar only product is produced with 2 minute latency. The MRMS QPE used in this study has an additional gauge correction step, whereby gauges are ingested and undergo quality control, after which they are compared against the collocated radar-only estimates—incorporating aspects such as gauge network density and distance from estimate location—to develop a bias grid that is subtracted from the radar-only estimates. This gauge-corrected MRMS QPE has approximately 1 hour of latency, still much less than ST4 and CCPA (Zhang et al. 2016), leading to more operational applicability in operational forecast settings. In this study, we compare how the high precipitation tail of each of these QPE sources compare with each other and with FFRs.

In addition to each of these QPE sources, two other factors are considered as well: accumulation interval (AI) length, and threshold source. Regarding AIs, threshold exceedances for 1-hour, 3-hour, 6-hour, and 24-hour QPEs are considered as flash flood proxies. All of these AIs are considered for ST4 and MRMS QPEs; only 6- and 24-hour QPEs are available from CCPA. Three different sources for flash flood thresholds are considered as well: 1) a fixed threshold (FT) across CONUS, 2) exceedances of ARI thresholds, and 3) exceedances of FFG. Each of these methods is explained further below; FT and ARI exceedances are available for all AIs, while FFG exceedances are available for 1-, 3-, and 6-hour accumulations. A full summary of the threshold exceedance comparisons made for each QPE source is provided in Table 2.1 for reference.

TABLE 2.1. Threshold sources examined as a function of AI and QPE source, using the symbology of the chapter text.

	1-Hour	3-Hour	6-Hour	24-hour
MRMS	FT,ARI,FFG	FT,ARI,FFG	FT,ARI,FFG	FT,ARI
ST4	FT,ARI,FFG	FT,ARI,FFG	FT,ARI,FFG	FT,ARI
CCPA			FT,ARI	FT,ARI

TABLE 2.2. FT thresholds examined as a function of AI. 'X' indicates that the given threshold, AI combination is examined.

	1"	1.5"	2"	2.5"	3"	3.5"	4"	5"	6"
1-Hour	X	X	X	X	X	X	X		
3-Hour		X	X	X	X	X	X	X	
6-Hour		X	X	X	X	X	X	X	
24-Hour			X	X	X	X	X	X	X

FT grids are extraordinarily simple, as they are constant across CONUS. A variety of different thresholds are considered to assess the relationship between precipitation severity and flash flooding. Applying more stringent thresholds will result in fewer false alarms but more misses, while more lenient thresholds will induce the opposite result; it is expected that an optimal balance exists between these extremes. Of course, precipitation sufficient to produce flash flooding when falling within an hour likely will not produce flash flooding when distributed across a longer period. Therefore, the exact thresholds considered must necessarily depend on the AI; the full set of thresholds evaluated are indicated in Table 2.2.

The ARI thresholds are generated using very similar methodology to Herman and Schumacher (2016a), where CONUS-wide thresholds are produced by stitching thresholds from several sources. NOAA's Atlas 14 thresholds (Bonnin et al. 2004, 2006; Perica et al. 2011, 2013, 2015), an update from older work and currently under development, are used wherever they were available at the time this research began. For five northwestern states—Washington, Oregon, Idaho, Montana, and Wyoming—updated thresholds are not available, and derived Atlas 2 threshold estimates are used instead (Miller et al. 1973; Herman and Schumacher 2016a). In Texas, which currently has Atlas 14 threshold estimate updates in progress but no finalized thresholds available, Technical Paper 40 (TP-40; Hershfield 1961)

estimates are used. Everywhere else uses the Atlas 14 ARI threshold estimates. All of these threshold estimates are based on many decades of gauge data based on the availability and density of historical data in the region. While sophisticated spatial statistics are applied to derive the estimates and downscale to ungauged locations, particularly in the case of Atlas 14 (e.g. Bonnin et al. 2004; Perica et al. 2011), it is possible that undersampling from use of exclusively gauges can result in uncertain or erroneous estimates, particularly in historically rural areas, in areas of complex terrain, and areas without updated thresholds. Threshold uncertainty is quantified in Atlas 14, and increases with increasing ARI; this study uses only the best estimate values provided for the 1-, 2-, 5-, 10-, 25-, 50-, and 100-year ARI thresholds for each different AI. Atlas 14 provides estimates for each of these AIs. TP-40 provides estimates for each of these ARIs, but only for 6- and 24-hour AIs. Furthermore, NOAA Atlas 2 has available in digitized form only 6- and 24-hour ARI thresholds for the 2- and 100-year ARIs. Herman and Schumacher (2016a) derived thresholds for those AIs for the other ARIs. For Texas, Washington, Oregon, Idaho, Montana, and Wyoming, 1-hour and 3-hour threshold estimates are thus not natively available, and had to be derived.

In NOAA Atlas 14, a generalized extreme value distribution is fit to an annual maximum series for each duration independently; the results are related to the extent that the underlying data are the same (e.g. a 3-hour accumulation is comprised of 1-hour accumulations), but ARI thresholds for different AIs are not directly computed in tandem (e.g. Bonnin et al. 2004). Here however, where Atlas 14 estimates have not yet been officially generated and the original underlying data had insufficient temporal resolution, a relationship must be derived between the threshold estimates that are available and the desired, unknown thresholds at shorter durations. Accordingly, an analytic equation is derived to relate the 6-hour and 24-hour thresholds for a given ARI to 3- and 1-hour estimates. The formula is composed of two components. One term is designed to exactly obey desired mathematical properties; the second, tunable term alters the formula to match the known relationships—where Atlas 14 estimates are available for all AIs—as well as possible whilst obeying the mathematical properties to the extent possible. Desired mathematical properties include: 1) threshold estimates go to zero in the limit as AI goes to zero; 2) threshold estimates go to infinity in the limit as AI goes to infinity; 3) the formula is valid for any positive AI; 4) the rate of change of threshold magnitude with increasing AI decreases with increasing AI; 5) when the ratio of known threshold estimates for two different AIs is exactly equal the ratio of those AIs, the threshold estimate for an AI with an equal ratio with one of the AIs with a known threshold should exactly preserve the same ratio with the threshold estimate corresponding to that AI

(e.g. if a 6-hour estimate is 10 and 24-hour estimate is 20, a 1.5-hour estimate should be 5); 6) using the formula to derive thresholds for one of the two known AIs being used returns those same threshold estimates; 7) the formula is reversible: it can be used to derive a third "known" estimate, and the use of any two can then be used to exactly recover the third; and 8) an arbitrary number of intermediate threshold estimates can be derived without altering the estimate for a given AI (e.g. deriving a 3-hour estimate from 6-hour and 24-hour estimates, and then using 3- and 6-hour estimates to derive a 1-hour estimate will produce the same result as deriving 1-hour estimates from the 6- and 24-hour values). It can be easily shown that for shorter AI S and longer AI L with known threshold estimates Θ_S and Θ_L, an equation for deriving an unknown estimate Θ_N for AI N that satisfies all of these properties is:

$$\Theta_N = \Theta_S \left(\left(\frac{\Theta_L}{\Theta_S} \right)^{\log_{\frac{L}{S}}\left(\frac{S}{N}\right)} \right)^{-1} \tag{2.1}$$

The tunable term is constructed as:

$$\left(\log_{\frac{L}{S}} \left(\frac{\sqrt{SN}}{\frac{S\sqrt{SL}}{L}} \right) - 1 \right) \log_{\frac{L}{S}} \left(\frac{S}{N} \right) \alpha \tag{2.2}$$

for tunable parameter α. That term is further decomposed into:

$$\alpha = \beta \left(\frac{\Theta_S}{\overline{\Theta_S}} \right)^{\kappa_S} \left(\frac{\Theta_L}{\overline{\Theta_L}} \right)^{\kappa_L} \tag{2.3}$$

Tuning in areas where Atlas 14 estimates are available and thus "truth" is known yielded:

$$\beta = \frac{4}{3}\left(1 - \log_{10} ARI\right); \kappa_S = 1.7; \kappa_L = 0.6 \tag{2.4}$$

The above expressions are used to derive estimates in locations where native estimates are currently unavailable, and stitched with the Atlas 14 estimates to produce the CONUS-wide threshold grids of Figure 2.1 (see also Fig. S1 in Herman and Schumacher (2018b)). The grids present a stark contrast to the spatially uniform FT grids, with values spanning near an order of magnitude across CONUS. Climatologically drier areas such as the Intermountain West have lower thresholds, while wetter regions such as the Gulf Coast have higher thresholds. As expected, thresholds are lowest for the smallest AIs and ARIs and become larger with increasing duration and rarity. However, the extent of change as a function of AI in particular is not spatially uniform, and instead reflects the climatological characteristics of the types of precipitation systems associated with locally extreme precipitation in the given

region. This is seen to some extent when comparing the left and right columns of Figure 2.1, but especially when comparing those of Figure 1 with those of S1. For example, while thresholds for the 24-hour AI are comparable between the Gulf Coast and Pacific coastal mountains, the former region has much higher thresholds at the 1-hour AI (e.g. ~125 mm vs. ~45 mm for the 100-year ARI in Fig. 2.1m). Further, locations much farther north and more distant from an ocean, such as over Iowa and Minnesota, have appreciably lower thresholds than the Pacific mountains for 24-hour accumulations, but are also much higher for 1-hour AIs. Over the Pacific Coast, extreme precipitation events are typically associated with long duration atmospheric river events, which can produce moderate to heavy rain for an extended duration (e.g. Rutz et al. 2014; Herman and Schumacher 2016a). In contrast, over the Southeast and Great Plains, most extreme precipitation is associated with smaller-scale convective systems, which can produce higher rain rates than their West Coast counterparts, but last for a shorter duration at any given point (e.g. Herman and Schumacher 2016a). The Gulf Coast region sustains high thresholds across the spectrum of AIs; at shorter AIs, this is predominantly associated with small-scale convective storms, while high thresholds at longer durations are predominantly supported by tropical cyclone rainfall (e.g. Kunkel et al. 2012).

FFG estimates the average precipitation amount required over an area in a prescribed amount of time to initiate flooding of small streams in that area (Sweeney 1992). FFG is calculated individually by each RFC, with each office maintaining independent code and algorithms for FFG calculation (Sweeney 1992; Barthold et al. 2015). RFC-generated FFG may be assembled to form a national grid covering all of CONUS, with the exception of Washington and Oregon west of the Cascades; the Northwest RFC does not calculate FFG for this small region of CONUS. FFG values are based on threshold-runoff calculations, which specify the minimum amount of runoff (not precipitation) into a stream or basin over a prescribed 1-, 3-, or 6-hour duration necessary to produce bank-full conditions (Sweeney 1992; Ntelekos et al. 2006). This is done offline for thousands of small basins and is independent of present conditions. These basin-specific threshold-runoff calculations are interpolated onto a ~4 km grid to provide a unified analysis. A hydrologic model, such as the Sacramento Soil Moisture Accounting Model (e.g. Carpenter et al. 1999) or Antecedent Precipitation Index models (e.g. Brocca et al. 2008), are then used in conjunction with current conditions to relate rainfall amounts to runoff amounts. The minimum rainfall to yield a runoff in the hydrologic model in excess of the gridded threshold-runoff values then constitute the gridded FFG values (Ntelekos et al. 2006).

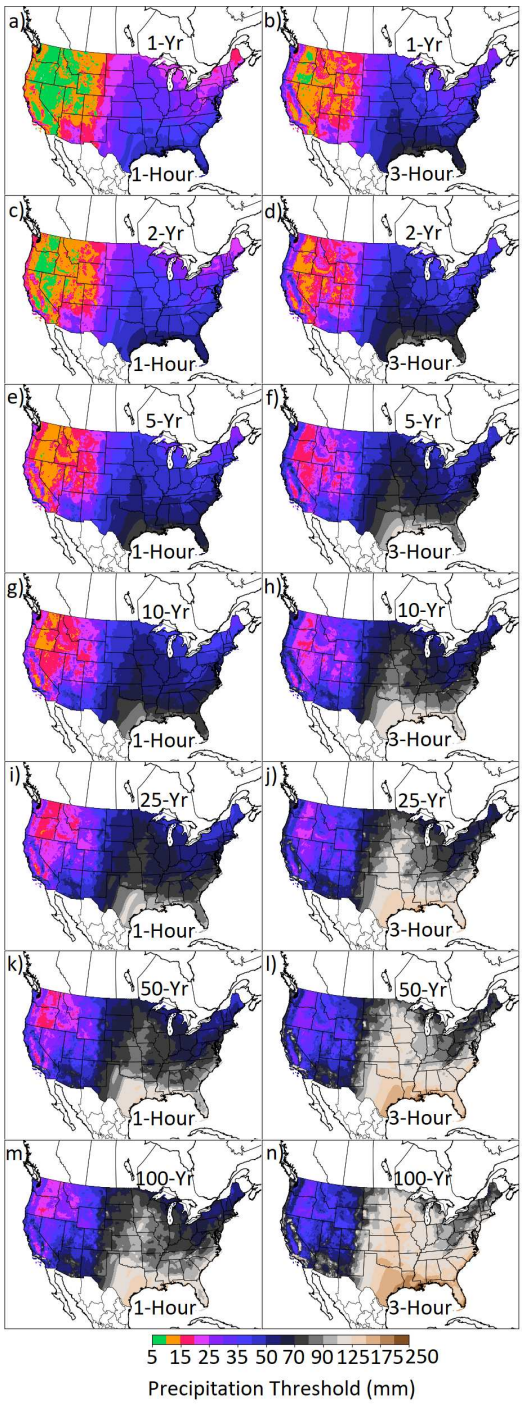


FIG. 2.1. ARI threshold estimates for 1- (left column) and 3-hour (right column) precipitation accumulations for 1-, 2-, 5-, 10-, 25-, 50-, and 100-year ARIs in panels (a)–(b), (c)–(d), (e)–(f), (g)–(h), (i)–(j), (k)–(l), and (m)–(n), respectively. Threshold estimates come primarily from NOAA Atlas 14, but are supplemented from other sources as described in the text.

Unlike ARI thresholds, FFG thresholds vary dynamically based on the antecedent conditions. While this makes it impossible to plot a single static plot depicting the FFG thresholds across the entire period

of record, the distribution of issued values can be considered. The median FFGs across the period of record (Fig. 2.2a–c) vary in a similar fashion to the tail of the precipitation climatologies as quantified by the ARI thresholds (Fig. 2.1), with very low values over the Intermountain West increasing progressively to very high values over the Gulf Coast and particularly Florida. However, while ARI thresholds reflect only the precipitation climatology and do not *directly* address the hydrologic component of flash flooding, FFG does account for these factors. This can result in large gradients in FFG climatologies in regions of rapid change in soil type or land use; one prominent example is in the Nebraska Sand Hills (e.g. Fig. 2.2a,b). Also evident is the large spatial discontinuities that occur even in the median across RFC boundaries. One glaring example in the 6-hour median FFG (Fig. 2.2c) is the border between the Northwest RFC and California-Nevada RFC near the southern borders of Oregon and Idaho. The same general findings exist on the high-risk tail of the FFG climatology, as evidenced by the tenth percentile FFGs (Fig. 2.2g–i). In general the difference between the fiftieth and tenth percentiles over the western RFC domains (cf. Fig. 2.2b,h) is small, while thresholds for the tenth percentile are appreciably—although not uniformly—lower across central and eastern CONUS. In particular, the Middle Atlantic RFC appears to be more responsive to antecedent conditions than its neighbors, resulting in locally lower thresholds in their domain and large spatial discontinuities in the tenth percentile FFGs at their RFC boundaries; this is especially pronounced for 1-hour FFG (Fig. 2.2g).

Both ARI and FFG exhibit strong and clearly apparent contrasts with FT methodology, but are quite different from each other as well. Median FFGs are, according to ARI thresholds (Figs. 2.2d–f), most commonly exceeded over the Great Plains and Mississippi Valley regions. There, ARI equivalents for median FFGs can be as low as 1-year in Iowa for 3-hour FFGs (Fig. 2.2e), and are between 2 and 5 years across most of the region. In contrast, median FFGs near and along the Atlantic Coast are generally appreciably higher, with values of 10–25 years. Higher still are typical thresholds in the West, with ARI equivalents mostly ranging from 25 to over 100 years. In the West, large differences in ARI equivalence are found depending on the AI used. In the northern Intermountain West, including Idaho, equivalent ARIs to the median 6-hour FFGs (Fig. 2.2f) are only 2–5 years, while being mostly 10–25 years in those same areas for 1-hour FFGs (Fig. 2.2d). The opposite, and even stronger, contrast is seen in the Arid Southwest, particularly Arizona. The ARI equivalent for the median 6-hour FFG (Fig. 2.2f) is at least 100 years, while it is 2–5 years over much of the state for 1-hour FFG (Fig. 2.2d), and is even as low as a 1-year ARI across the southeast portion of the state. These AI-dependent contrasts suggest, for example, that most floods in the Southwest are associated with short-lived rain events, and that

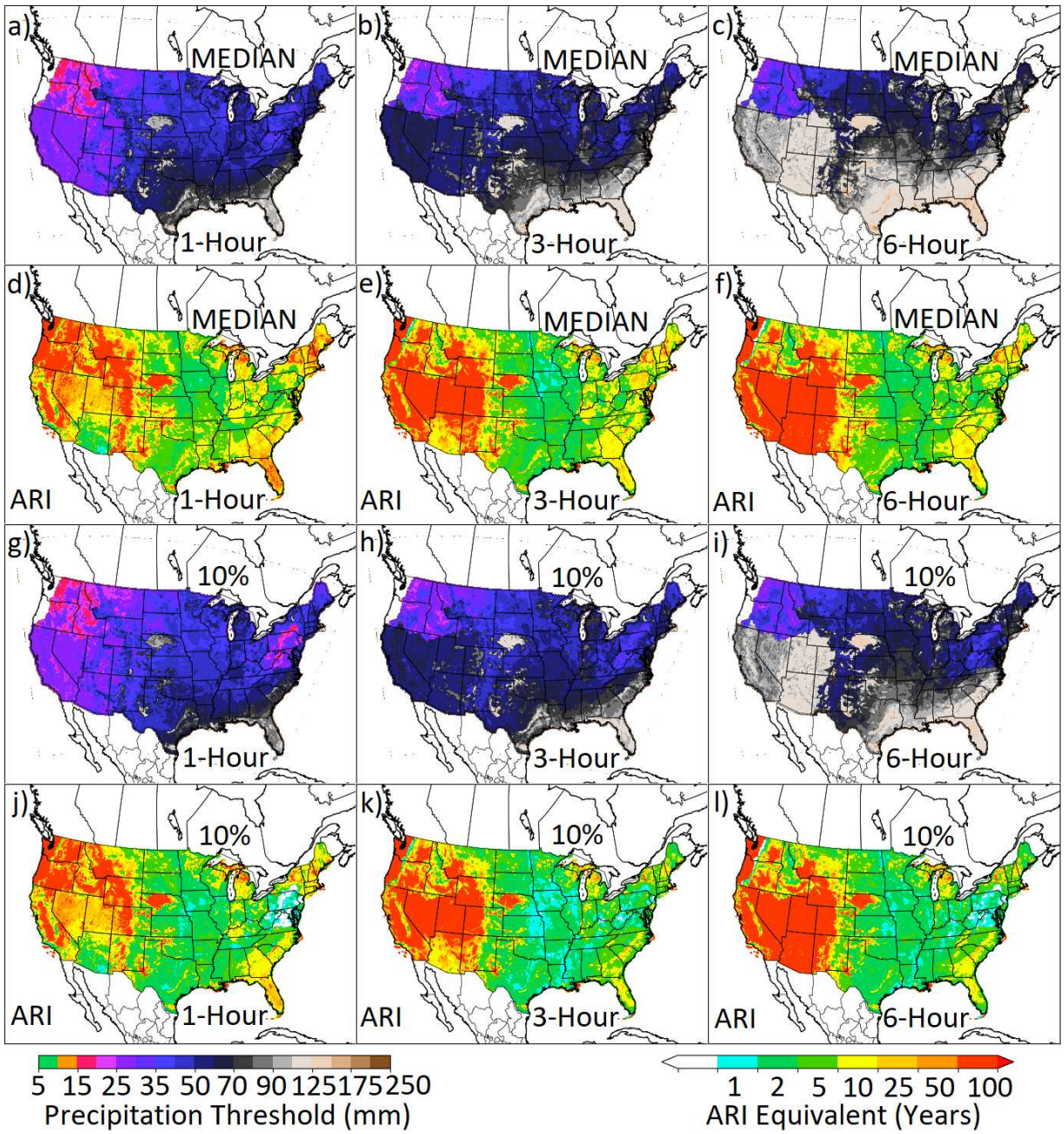


FIG. 2.2. Median (panels (a)–(f)) and tenth percentile (panels (g)–(l)) FFG estimates over the 2.5 year period of record. The left column (panels a,d,g,j) corresponds to 1-hour FFG values, center column (panels b,e,h,k) to 3-hour FFGs, and right column (panels c,f,i,l) to 6-hour FFGs. Panels (a)–(c) and (g)–(i) correspond to the actual threshold estimates, while (d)–(f) and (j)–(l) correspond to the equivalent ARIs to those thresholds for the particular grid point.

for 1- and 6-hour rain events of equal rarity in the Southwest, the 1-hour event rates to have greater hydrometeorological impact. All the same general findings are also found comparing the ARI framework with the tenth percentile FFG thresholds, just with lower ARI equivalent thresholds. The one very

prominent difference is again in the Middle Atlantic RFC; here, tenth percentile 1-hour FFGs are below the 1-year ARI across much of their domain (Fig. 2.2j), while the tenth percentile FFGs in neighboring areas largely correspond to 5–10 year ARIs.

2.3 Analysis Methodology

All grids are first regridded if necessary onto the ST4 Hydrologic Rainfall Analysis Project (HRAP; Fulton et al. 1998) ~4 km grid. ST4 and CCPA QPE is already provided on this grid; MRMS QPE is regridded onto this grid using a first-order conservative scheme (Ramshaw 1985). ARI and FFG thresholds are regridded bilinearly onto this grid. FFRs and FFWs are not gridded products at all, the former being points in space and the latter polygons in space. FFRs are remapped onto the HRAP grid using a 40km radius of influence, projecting a single report onto numerous points on the grid. Events are defined for 24-hour 1200–1200 UTC "meteorological days", similar to current operational practice at the Weather Prediction Center and Storm Prediction Center (e.g. Edwards et al. 2015; NWS 2017a, , and discussed more in Chapter 5). As such, despite both having 1-minute resolution, an FFR "event" for the purpose of this study is defined as one or more reports within 40 km of the point occurring anytime within the meteorological day. An FFW event is similarly defined as any FFW enclosing the given HRAP grid point valid at any time during the meteorological day.

Once this is performed and all fields are assembled on a uniform grid, slight additional quality control is performed following Herman and Schumacher (2016a) to remove QPEs that are clearly non-physical, and then binary comparison between QPE and selected thresholds is made. Comparisons are first made on the ST4 HRAP grid to generate binary exceedance grids. For sub-daily AIs, there are multiple grids with valid times falling within a given 1200–1200 UTC period. In these instances, the maximum of all grids with the same AI and corresponding meteorological day is taken to form a single daily exceedance grid for the series; all subsequent analysis uses these aggregated daily grids. In this way, daily grids for sub-daily AIs correspond to one or more of the given type of QPE exceedance occurring at that point during the meteorological day, regardless of the exact number of exceedances. Tolerance to small spatial displacements is provided by using a maximum nearest neighbor upscaling from the HRAP grid to a 0.5°×0.5° grid. For each day, all HRAP grid points are mapped to their nearest point on the 0.5° grid; at each grid point on this coarser grid, an event is recorded if any of their mapped HRAP points indicated an exceedance event for that meteorological day. QPEs are only considered for periods centered about the meteorological day, and are not considered for any interval spanning across

meteorological days (e.g. 24-hour 0000-0000 UTC accumulations). Based on data availability, a 2.5 year period of record spanning 2 January 2015–23 June 2017 is used for most verification in this study, with slightly truncated period beginning 19 March 2015 for MRMS QPE comparisons, again limited by data availability.

After exceedance grids have been computed, they are compared to assess how the characteristics vary as a function of threshold source, accumulation interval, threshold magnitude, and QPE source. Despite the aforementioned limitations of FFRs and FFWs, evaluation is made using each of these sources as a reference truth. Although these are not believed to completely embody "true" occurrences and non-occurrences of flash floods, it is performed in order to provide a common framework for comparison between different QPE exceedances. In this framework, the reference—either FFRs or FFWs—serves as a deterministic truth, and the QPE exceedances serve as deterministic predictions. The analysis framework employed here, namely deterministic binary predictions and binary observations, lends itself well to the use of contingency table statistics (Wilks 2011). Given the number of different thresholds, intervals, and sources considered, it is convenient to represent the comparison statistics succinctly in a single plot. One popular way to present the full dimensionality of the contingency table verification for many different forecast sets in a single plot is through the so-called performance diagram (PD; Roebber 2009). The PD succinctly places a forecast set in the context of these verification statistics on one plot, with success ratio (SR) increasing on the x-axis, probability of detection (POD) increasing on the y-axis, frequency bias (FB) increasing from 0 at the lower right corner to infinity at the upper left, and critical success index (CSI) increasing from 0 at the lower left corner to unity—a perfect score—at the upper right. In addition to PDs, spatial maps of CSI are assessed to provide context of where correspondence between the QPE exceedances and reference truth are better and worse across CONUS. Finally, a single geometric mean equitable threat score (ETS) is computed between the comparisons with the two reference datasets, and is so chosen over CSI to alleviate concerns that the latter exhibits with varying underlying event frequencies (Gandin and Murphy 1992; Marzban 1998) and produce a skill metric more independent of the event climatology (e.g. Jolliffe and Stephenson 2003). A geometric mean is chosen over a conventional one to more strongly penalize lack of correspondence with either observation set.

2.4 Results: Exceedance Climatologies

Examination of simple exceedance, report, or warning counts, as the case may be, over the period of record in Figure 2.3 illuminates several interesting contrasts between the datasets. A heat map of FFRs over the period (Fig. 2.3a) illustrate several of the aforementioned limitations of reports as representations of true flash floods. The population bias is clearly evident in some regions of the country. For example, in Texas, far more reports are observed in the Houston, Dallas, Austin, and San Antonio metro areas than in surrounding areas despite them having similar rainfall climatologies (e.g. Fig. 2.1). Some of this is likely a legitimate reflection of urban environments flooding more easily than rural ones due to land use and other factors, but extracting the component of legitimate spatial variation from a population-based reporting bias is challenging. Spatial variations attributable to human factors can be discerned as well, with discontinuities in report counts across CWA boundaries in places, particularly entering Florida and Michigan and to a lesser extent in Georgia and Alabama (Fig. 2.3a). These political effects from WFO tendencies are even more prominent in FFW issuances (Fig. 2.3b) in those same locations. Additionally, there are several local "hot spots", such as Las Vegas, wherein one WFO issues far more FFWs than its neighbors. Again, some of this is certainly meteorological due to different climatological flood susceptibility of neighboring CWAs and from limited sampling due to the finite period of record. For example, FFWs and especially FFRs (Fig. 2.3a) are more concentrated in the Southern Plains and Midwest, with far fewer events over the Northern Plains and farther west over the Rocky Mountains and Pacific Coast. These large-scale spatial variations accord with previous studies of flash flooding (e.g. Brooks and Stensrud 2000; Hitchens et al. 2013; Schumacher and Johnson 2006), and are likely legitimate rather than an artifact of the datasets. The magnitude of the differences suggest, however, that at least some contribution to these *local* spatial variations across CWA boundaries is political rather than purely hydrometeorological. FT exceedances, in contrast, do not exhibit any of these spatial discontinuities at political boundaries. For 1-hour accumulations (e.g. Fig. 2.3e), they instead exhibit a prominent, relatively smooth gradient from almost no exceedances of 1.5 in. $hr.^{-1}$ occuring over northwestern CONUS, to being extremely common over southeastern CONUS near the Gulf Coast. The spatial distribution of events over central and eastern CONUS remain similar with increasing AI (cf. Fig. 2.3e, 2.3h), but the number of exceedances along the Pacific Coast increases dramatically, with almost no exceedances for 1-hour accumulations (e.g. Fig. 2.3e), and as many exceedances as the Gulf Coast for 24-hour accumulations (e.g. Fig. 2.3h). This largely accords with the ARI thresholds, which

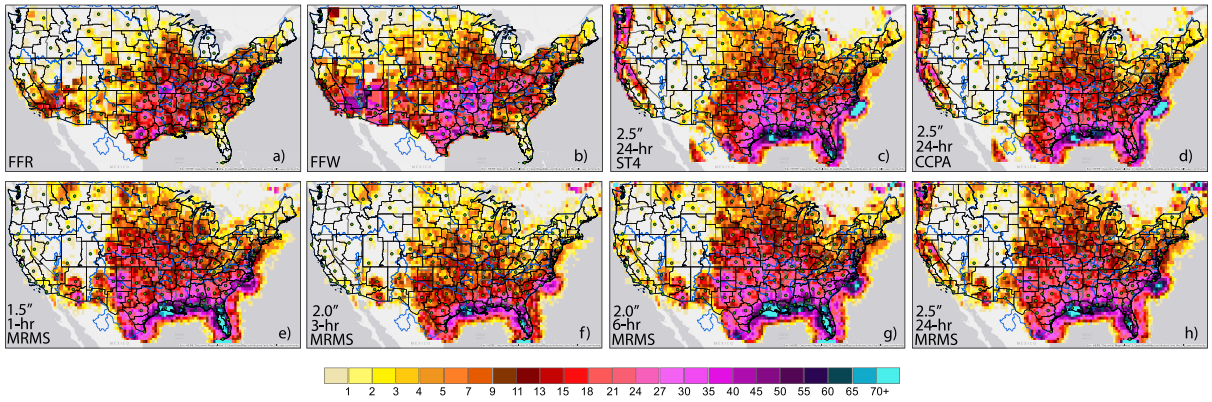


FIG. 2.3. Heat maps for FT exceedances, FFRs, and FFWs during the relevant period of record (see text). Panels (a) and (b) correspond respectively to FFRs and FFWs reported and issued during the period of record, gridded as described in the chapter text. Panels (e)–(h) depict MRMS QPE FT exceedances, where columns from left to right correspond to 1-, 3-, 6-, and 24-hour precipitation accumulations and 1.5, 2.0, 2.0, and 2.5 in. (38, 51, 51, 64 mm). Panels (c) and (d) are also for 24-hour 2.5 in. exceedances as in panel (h), but for ST4 and CCPA, respectively. Thick black outlines depict CWA boundaries; blue lines indicate RFC domain boundaries, and green circles indicate locations of NEXRAD radar sites.

are similar in these two regions for 24-hour thresholds (Herman and Schumacher 2016a) and much higher over the Gulf Coast for 1- and 3-hour accumulations (Fig. 2.1).

Politically-attributable exceedance count discontinuities are also evident in FFG exceedance heat maps (Fig. 2.4), in this case primarily across RFC boundaries. For 1-hour FFGs (Fig. 2.4a), this is most readily apparent with respect to the Middle Atlantic RFC; there are far more exceedances in their domain than either their Northeast or Southeast RFC neighbors. For 3- and 6-hour FFGs (Fig. 2.4e,f), a very large discontinuity is seen between the Colorado Basin RFC and its neighbors to the north and east, including the Northwest, Missouri Basin, and West Gulf RFCs, with almost no exceedances of 3- or 6-hour FFGs in the Colorado Basin domain but numerous exceedances immediately adjacent in other RFC domains. These discrepancies are consistent with the FFG threshold climatologies (Fig. 2.2). There are also far fewer total FFG exceedances across CONUS than exceedances of the FT thresholds presented in Figure 2.3, with a prominent exception of 1-hour MRMS QPE FFG exceedances in Arizona (Fig. 2.4a). There is also in general far less spatial gradient in exceedance counts of FFG compared with the FT exceedances. In places, such as the Southeast and particularly Florida, the anomalously low number of reports (Fig. 2.3i) and warnings (Fig. 2.3j) in the area compared with its surroundings is corroborated by a relatively low number of FFG exceedances (e.g. Fig. 2.4d), while in other areas, such as Michigan (e.g. Fig. 2.4c), it is not.

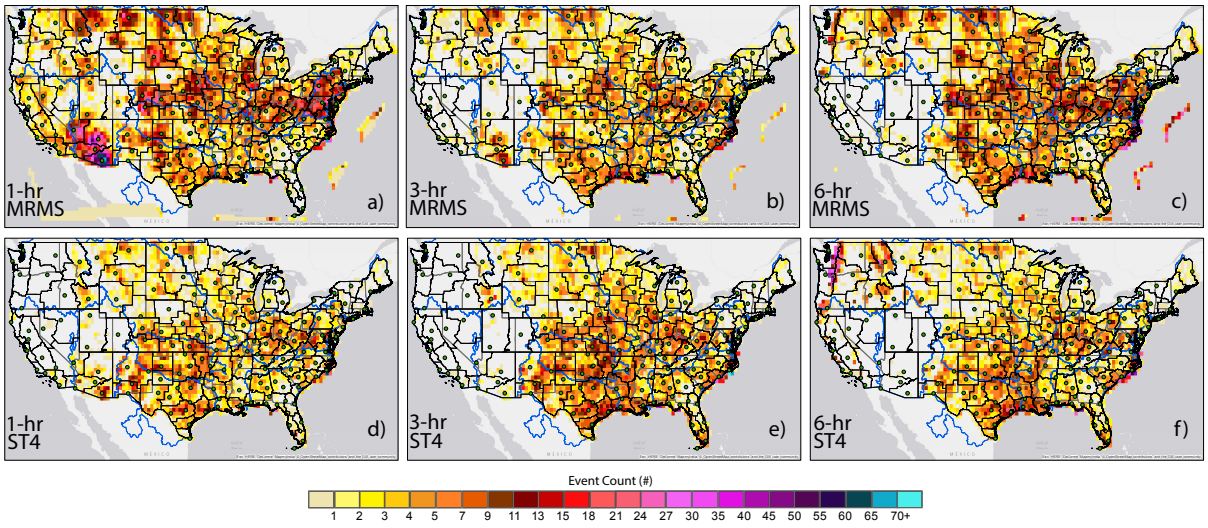


FIG. 2.4. Heat maps for FFG exceedances across CONUS during the relevant period of record (see text). Panels (a)–(c) correspond to exceedances of FFG based on MRMS, (d)–(f) to ST4 QPE exceedances for 1-hour, 3-hour, and 6-hour FFGs in panels (a) and (d), (b) and (e), and (c) and (f), respectively. Thick black outlines depict CWA boundaries; blue lines indicate RFC domain boundaries, and green circles indicate locations of NEXRAD radar sites.

Aside from sampling noise associated with using a finite and relatively short period of record, ARI threshold exceedance heat maps (Fig. 2.5) should definitionally be uniform across the entire spatial domain. Departures from uniformity must then be attributable to either 1) sampling noise, 2) inaccurate ARI threshold estimates, or 3) systematic error in the QPE source. Comparison across different QPE sources and threshold sources can help identify root causes. Some notable spatial variations can be seen—some are consistent across QPE sources, while others are particular to one. For example, MRMS QPE (Fig. 2.5a–b,2.5e–f) exhibits a glaring anomaly in exceedance counts in the West: there are far more ARI exceedances observed in the immediate vicinity of radar sites compared with their surroundings. While this is evident for all AIs and across different levels of severity, it is especially apparent for shorter intervals (e.g. Fig. 2.5a,b). This phenomenon, which is also seen in the FT (e.g. Fig. 2.3e) and FFG (e.g. Fig. 2.4a) exceedances, is considerably alleviated or even entirely eliminated to the east of the Rocky Mountains. ST4 QPE ARI exceedances (Fig. 2.5c–d,2.5g–h) exhibit two sharp local maxima far above any other location: one in Wyoming and the other in New Mexico. This is especially prominent for the 24-hour AI (Fig. 2.5h); it is seen to a lesser extent with MRMS QPE (Fig. 2.5f) as well. Interestingly, the discontinuity in FFG exceedance counts across the Colorado Basin RFC boundary is replicated in the ARI exceedances—especially prominent in ST4 but evident in all QPE sources. This suggests that the discontinuity may be largely attributable to artifacts of the native QPE rather than

politically-based discontinuities in FFG thresholds. ST4 exceedances at short 1- and 3-hour AIs (e.g. Fig. 2.5c,d) have a substantial reduction in ARI exceedances in the West due to the aforementioned limited productionof 1-hr ST4 QPE in the western RFCs. Lastly, for CCPA QPE (Fig. 2.5i–j), the maximum in Wyoming remains clearly apparent, but the maximum in New Mexico is greatly muted. The consistency of overexceedances in Wyoming suggests that the ARI threshold estimates, which are now several decades old, may be too low in this area. That the New Mexico maximum is largely removed with the bias correction applied by CCPA suggests that the New Mexico issue may be more attributable to deficiencies with ST4 and MRMS QPE in complex terrain, with small areas of large radar estimated values unable to be properly corrected due to insufficient gauge data in the region.

2.5 Results: Flash Flood Correspondence Skill

PDs for CONUS-wide verification for the complete set of QPE to observation reference comparison verification in Figure 2.6 illustrate that, for a given QPE source, AI, and threshold method, a curve sweeps from the top left corner of the PD to the bottom right corner with increasing threshold magnitude. The lowest thresholds jointly exhibit a high POD, high FB, and low SR, while high thresholds possess the opposite characteristics. The curve is not, however, parallel with the curved skill (CSI) lines in the PD, and instead attains a maximum CSI for some middle threshold magnitude. Surpisingingly, out of all the different QPE comparisons against FFRs (Fig. 2.6), maximum CSI values are obtained for 2.0 in. (50.8 mm) 6 hr^{-1}, 2.5 in. (63.5 mm) day^{-1}, and 3.0 in. (76.2 mm) day^{-1}—all FT threshold sources (warm-colored interior symbols). The maximum CSI obtained using ST4 (blue outlined symbols) and CCPA-based (green outlined symbols) QPE comparisons exceed that reached using MRMS QPE exceedances (red outlined symbols), though the CSI differences are small, with maximum values all around 0.23. Across the range of threshold comparisons considered, the highest FBs extend over 5, and the lowest are well under 0.1; SRs range from 0.1–0.65, and PODs range from almost 0 to near 0.85. FBs for FFG exceedances are all near or below unity, consistent with the findings of Clark et al. (2014) which also found raw FFG to be too stringent and found better correspondence using fractional FFG. For a given threshold method, magnitude, and AI, there are similar common differences between QPE source comparisons. CCPA QPE consistently exhibits a lower FB, higher SR, and lower POD than ST4 or MRMS; MRMS usually exhibits the highest FB, highest POD, and lowest SR of the three. When compared against FFWs (Fig. 2.7), the general scatter of QPE exceedance verifications in the PD phase space remain the same, but CSIs are generally somewhat higher and differences in the specifics emerge.

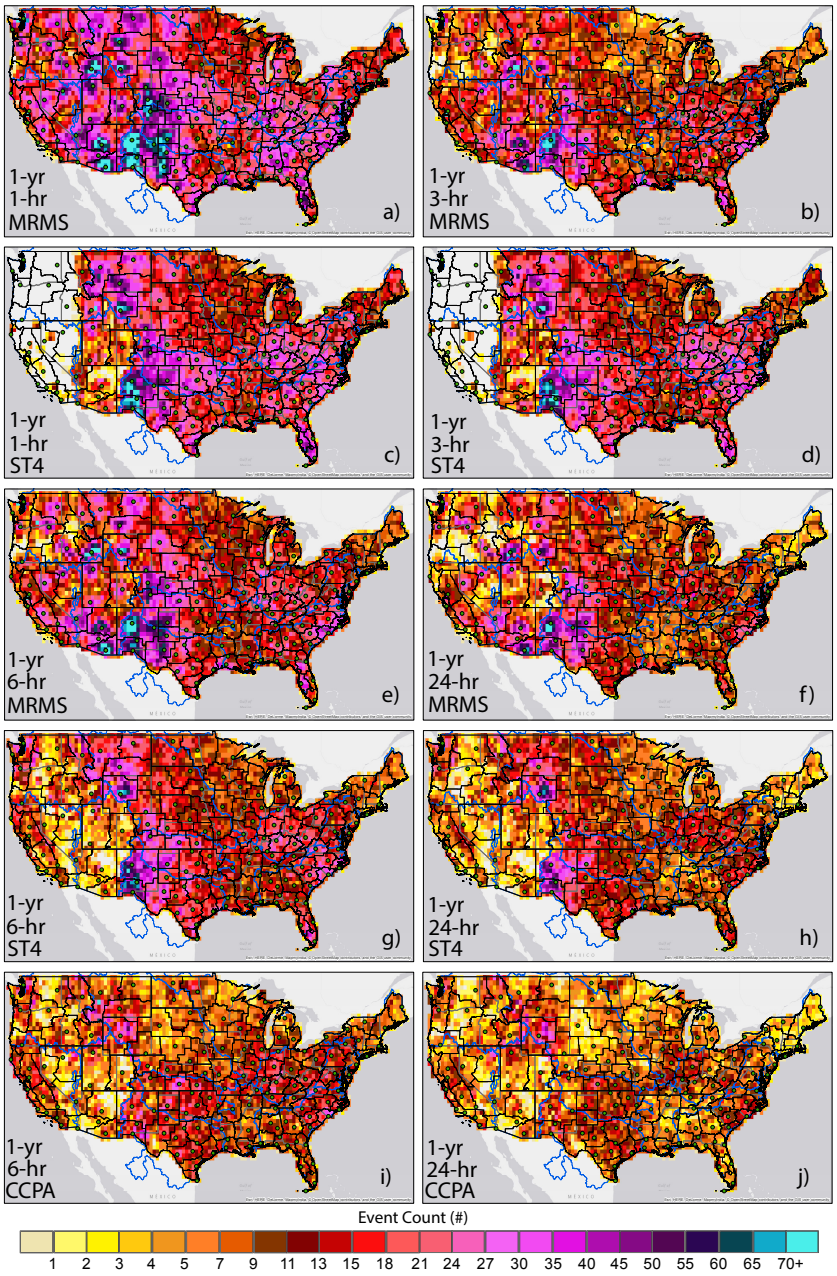


FIG. 2.5. Heat maps for ARI exceedances for different ARIs, AIs, and QPE sources across the period of record. Panels (a) and (b) correspond respectively to MRMS QPE exccedances of 1-year 1- and 3-hour ARIs; panels (c) and (d) are the same as (a) and (b), respectively, except for from ST4 QPE. Panels (e) and (f) illustrate MRMS QPE exceedances of the 1-year ARI for 6- and 24-hour accumulations, respectively; (g) and (h) show ST4 exceedances and (i) and (j) CCPA, both respectively for 6- and 24-hour accumulations. Symbology otherwise as in Figure 2.3.

Specifically, while the highest CSI values are obtained for ST4 and CCPA QPEs with FFRs, correspondence with FFWs is maximized using MRMS QPE exceedances. Consistent with using FFRs for truth,

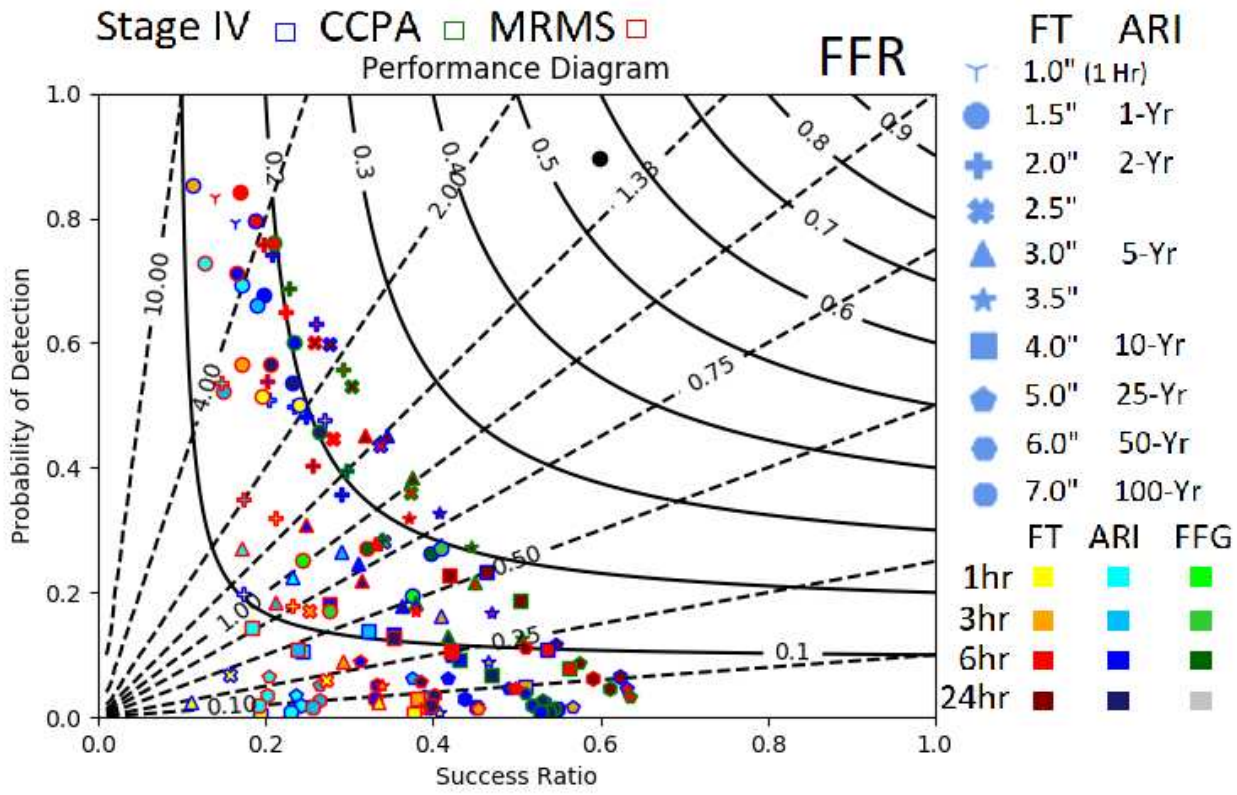


FIG. 2.6. Performance Diagrams as per Roebber (2009) evaluated over the entire period of record and across all of CONUS. Verification made with respect to FFRs for many different QPE threshold exceedances. The symbol shape corresponds to the threshold magnitude, as indicated in the top table of the panel legend. All FFG exceedances use a circular symbol. The inner color to each symbol indicates the accumulation interval associated with the comparison, as indicated in the middle table of the figure legend. Outer edge colors indicate the QPE source of the marker, with blue corresponding to ST4, green to CCPA, and red to MRMS as indicated at the bottom of the figure legend. The black circle depicts the verification of FFWs with respect to FFRs. Further description to aid with interpretation of PDs is included in the chapter text.

the 2.5 in. day^{-1} threshold provides the maximum CSI among the various QPE exceedance comparisons evaluated.

The PD view also allows for straightforward identification of some flaws and deficiencies in the QPE products. For example, as noted above, 3-hour ST4 QPE, which is acquired by summing 1-hour ST4 QPE, has less quality control and thus more spurious high values than compared with 6-hour ST4 QPE, which is a separate, independent grid and not necessarily equal to the sum of the six 1-hour QPEs that fall within the six hour period. This is evident in Figure 2.6, for example when comparing 3-hour FT exceedances with the same magnitude 6-hour FT exceedances. For the same QPE source and period of record, there should necessarily be as many or more exceedances of a given precipitation amount occurring over a six hour period than a three hour one. However, the orange circle surrounded by

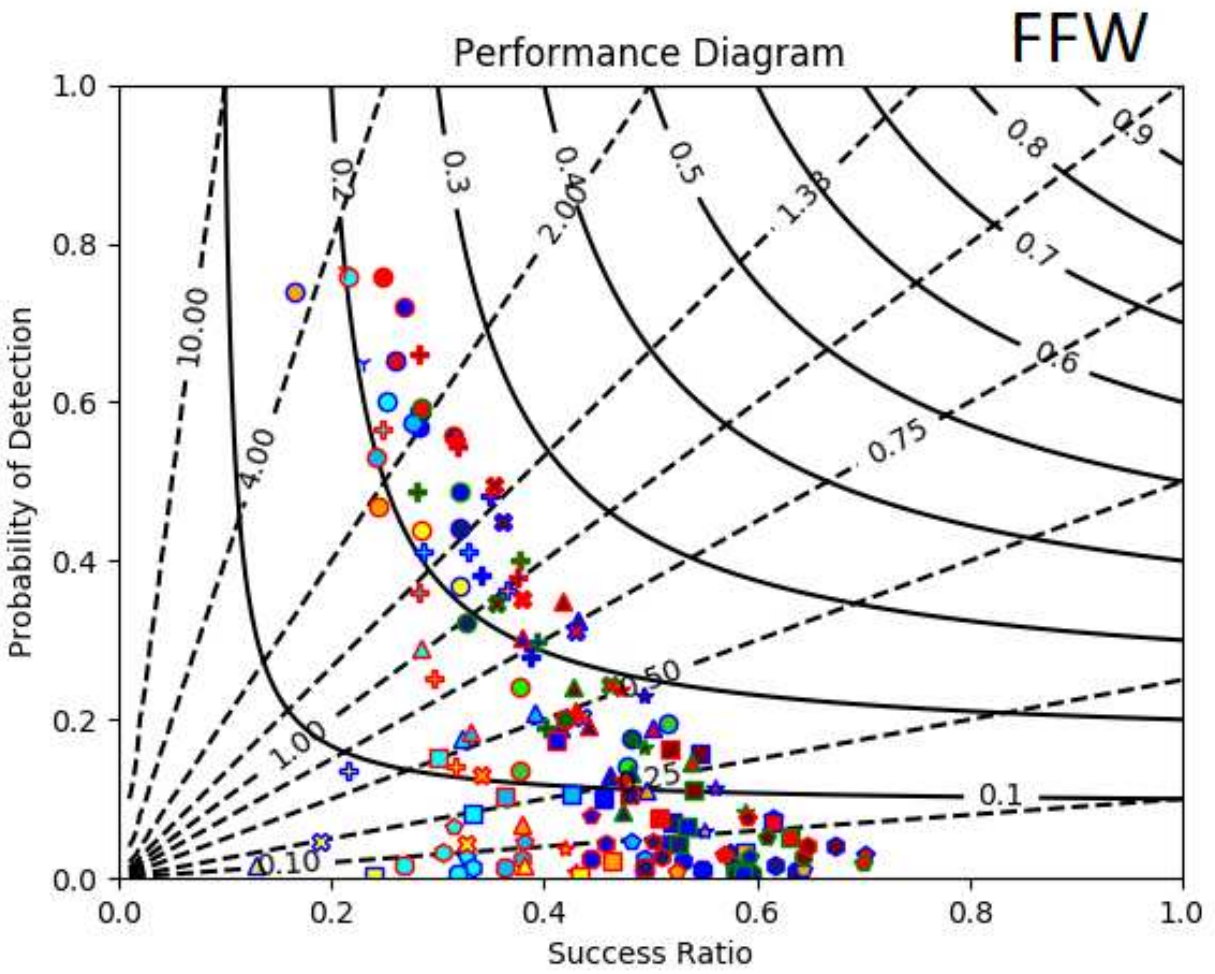


FIG. 2.7. Same as Figure 2.6, except for using FFWs as reference "truth".

blue, denoting ST4 QPE exceedances of 1.5 in. (38.1 mm) 3 hr.$^{-1}$ or higher, has a *higher* FB than the red circle surrounded by blue, denoting exceedances of the same threshold over a six hour period. Other deficiencies and limitations of QPE sources exist on a regional basis as well, some of which are discussed below.

Maps of CSI (Fig. 2.8) reveal considerable spatial variability in correspondence between different QPE exceedances and FFRs. FFWs (Fig. 2.8g) have much better correspondences with FFRs than any QPE threshold exceedance—as evidenced by the black dot of Figure 6—including FFG exceedances (Fig. 2.8a–f). Highest FFR-FFW CSI is found across much of CONUS east of the Rocky Mountains. Exceptions include central North Dakota, southern Florida, and Michigan, where there is an overall lack of reports (Fig. 2.3a). The number of reports is similarly scarce over much of the northern Intermountain West, resulting in low CSI scores there as well; in the extreme of no reports over a grid point, the

CSI is necessarily 0, since it is impossible to hit. In the West, scores are higher than surrounding areas in southern California and Nevada, where there are both more reports (Fig. 2.3a) and many more warnings (Fig. 2.3b) than in adjacent locations. FFG exceedances all exhibit somewhat similar CSI spatial distributions. 1-hour MRMS QPE exceedances of FFG (Fig. 2.8a) appear to perform the best of the six over the West, particularly in the southern Nevada and California vicinity. Correspondence with FFRs is generally highest over the Midwest and Southern Great Plains, with maximum correspondence for longer AIs (e.g. 6-hour, Fig. 2.8c,f) and with ST4 QPE providing better correspondence compared with MRMS (cf. Fig. 2.8b,e). CONUS-wide FFG-based CSIs, 0.1–0.2 depending on various choices, are quantitatively consistent with past findings over different study periods (Clark et al. 2014).

Comparing QPE exceedances of FFG (Fig. 2.8) with a sample of evaluated FT (Fig. 2.9) and ARI (Fig. 2.10) thresholds yields several interesting findings. Overall, largely because the base QPEs are the same and the only difference is the threshold discriminator between flash flood and non-event, the spatial character of correspondence between QPE exceedances and FFRs is broadly similar for each set of exceedances. MRMS QPE exceedances consistently yield the best correspondence with FFRs in the Southwest in southern California and Nevada, and the highest CSI appears to be achieved for 1-year 24-hour ARI exceedances (Fig. 2.10d) in that local area. ST4 QPE exceedances, particularly for longer AIs such as the 2.5 in. day^{-1} exceedances (Fig. 2.9h), appear to achieve the highest CSI across the broader Western region. Longer AIs, particularly in the ARI framework (Fig. 2.10), appear to exhibit improved correspondence compared with 1- and 3-hour QPEs across central and eastern CONUS as well (cf. Fig. 2.10i,l). The lower 1-year ARIs demonstrate superior correspondence with FFRs across CONUS compared with more extreme 10-year thresholds, with the one exception of 10-year 1-hour MRMS QPE exceedances, which provide optimum correspondence in the aforementioned small region surrounding Las Vegas and vicinity (Fig. 2.9e).

CSI maps illustrate that conclusions from aggregate nation-scale statistics do not always hold when zoomed to regional scales; PDs subsetted to particular regions allow for more quantitative analysis across the full spectrum of threshold comparisons. For example, using the region definitions shown in Fig. 2.11, the Northeast (NE; Fig. 2.12a) and Southeast (SE; Fig. 2.12b) regions exhibit appreciably different verification results to the national total. In particular, in both of these regions, ARIs clearly outperform the use of either FT or FFG thresholds, and ST4 outperforms MRMS and to a lesser extent CCPA, as evidenced by the blue interior and exterior symbols, respectively placed farther towards the upper right corner of each panel. However, the exact thresholds that obtain optimal skill vary between

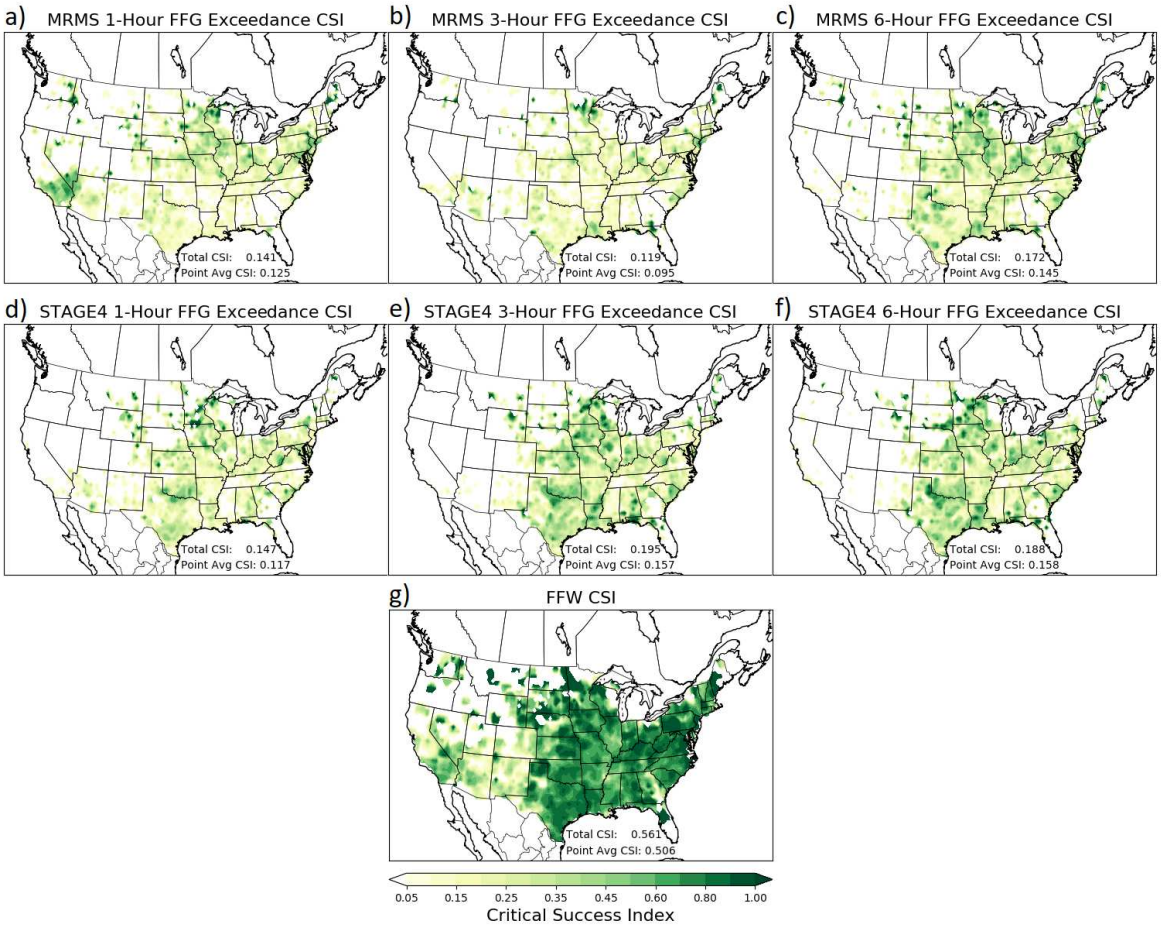


FIG. 2.8. CONUS-wide maps of CSI for comparisons between several sources with FFRs used as reference for "truth". Panels (a)–(f) correspond to exceedances of FFG based on MRMS (panels (a)–(c)) and ST4 (panels (d)–(f)) QPE for 1-hour, 3-hour, and 6-hour FFGs in panels (a) and (d), (b) and (e), and (c) and (f), respectively. Panel (g) depicts correspondence between NWS FFWs and FFRs over the period of record, again based on CSI. The top number in the bottom right of each panel shows the CSI for the corresponding threshold comparison when all observations contribute to a single set of hits, misses, and false alarms. The bottom number instead shows aggregate performance when the aggregate scores over the period of record are calculated individually for each grid point, and then averaged between all grid points.

the two regions; in the NE (Fig. 2.12a), the 2-year ARI achieves optimum CSI, while in the SE region (Fig. 2.12b), the 1-year ARI produces better results, with the 2-year exceedances being negatively biased. Moreover, while 6-hour AI exceedances produce maximum skill in comparison with FFRs across the NE, 24-hour accumulations are more predictive across the SE region. In the SE region, the 1-year 24-hour ARI exceedances are nearly equally skillful using ST4 and CCPA QPE, but ST4 is positively biased while CCPA is negatively biased. In both regions, FFWs correspond very well to FFRs, with CSIs of 0.76 and 0.63 in the NE (Fig. 2.12a) and SE (Fig. 2.12b) regions, respectively.

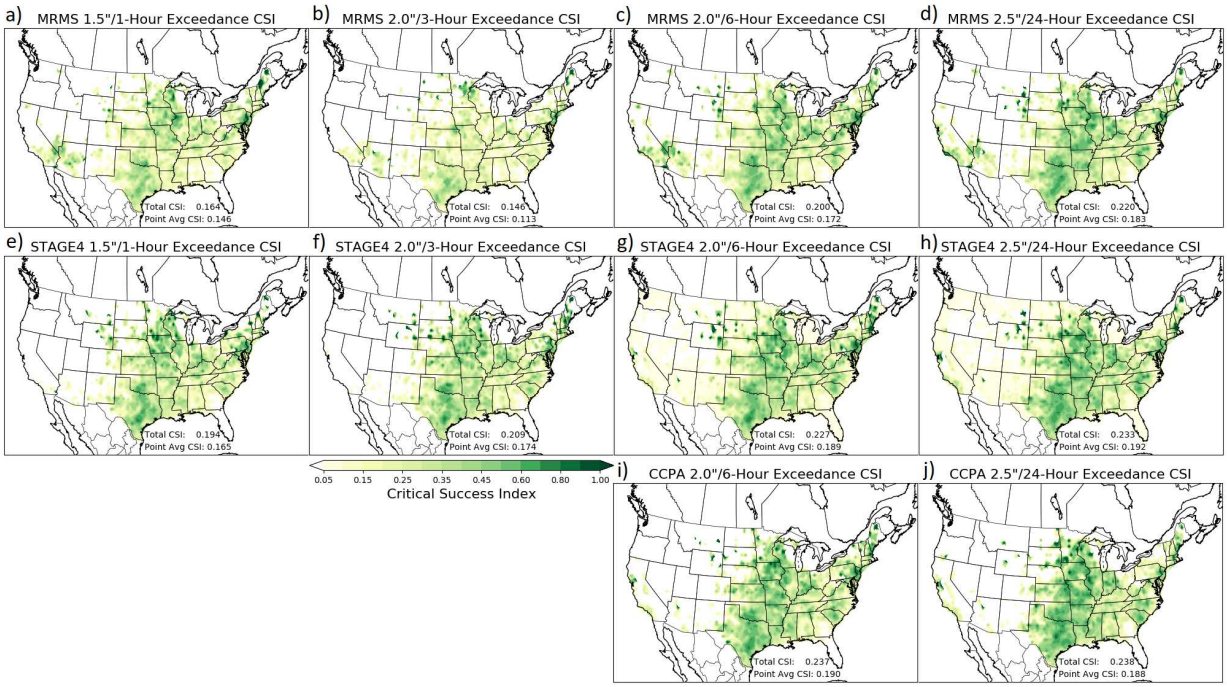


FIG. 2.9. Maps of CSI for comparisons between several QPE FT exceedances with FFRs used as reference for "truth". The top row (panels (a)–(d)) correspond to MRMS-based FT exceedances, the center row (panels (e)–(h)) depicts ST4-based FT exceedances, and the bottom row (panels (i)–(j)) compares CCPA-based FT exceedances with FFRs. Columns from left to right correspond to 1-, 3-, 6-, and 24-hour precipitation accumulations and 1.5, 2.0, 2.0, and 2.5 in. (38, 51, 51, 64 mm).

Over the Great Plains regions, NGP (Fig. 2.13a) and SGP (Fig. 2.13b), some mixed signals are found. In NGP, the 2.5 in. (63.5 mm) day^{-1} threshold using CCPA QPE attains the highest CSI score of all of the QPE threshold exceedance comparisons using FFRs as a reference. However, this scores very similar to 3.0 in. (76.2 mm) day^{-1} and 2.5 in. (63.5 mm) 6 hr^{-1} thresholds for ST4 QPE, with the 2.5 in. day^{-1} threshold suffering from too many false alarms. While the 2-year ARI produces the best results among the ARI thresholds considered, all ARI comparisons lag the best FT CSI values. FFG exceedances, while more competitive than in the eastern regions (Fig. 2.12), still lag FT exceedances in NGP (Fig. 2.13a). In SGP (Fig. 2.13b), the 1-year 24-hour ARI exceedances using CCPA QPE attain the higest CSI of any QPE comparison. 3.5 in. (88.9 mm) day^{-1} with ST4 QPE performs almost equally well—a higher threshold than in NGP owing to the wetter precipitation climatology in the region (e.g. Fig. 2.1). FFG is again somewhat competitive, especially for 6-hour accumulations, but lags the other methods.

The SW region (Fig. 2.11) displays very different verification characteristics (Fig. 2.14) to both the CONUS-wide perspective and the other individual regions examined above. Correspondence between

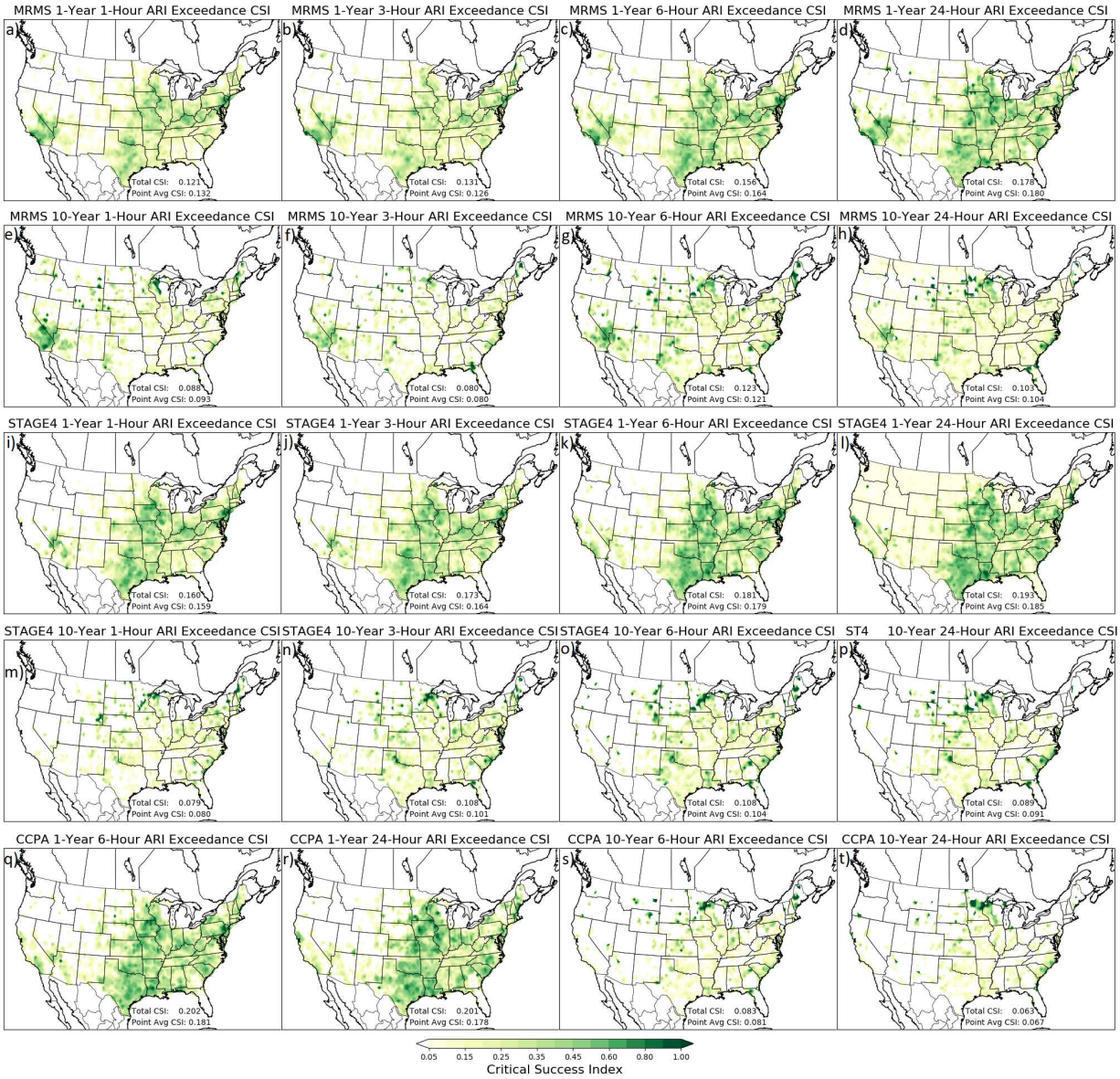


FIG. 2.10. Maps of CSI for comparisons between several QPE ARI exceedances with FFRs used as reference for "truth". The top two rows correspond to MRMS-based ARI exceedances, the bottom row depicts CCPA-based ARI exceedances, and the remaining two rows are associated with ST4-based ARI exceedances. Panels (a)–(d), (i)–(l), and (q)–(r) correspond to 1-year ARI exceedances, while (e)–(h), (m)–(p), and (s)–(t) are for 10-year exceedances. Columns from left to right correspond to 1-, 3-, 6-, and 24-hour precipitation accumulations, except for panels (q) and (r), which correspond to 6- and 24-hour accumulations, respectively.

all QPE exceedances and the reference truth are much worse than across the nation as a whole with both FFWs (Fig. 2.14b) and especially FFRs (Fig. 2.14a) serving as reference. The correspondence between the two reference truths is also particularly poor (Fig. 2.14a), with a CSI of only 0.3. The relative

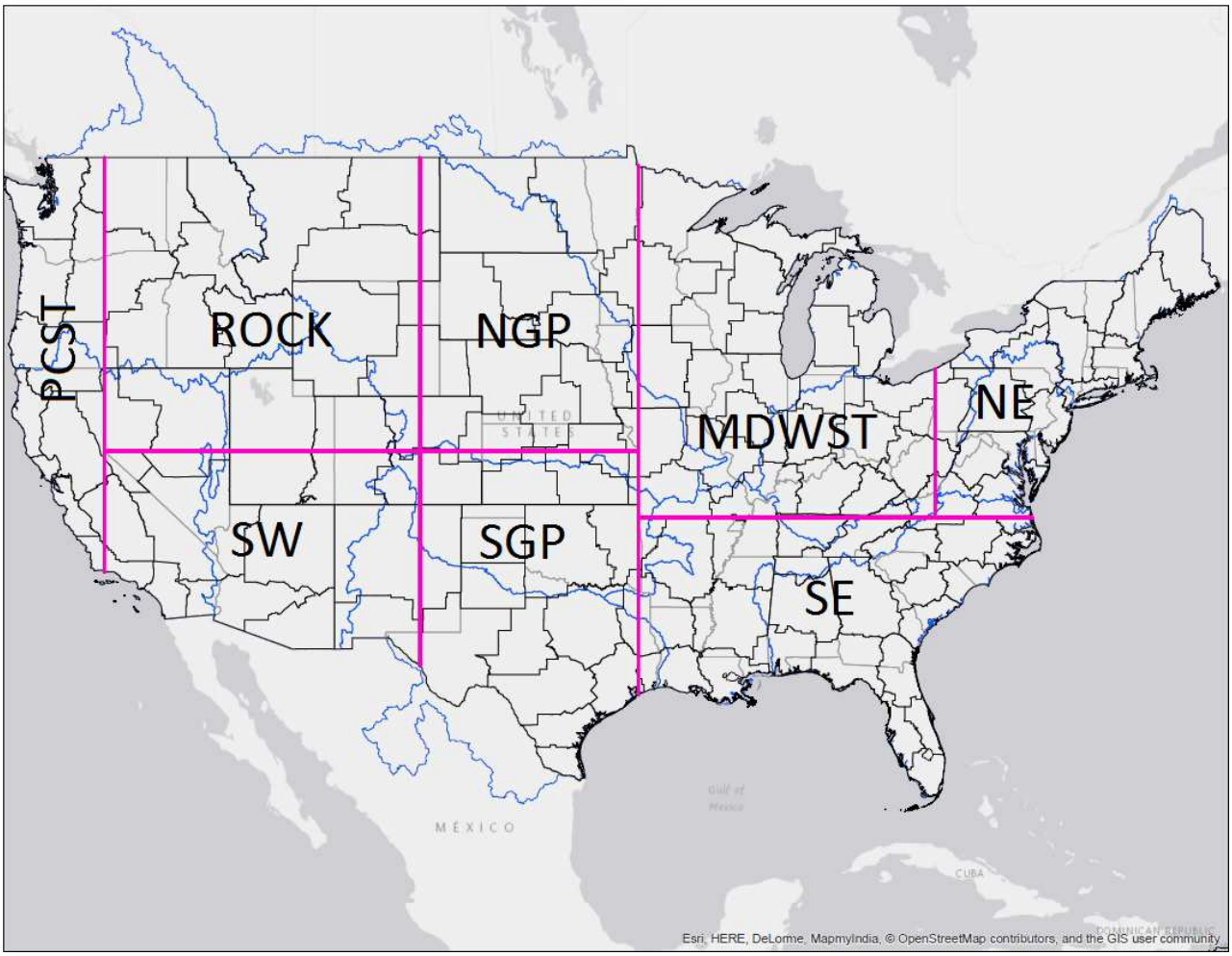


FIG. 2.11. Map depicting the regional partitioning of CONUS used in this study, and the labels ascribed to each region.

verification of the QPE exceedances also contrasts sharply with the results of other regions. Unlike over the East and Great Plains, MRMS QPE provides much better correspondence with both FFRs and FFWs than ST4 or CCPA QPE. Like in the East (Fig. 2.12), the ARI exceedances demonstrably outperform the FT and FFG exceedances in correspondence with both FFRs and FFWs (Fig. 2.14). But unlike other regions, especially for correspondence with FFRs (Fig. 2.14a), maximum CSI is attained for much shorter 1-hour AIs, and also for much higher ARIs, with maximum correspondence for the 10-year 1-hour ARI exceedances. Correspondence with FFWs (Fig. 2.14b) is higher across all comparisons. Furthermore, 3- and 6-hour exceedances attain similar CSI scores with 1-hour exceedances, and the highest CSI value is achieved with a much lower 2-year ARI. Much of this is attributable to the fact that there are many more—nearly three times as many (Fig. 2.14a)—FFWs than reports in this region, with a frequency bias

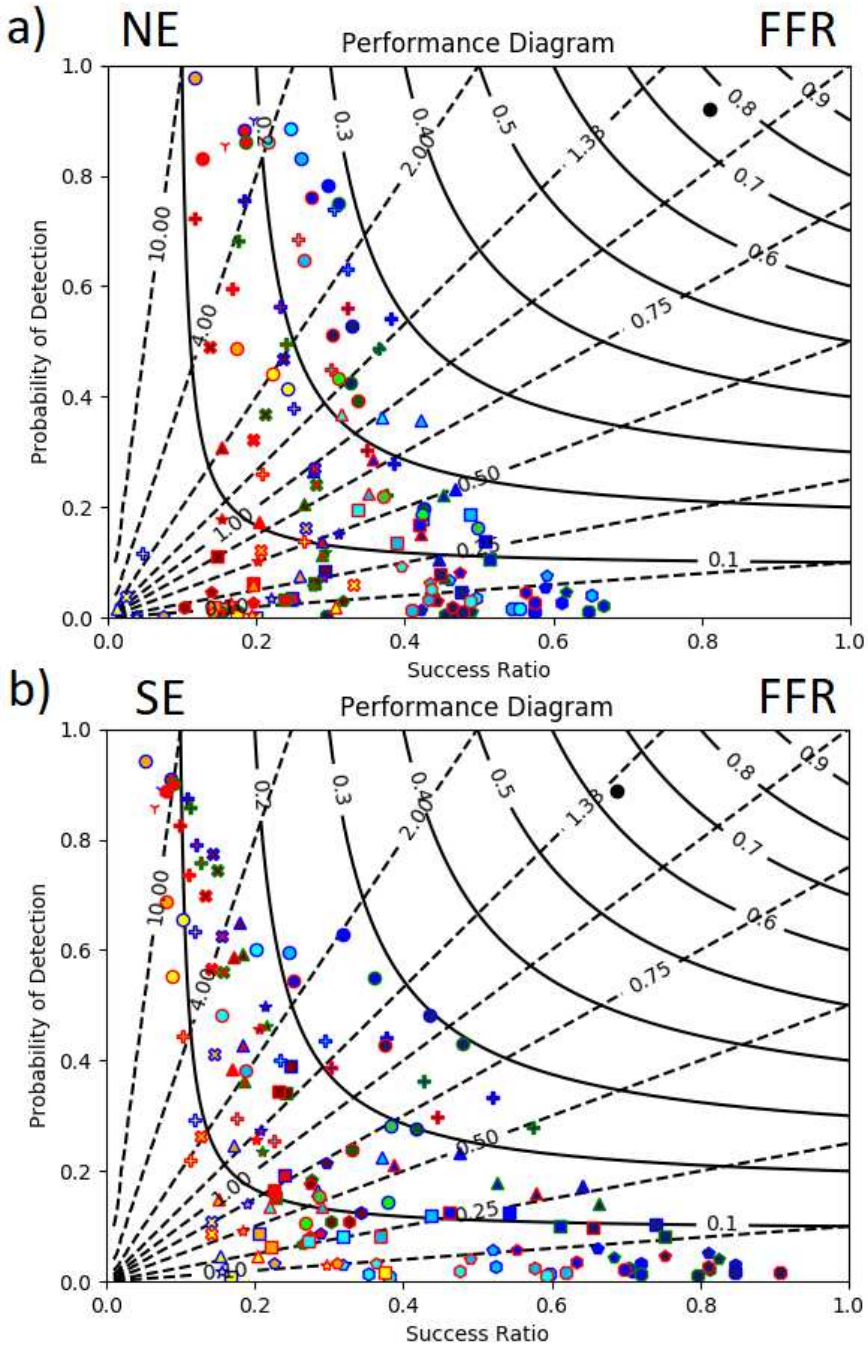


FIG. 2.12. Same as Figure 2.6, but with verification restricted to the NE region (panel (a)), and SE region (panel (b)), with region definitions as depicted in Figure 2.11.

of near 3 (Fig. 2.14a). Overall, each region exhibits unique verification characteristics, with optimum QPE sources, AIs, and threshold levels all varying by region.

Synthesizing every comparison into a single ETS for each QPE exceedance set (Fig. 2.15) yields concrete quantitative conclusions largely consistent with the CONUS-wide findings discussed above.

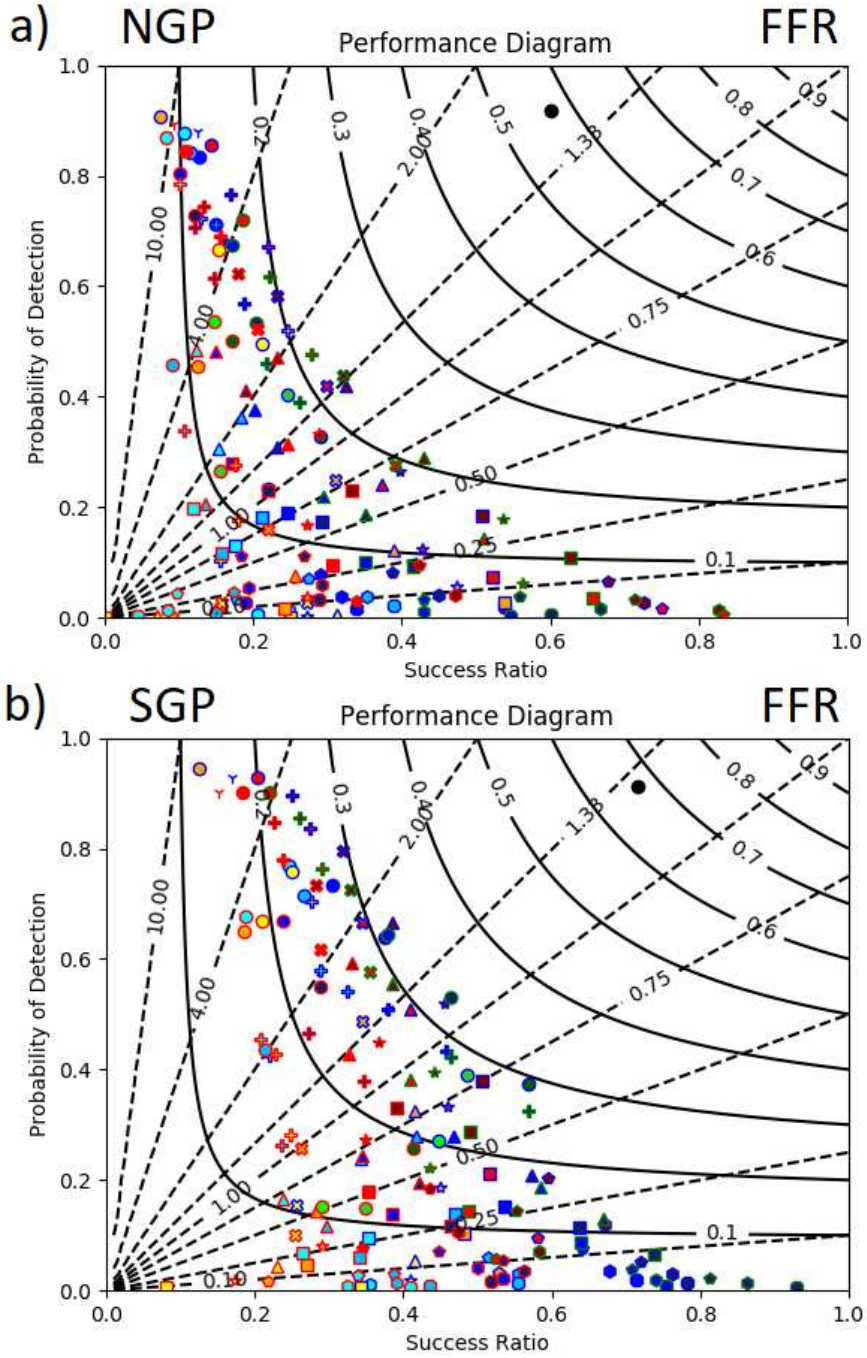


FIG. 2.13. Same as Figure 2.6, but with verification restricted to the NGP region (panel (a)), and SGP region (panel (b)), with region definitions as depicted in Figure 2.11.

Higher ETSs are found for longer 6-hour and 24-hour AIs for ST4 and MRMS QPE. The highest score using a fixed threshold attains a higher ETS than the maximum comparison with ARIs, which in turn outperforms the best corresponding FFG exceedance set with the reference truths. Overall, despite

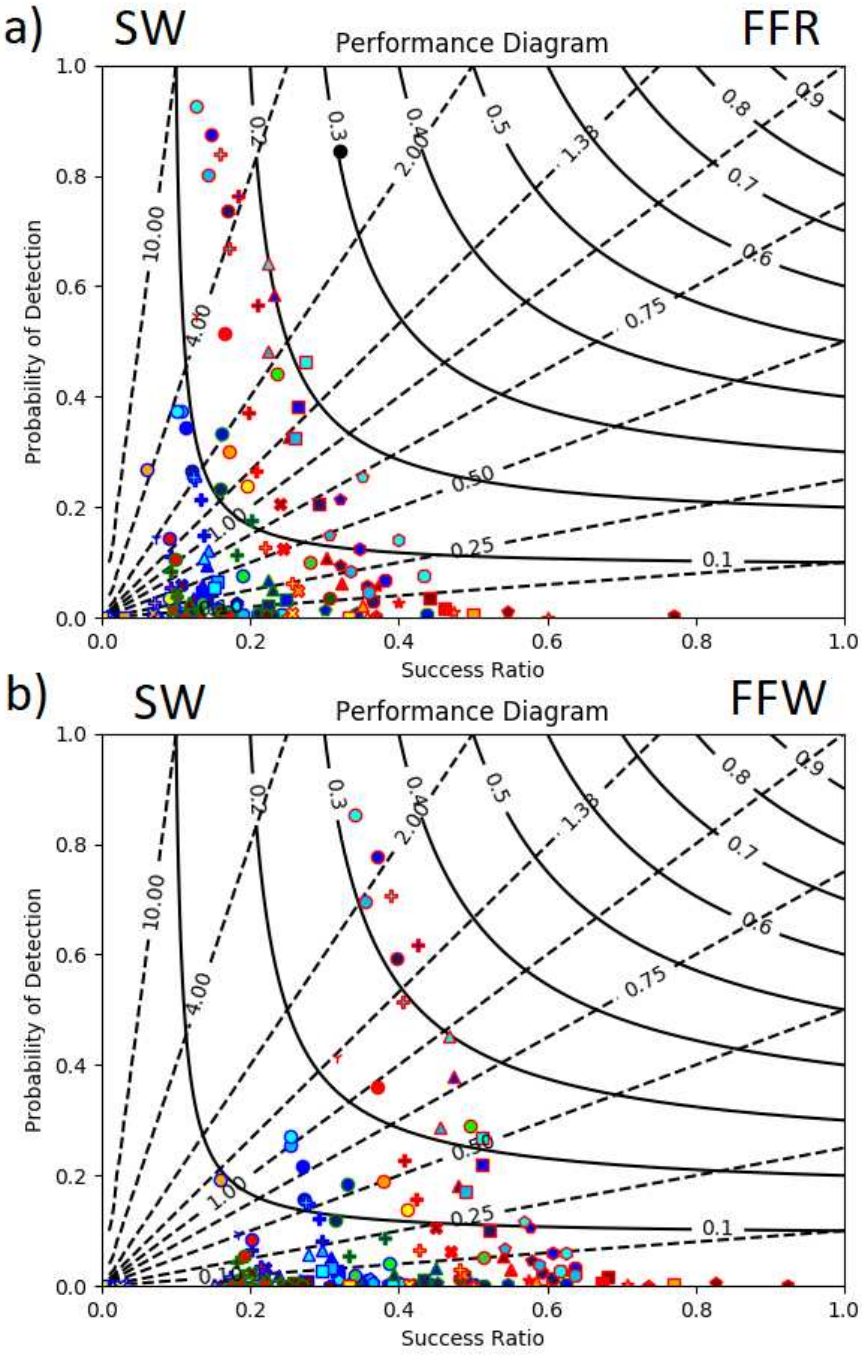


FIG. 2.14. Same as Figure 2.6 (panel a) and 2.7 (panel b), but with verification restricted to the SW region as defined in Figure 2.11.

very different characteristics, when averaged across CONUS, each of the three QPE sources evaluated achieved similar overall verification scores, all achieving a maximum ETS of almost 0.24. ST4 and MRMS achieved maximum ETS for a 2.5 in. day^{-1} threshold, while CCPA's maximum ETS was

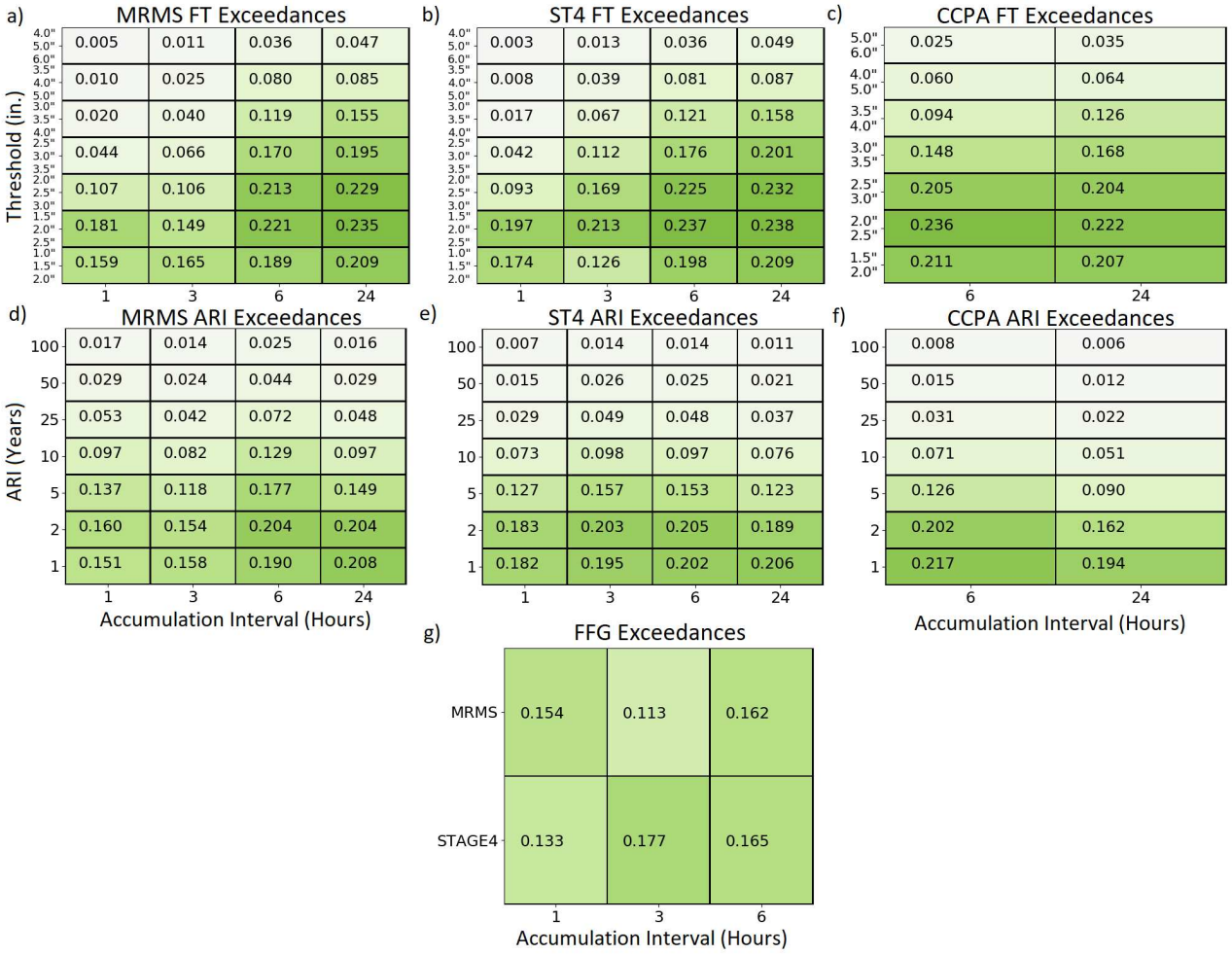


FIG. 2.15. Mean Equitable Threat Scores for each QPE exceedance method compared against both FFRs and FFWs, calculated as described in the chapter text. Panels (a)–(c) depict scores for FT QPE exceedance verifications for the MRMS, ST4, and CCPA QPE sources, respectively. Like accumulation intervals are organized by column, while thresholds are organized by row. For panels (a) and (b), the top number of each row label corresponds to the threshold for the 1-hour QPE exceedances, the middle number applies to the 3- and 6-hour accumulation comparisons, and the bottom number to the 24-hour QPEs. In panel (c), the top number corresponds to 6-hour QPEs and the bottom number to 24-hour QPEs. Panels (d)–(f) depict scores for QPE exceedances of ARI thresholds again respectively for the MRMS, ST4, and CCPA sources. Rows of these tables have a common ARI value, labeled in years; columns are again organized by accumulation interval. Panel (g) shows scores for FFG exceedances, with 1-, 3-, and 6-hour FFGs in the leftmost, center, and right columns, respectively, and comparisons with MRMS and ST4 QPEs respectively in the top and bottom rows.

for 2.0 in. 6 hour^{-1} exceedances. Among ARIs, best correspondence was obtained for 1-year 24-hour exceedances for MRMS and ST4 QPEs, and 1-year 6-hour exceedances for CCPA QPEs.

2.6 Summary and Conclusions

This study performed an expansive comparison using different QPE-based threshold exceedances as a proxy for flash flooding, as quantified through flash flood reports and NWS flash flood warnings. Comparisons were conducted across CONUS for an evaluation period spanning January 2015 through mid June 2017. Many different factors were considered, including the QPE accumulation interval, with 1, 3, 6, and 24-hour accumulations evaluated; the QPE source, with three leading QPE sources—ST4, MRMS, and CCPA—each compared; and the method for deriving local QPE thresholds. In addition to considering FT exceedances, exceedances of ARIs ranging from 1–100 years are considered, as well as exceedances of NWS RFC-generated FFG. For each of these binary observation sets, climatologies based on the study period were constructed, and skill in correspondence between the threshold exceedances and FFRs and/or FFWs was assessed on both national and regional scales. Ultimately, the study investigates the characteristics of the high tail of the probability distribution of QPEs, and the relation these QPEs have with observed flash flood impacts across CONUS.

Some of the findings from the study confirm prior knowledge of the hydrometeorological community, while in other areas, they introduce surprising and somewhat counterintuitive results. Even in the former case, the this study gives concrete, quantitative numbers to some of these differences previously known or quantified only qualitatively. The principal findings of this study can be summarized as follows:

- The question of whether a flash flood has occurred is much more involved than a simple binary comparison between local QPE and a flash flood threshold. No QPE threshold exceedance corresponded well with either FFRs or FFWs.
- While the aggregate skill statistics across CONUS were similar for each QPE source, significant regional differences emerged, with diminishing correspondence from east to west across CONUS. MRMS does outperform ST4 and CCPA in FFR and FFW correspondence in the Southwest, while ST4 performs best in the East and CCPA the best over central CONUS. With significant regional dependence, identification of existing deficiencies and areas for future product improvements can require regional, rather than purely national, analysis.
- Each QPE source has recurring deficiencies and biases. ST4 systematically reports heavy QPEs too frequently over much of the Intermountain West, but particularly in New Mexico and Wyoming, much more than other precipitation climatologies such as ARIs would indicate. It also suffers from numerous spurious very large values in its 1-hour QPEs that are not removed during quality control,

occurring especially but not exclusively in the West. CCPA corrects for many of these issues, but in its linear calibration, resultantly removes many legitimate extreme events. It consequently has a much lower frequency bias than either ST4 or MRMS for a given QPE threshold set. MRMS also experiences many of the biases observed with ST4 in the West, but to a lesser degree. However, it additionally exhibits a strong sensitivity to radar location in that region, with many more QPE exceedance events occurring near radar sites compared with more distant locations.

- Regardless of the threshold framework, very high thresholds often employed in extreme rainfall studies and analyses are too stringent to provide optimal correspondence with FFRs or FFWs owing to too many missed flash flood events. In general, the least severe thresholds examined had among the best correspondence with the reference records.
- Contrary to expectations given the definition of a flash flood, correspondence between QPE exceedances and the reference records tended to improve with increased accumulation interval. Minimum correspondence was generally obtained for threshold exceedances of 1-hour QPEs, and maximum correspondence for 24-hour QPE exceedances. On a regional basis, there were exceptions where shorter accumulations provided more skillful predictions, particularly in the Arid Southwest.
- Also surprisingly, FT exceedances provided slightly superior correspondence to FFRs and FFWs compared with FFG or ARI exceedances when a uniform threshold method was applied across CONUS. Overall, 2.5 in. day^{-1} provided the best correspondence with FFRs and FFWs of any threshold QPE exceedance examined, although 2.0 in. 6 hr^{-1} and others provided nearly equal ETS.
- In some regions of CONUS, FFG and/or ARI exceedances outperformed any FT exceedance, but the optimal ARI varied between 1- and 5-years, and occasionally higher for certain subregions, such as Florida and New Mexico.
- Among ARIs, the 1-year ARI provided the best predictions of FFRs across CONUS. For ST4 and MRMS, 24-hour accumulations performed best; for CCPA, 6-hour accumulations performed better. Among FFGs, 6-hour FFGs provided the best correspondence, but agreement was appreciably worse than with ARIs, which were in turn worse than the FT exceedances when applied uniformly across CONUS.
- Via their warnings, the NWS is able to add substantial value over automated QPE exceedances in projecting where heavy rain will produce reported flash flooding.

There are several limitations worthy of reemphasizing. Ultimately, this study has examined how well different QPE threshold exceedances correspond with flash flood *reports* (or warnings), and not true flash floods. FFRs have numerous non-physical reporting biases, including a tendency to underreport

flash floods in rural areas and at night. FFR frequency also varies by encoding practices of local WFOs, with some preferring to encode as a flood what another office may encode as a flash flood. A perfect "truth" does not exist, a fact which also serves as much of the motivation for conducting this comparative analysis. Compared with true incidences of flash flooding, FFRs likely underrepresent the true flash flooding climatology due to the aforementioned reporting and encoding practices. In particular, FFRs likely have few false alarms—most reports are indeed true events—but have numerous missed events. As a result, while verification is traditionally treated symmetrically, as it is also in this study, there is reason to believe those comparisons when evaluated against FFRs with frequency biases above unity likely have better correspondence with true flash flooding than those with biases below unity for the same CSI or ETS. It may be desirable in future work to incorporate the uncertainty of observations in the evaluation framework, penalizing non-hits in densely populated areas more than those in rural ones where "truth" is more uncertain, similar to that suggested in Weijs and Van De Giesen (2011) and elsewhere. Relatedly, some of the improved correspondence to FFRs illustrated by the FFWs over QPE threshold exceedances is likely artificial. WFOs have different proclivities to warn flash floods, and can for example choose to not warn storms that are likely to produce unrepported flash flooding, such as those confined to highly remote areas. They can also adopt different practices on encoding reports, and adopt different verification practices for warned and unwarned events (e.g. Barnes et al. 2007).

Nevertheless, there are several important implications from this analysis. Several prominent deficiencies are observed for each QPE source—some deficiencies are found in common between data sources, while others are unique to a particular source. In particular, all sources struggle with QPEs in the West. ST4 suffers from spurious very high values in areas of complex terrain, particularly in its 1-hour QPEs. This is especially prominent in the complex terrain of New Mexico. This phenomenon is seen, albeit to a lesser extent, across the rest of CONUS as well. MRMS has the same spurious high values over New Mexico, most prominently seen with ARI exceedances. While the root issues likely share commonalities with the ST4 deficiencies, MRMS appears to suffer from another major issue. Across the West, extreme QPEs occur with much larger frequency near radar sites compared with more distant locations—surely an artifact of the QPE derivation rather than a true spatially varying climatology given the number of sites exhibiting this behavior and the extent of correspondence. CCPA alleviates many of these problems, but removes many extreme events correctly identified by both ST4 and MRMS. This deficiency is prominent across all of CONUS. The lack of 1-hour CCPA QPEs also limit its utility in identifying flash flood scenarios in the SW and other regions where the best QPE-FFR correspondence

was identified for shorter accumulation intervals. Developers of QPE products may wish to further investigate some of these identified issues and adopt methods to alleviate them in order to generate more accurate and operationally useful products. A number of measures may assist with this, including improved quality control, particularly in the West, and statistical corrections tailored specifically for extremes, perhaps as a function of radar distance for MRMS, and to counteract the linear corrections made that necessarily and undesirably regress towards the climatological mean in the case of CCPA. Lastly, the verification in this study was limited to daily 1200–1200 UTC timescales. While this does not directly harm the verification performance of shorter AI exceedances compared with longer ones, the verification framework does not account for the fact that shorter AI QPE exceedances may provide additional information about the timing of flash flooding that the longer AI QPE exceedances cannot. Flash flood timing can be an important component of flash flood analysis, and gives the shorter AI QPE exceedances an advantage unaccounted for in this study's verification framework.

The analysis also lends some insight into current deficiencies with the ARIs. For example, there being nearly an order of magnitude more ARI exceedances in all three QPE sources over Wyoming and to a lesser extent in Montana, when ostensibly they should be the same everywhere over an infinitely long period of record, indicates that the ARI threshold estimates from the old NOAA Atlas 2 are likely too low in this region. Many of these places are very rural—and even moreso prior to Miller et al. (1973)—and the threshold estimates were likely inaccurate and highly uncertain in these areas at the time the threshold estimates were derived due to lack of sufficient robust observational data in these areas. It is expected that updated estimates through the NOAA Atlas 14 project will likely increase over the prior NOAA Atlas 2 estimates in these areas. In contrast, CCPA did not exhibit the high bias in New Mexico seen with MRMS and ST4 QPEs and were also seen for other thresholds sources, suggesting that the issues encountered in that region are more likely attributable to QPE limitations in those two sources rather than a fundamental ARI threshold estimate issue in this area.

The low frequency biases observed comparing FFG exceedances with FFRs seen across much of CONUS suggests that FFGs may be too high in many situations. It may be advisable to consider FFG calculation practices and evaluate whether any revisions can be made to increase spatial homogeneity across RFC boundaries and lower thresholds when appropriate to improve bias characteristics with respect to reported flash floods, similar to recommendations made in recent decades (Sweeney 1992; Carpenter et al. 1999). The findings of this study also raise implications about operational flash flood forecasting across a range of time scales. On longer time scales for example, the Weather Prediction

Center issues Excessive Rainfall Outlooks providing probabilistic guidance across CONUS for the current day out to two days ahead that sufficiently heavy rainfall will occur to produce flash flooding. Currently, forecast probabilities are defined with respect to exceedances of FFG. This provides a concrete framework for evaluating their outlooks and avoids many of the pitfalls associated with directly using FFRs or similar observational sources. The findings of this study suggest that, contrary to conventional wisdom, the product may have more utility in relating directly to precipitation impacts if instead one defines the outlook with respect to longer accumulation intervals, such as 6 or 24 hour QPE exceedances compared with the 1- and 3-hour FFG exceedances used in operations, and perhaps even using a homogeneous threshold, such as 2.5 in. (63.5 mm) day^{-1}. At shorter time scales, the findings further suggest that in assessing flash flood potential from a warning perspective, operational forecasters may wish to rely more heavily on one QPE source than another depending on their location. In the East, forecasters may wish to employ ST4 QPE more heavily, while relying more on MRMS QPE in the West and CCPA QPE across the Great Plains. Lastly, the findings shed insight into how QPF verification (e.g. Herman and Schumacher 2016a) and heavy precipitation forecast product development—such as those discussed in Chapters 3 and 4—may be conducted to be more physically relevant towards the impacts of heavy rainfall.

New flash flood analysis tools such as those described in Gourley et al. (2017), which use hydrologic models to provide additional insights, are becoming available in forecast operations. These tools have the potential to instill hydrometeorological insights beyond what can be gleaned from a simple inspection of QPE with respect to a threshold or thresholds. However, even in this framework, hydrologic guidance is only useful to the extent that its QPE input is accurate. It is hoped that the findings from this study helped to identify specific issues and areas the QPE products can be improved to alleviate recurring errors and biases, resulting in more representative outputs from analysis tools based on QPEs. In the meantime, knowledge and quantification of these deficiencies can improve human interpretation of derived analysis products by increasing (decreasing) confidence in areas that QPE is (not) skillful and damping (raising) perceived risk in areas that systematically have QPEs that are too high (low).

More investigation is required to further validate and constrain these findings. In addition, this work has not attempted to combine information from different sources to provide better correspondence between QPE exceedances and flash flood observations. Future work should examine these joint distributions to ascertain whether the full suite of QPE information can be used more effectively for

flash flood forecasting and analysis. This study also did not attempt to recalibrate QPEs with the specific focus of removing apparent systematic biases and improving their overall accuracy in heavy precipitation scenarios. Producing a CCPA-like correction to ST4 QPE, but employing different methodology geared towards the tail of the QPE distribution rather than the entire distribution would likely be a worthwhile and fruitful endeavor.

CHAPTER 3

MONEY DOESN'T GROW ON TREES, BUT FORECASTS DO: FORECASTING EXTREME PRECIPITATION WITH RANDOM FORESTS

3.1 INTRODUCTION

Locally extreme precipitation can cause a variety of costly, disruptive, and endangering impacts, including flooding, flash flooding, and landslides. In 2016 alone, these hazards combined caused more than 120 fatalities and $10 billion in damages over the United States (NWS 2017c). The prediction of flash floods is a notoriously challenging forecast problem, requiring not only accurate prediction of heavy rainfall magnitudes, but also of the spatiotemporal distribution of that rainfall; the hydrologic interactions between precipitation, terrain, and the land surface; and also of antecedent precipitation and its effects on soil conditions. Forecasting precipitation processes responsible for most observed extreme rainfall over the contiguous United States (CONUS) is often considered among the most challenging problems in contemporary numerical weather prediction (NWP; e.g. Fritsch and Carbone 2004; Novak et al. 2014). Given that the rainfall forecast alone presents such a considerable challenge, the additional hydrologic considerations in the flash flood forecast problem present an even more daunting task. While recent advances in heavy rainfall and flash flood forecasting have been made (e.g. Hapuarachchi et al. 2011; Novak et al. 2014; Barthold et al. 2015), forecasts still struggle in many situations (e.g. Delrieu et al. 2005; Lackmann 2013; Schumacher et al. 2013; Gochis et al. 2015; Nielsen and Schumacher 2016, among many others) and substantial progress remains to be made.

Contemporary operational dynamical forecast models often struggle to simulate accurately the physical processes responsible for extreme precipitation production. For example, models with parameterized convection often have a variety of persistent errors and biases associated with their depiction of convective systems, which are responsible for the majority of flooding rains over much of CONUS (e.g. Schumacher and Johnson 2006; Stevenson and Schumacher 2014; Herman and Schumacher 2016a). These include a tendency to underpredict total rainfall from convective systems (e.g. Schumacher and Johnson 2008; Herman and Schumacher 2016a); produce systems displaced too far to the north and west from where they are observed (e.g. Grams et al. 2006; Wang et al. 2009; Clark et al. 2010); initiate convection too early (e.g. Davis et al. 2003; Wilson and Roberts 2006; Clark et al. 2007); generate systems with too large an areal extent (e.g. Wilson and Roberts 2006); and propagate them incorrectly, too slowly, or not at all (e.g. Davis et al. 2003; Pinto et al. 2015). While convection-allowing

models (CAMs) can better resolve the physical processes responsible for heavy rainfall generation (e.g. Kain et al. 2006; Weisman et al. 2008; Duda and Gallus 2013), they too can suffer from many of these biases (e.g. Kain et al. 2006; Lean et al. 2008; Kain et al. 2008; Weisman et al. 2008; Herman and Schumacher 2016a). Furthermore, although there is a plethora of CAM guidance out to the day-ahead time frame (out to 36 hours to perhaps 48 hours after initialization), due to current computational constraints, there is almost no operational CAM guidance running out to two days ahead, and nothing operational that runs to three days ahead or beyond. Instead, global ensembles with parameterized convection serve as the primary source of forecast information and uncertainty quantification at these lead times. Nevertheless, there is considerable utility in skillful extreme precipitation forecasts at these longer lead times, since many mitigative actions that may not be feasible to execute in a matter of hours, but easily accomplished with a day or more of warning. Statistical post-processing of global ensemble output can potentially alleviate many of these dynamical model deficiencies and provide skillful extreme precipitation guidance at medium-range time scales. A specific focus on the Day 2–3 period is warranted due to the increased existing operational emphasis on these lead times compared with even longer ones, such as the Excessive Rainfall Outlooks produced by the Weather Prediction Center (Barthold et al. 2015) which forecast locally excessive rainfall across CONUS for Days 1–3.

There is a long history of successful application of statistical post-processing to dynamical model output (e.g. Klein et al. 1959; Glahn and Lowry 1972). Model Output Statistics (MOS; e.g. Glahn and Lowry 1972), is a simple, effective multivariate linear regression technique relating a set of dynamical model predictors to sensible weather predictands such as minimum and maximum temperature, wind speeds, and precipitation probability. This basic technique has long demonstrated skill over both the underlying models and even human forecasters (e.g. Jacks et al. 1990; Vislocky and Fritsch 1997; Hamill et al. 2004; Baars and Mass 2005), but is inherently limited by the linear assumptions underlying the method. Statistical post-processing techniques have also been successfully applied to QPFs, from early linear approaches (e.g. Bermowitz 1975; Antolik 2000) to more contemporary techniques that can exploit more complex variable relationships, including neural networks (e.g. Hall et al. 1999), reforecast analogs (e.g. Hamill and Whitaker 2006; Hamill et al. 2015), logistic regression (LR; e.g. Applequist et al. 2002; Whan and Schmeits 2018), random forests (RF; e.g. Gagne et al. 2014; Ahijevych et al. 2016; Gagne et al. 2017; Whan and Schmeits 2018), and other parametric techniques (e.g. Scheuerer and Hamill 2015; Whan and Schmeits 2018). For other meteorological applications, other machine learning algorithms, such as support vector machines (e.g. Zeng and Qiao 2011; Herman and Schumacher 2016b)

and boosting (e.g. Herman and Schumacher 2016b; Hong et al. 2016) have also been successfully applied. Related techniques have also been applied to forecasting related high-impact phenomena, such as severe hail (Brimelow et al. 2006; Gagne et al. 2015) and tornadoes (Alvarez 2014). One of the most powerful aspects of machine learning algorithms—and RFs in particular—is finding patterns and nonlinear interactions in the supplied training data (e.g. Breiman 2001). Depending on the extent and diversity of the data supplied in these experiments, trained RFs pose the theoretical capability of diagnosing and automatically correcting for various kinds of model biases, including context-dependent quantitative biases, such as QPF being systematically too high or too low; spatial displacement biases in the placement of extreme precipitation features; and, to some extent, temporal biases in the initiation or progression of extreme precipitation features.

This study makes a comprehensive investigation of using a global reforecast dataset to produce skillful and reliable probabilistic forecasts of locally extreme precipitation using the RF statistical post-processing technique in the medium-range. The following section provides further background and rigorously describes the data and methods used, algorithms employed, models trained, and experiments performed. Section 3 presents results of the sensitivity experiments conducted, while Section 4 presents the final results of the trained models and provides two brief case studies illustrating the process. Section 5 summarizes the findings of this study, outlines complementary analysis of these models, identifies avenues for further research, and discusses the broader implications of the results on numerical weather prediction and post-processing.

3.2 Data and Methods

There are several successive steps applied in creating the final forecasts evaluated in this study. A schematic overview of the forecast pipeline for the models trained in this study is depicted in Figure 3.1. Many types of hydrometeorological information are first taken, then assembled in a methodical manner, further pre-processed for subsequent analysis, analyzed using a statistical machine learning algorithm, and finally, extreme precipitation forecast guidance is produced and evaluated. This section details each of these setps in the model development and evaluation process.

3.2.1 Datasets

Dynamical model data used for training the RF models in this study comes from NOAA's Second-Generation Global Ensemble Forecast System Reforecast (GEFS/R; Hamill et al. 2013) dataset. The

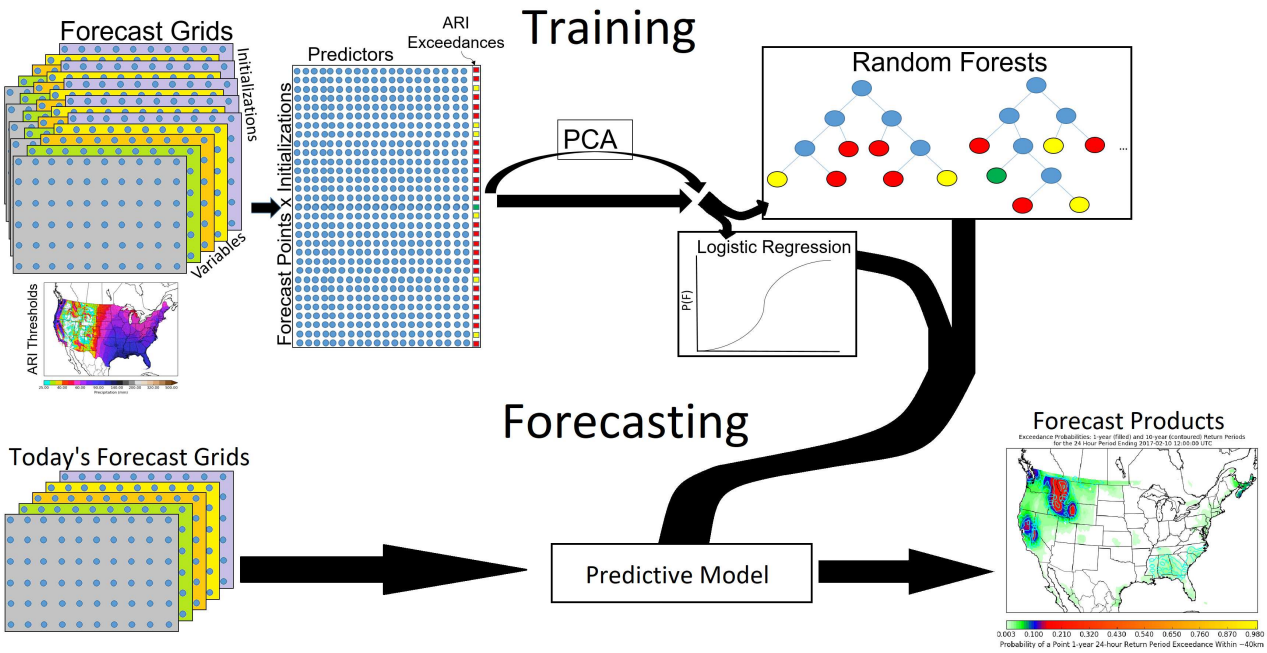


FIG. 3.1. Schematic representation of the forecast process for this study. GEFS/R forecasts are taken, assembled across fields, space, and time to form a training matrix, and past observations are used to associate a label with each forecast initialization, forecast day, forecast point triplet. The training matrix optionally undergoes pre-preocessing through principal component analysis, and then is input to one or more machine learning algorithms. From here, probabilistic ARI exceedance forecasts may be readily generated.

GEFS/R is a global 11-member ensemble with parameterized convection and T254L42 resolution—which corresponds to an effective horizontal grid spacing of ∼55 km at 40° latitude— initialized once daily at 0000 UTC back to December 1984. Perturbations are applied only to the initial conditions, and are made using the ensemble transform with rescaling technique (Wei et al. 2008). The ensemble system used to generate these reforecasts is nearly static throughout its 30+ year period of coverage, though updates to the operational data assimilation system over time have resulted in some changes in the bias characteristics of its forecasts over the period of record (Hamill 2017). Some forecast fields are preserved on the native Gaussian grid (∼ 0.5° spacing), while others are available only on a 1° × 1° grid. Temporally, forecast fields are archived every three hours out to 72 hours past initialization, and are available every six hours beyond that. This study employs an almost 11-year period of record to explore this forecast problem, using daily initializations from January 2003 through August 2013.

In creating probabilistic extreme precipitation forecast guidance, the predictand must first be concretely specified and a robust, consistent verification framework established. One of the many challenges in heavy rainfall and flash flood forecasting is the considerable difficulty in verifying events (e.g.

Welles et al. 2007; Gourley et al. 2012; Barthold et al. 2015), as every approach has its deficiencies and limitations. It is attractive to consider the problem from a simple perspective of quantitative precipitation estimate (QPE) exceedances of some temporally static threshold. In particular, a fixed threshold (e.g. 50 mm hr^{-1}) can be used as a proxy for flash flooding (e.g. Brooks and Stensrud 2000; Hitchens et al. 2013), as can exceedances of thresholds defined relative to the local precipitation climatology (e.g. Schumacher and Johnson 2006; Stevenson and Schumacher 2014; Herman and Schumacher 2016a), such as average recurrence intervals (ARIs). An ARI defines a fixed frequency relative to the hydrometeorological climatology of the region; in particular, it corresponds to the expected duration, given the local climatology, between exceedances of a given threshold. For example, the 1-year ARI for 24-hour precipitation accumulations describes the accumulation amount for which one would expect the mean duration between exceedances of said amount to be one year. Past research has shown that a fixed-frequency ARI-based framework has better correspondence with heavy precipitation impacts than the use of any fixed threshold across the hydrometeorologically diverse regions of CONUS (e.g. Reed et al. 2007). From the perspective of forecast verification, defining extreme precipitation with respect to a fixed threshold exceedance raises challenges when applied uniformly across CONUS. For example, skill differences observed between regions may simply be an artifact of a regionally varying event climatology rather than "true" regional differences in forecast skill (e.g. Hamill and Juras 2006). The ARI framework avoids this issue and provides reasonable correspondence with precipitation impacts while avoiding the additional complications such as antecedent conditions, local hydrology, and urban effects (e.g. Herman and Schumacher 2016a) and is consequently used to quantify extreme rainfall for this study.

Specifically, forecast probabilities are issued for 24-hour ARI exceedances at each GEFS/R archive grid point on its native Gaussian grid at all points across CONUS, using a predictand with three categories: 1) No 1-year ARI exceedance at any point within the grid point domain, 2) At least one 1-year ARI exceedance, but no 10-year ARI exceedances within the grid point domain, and 3) At least one 10-year ARI exceedance within the grid point domain. For evaluation, probabilities from the middle and most severe categories are often aggregated to produce a 1-year ARI exceedance probability. This approach has the advantage of retaining aspects of the anticipated event severity as would be retained in a regression context but is largely lost when performing single category classification. While there can be some additional complications especially with respect to calibration, formulating the prediction problem as a single multicategory classification task rather than multiple distinct binary category

models also ensures mathematical consistency of the exceedance probabilities within the generated probability mass functions in a way that the latter approach would not.

In aggregating multiple QPE-to-ARI threshold grid point comparisons in a single predictand, the forecasts issued correspond to neighborhood event probabilities, an increasingly popular method of communicating probabilistic high-impact weather information in forecast operations (e.g. Barthold et al. 2015; NWS 2017b). Counting any one of several possible point exceedances as an "event" results in the event having a higher observed relative frequency relative to that of any of the individual point exceedances; the event frequency in this framework thus exceeds the purported frequency suggested by the ARI. However, the fixed-frequency property, and thus many of the aforementioned desirable properties of the framework, are approximately retained. For this study, focus is placed exclusively on two 24-hour forecast periods: the 1200–1200 UTC period corresponding to forecast hours 36–60 from the GEFS/R forecast fields and the subsequent 24-hour period encompassing forecast hours 60–84, denoted respectively as Days 2 and 3. At these times, there is typically some knowledge to characterize the environmental conditions in which precipitation may form, but it is beyond the current range of operational CAM guidance.

Verification comes from the National Centers for Environmental Prediction (NCEP) Stage IV Precipitation Analysis (Lin and Mitchell 2005) QPE product, created operationally since December 2001. Stage IV provides 24-hour analyses over the CONUS on a ∼4.75 km grid. It uses both rain gauge observations and radar-derived rainfall estimates to generate an analysis, and is further quality controlled via NWS River Forecast Centers (RFCs) to ensure stray radar artifacts and other spurious anomalies do not appear in the final product. Despite some limitations (Herman and Schumacher 2016a; Nelson et al. 2016), its analysis quality; resolution, allowing better ability to capture precipitation extremes compared with other QPE products (e.g. Hou et al. 2014); and data record length make it preferable to analogous products.

The ARI thresholds associated with the 1- and 10-year ARIs for 24-hour precipitation accumulations are generated using the same methodology as Herman and Schumacher (2016a), where CONUS-wide thresholds are produced by stitching thresholds from several sources. NOAA's Atlas 14 thresholds (Bonnin et al. 2004, 2006; Perica et al. 2011, 2013), an update from older work and currently under development, are used wherever they were available at the commencement of this study. For five northwestern states—Washington, Oregon, Idaho, Montana, and Wyoming—updated thresholds are

not available, and derived NOAA Atlas 2 threshold estimates are used instead (Miller et al. 1973). Additionally, in Texas and the Northeast—New York, Vermont, New Hampshire, Maine, Massachusetts, Connecticut, and Rhode Island—Technical Paper 40 (TP-40; Hershfield 1961) thresholds are used[1]; everywhere else uses the Atlas 14 threshold estimates. The 10-year ARI thresholds (Fig. 3.2b) show a similar spatial pattern to the 1-year ARI thresholds (Fig. 3.2a), but are substantially higher everywhere. More significantly, it is apparent that at both severity levels, there are large regional disparities in the threshold magnitudes. Over climatologically wet regions of CONUS, such as the Pacific coastal mountains and immediately along the Gulf Coast, thresholds are as high as 100–150 mm and 250–300 mm for 1-year and 10-year ARIs, respectively. Over central and eastern CONUS, thresholds tend to decrease smoothly with increasing latitude and distance from major bodies of water. Sharper variations are seen in areas of complex terrain over western CONUS. In the driest parts of the Arid Southwest and Intermountain West, thresholds can be as low as 10–15 mm and 25–30 mm for the two ARI levels—a full order of magnitude difference from the largest thresholds at the same intensity level.

Forecast models in this study are trained separately for eight distinct, yet cohesive and internally fairly hydrometeorologically homogeneous regions of CONUS, using the delineation indicated in Figure 2.11. Observed 1- and 10-year ARI exceedance events that occurred during the period of record (Fig. 3.2c,d) highlight important regional differences in the seasonal climatology of ARI exceedances across CONUS. In the Pacific Coast (PCST) region, the vast majority of exceedances at both the 1-year and 10-year severity levels occur in the cool-season, and occur largely from atmospheric river events with large moisture transport impinging on coastal topography (e.g. Rutz et al. 2014; Herman and Schumacher 2016a). This seasonality holds to a lesser extent in the neighboring Southwest (SW) region, with some signal carrying over to the Rockies (ROCK) region as well. In the central and eastern regions, the majority of events occur during the warm-season from more scattered convective-scale processes, particularly in the months of May, June, and July (e.g. Schumacher and Johnson 2006; Herman and Schumacher 2016a). Tropical cyclones can cause widespread and very significant rainfall, and comprise a substantial portion of the extreme precipitation climatology, especially in the Northeast (NE) and Southeast (SE) regions. Due to the spatial extent of their impacts and immense rainfall totals they can produce, they form a much larger fraction of the climatology of 10-year ARI exceedances (Fig. 3.2d) than 1-year events (Fig. 3.2c). Additionally, the numbers are lower than would be expected; by the explicit exceedance frequencies associated with the thresholds, one would expect an average of one

[1]The northeastern states did receive updated Atlas 14 estimates in October 2015, but TP-40 thresholds were retained for consistency with prior work.

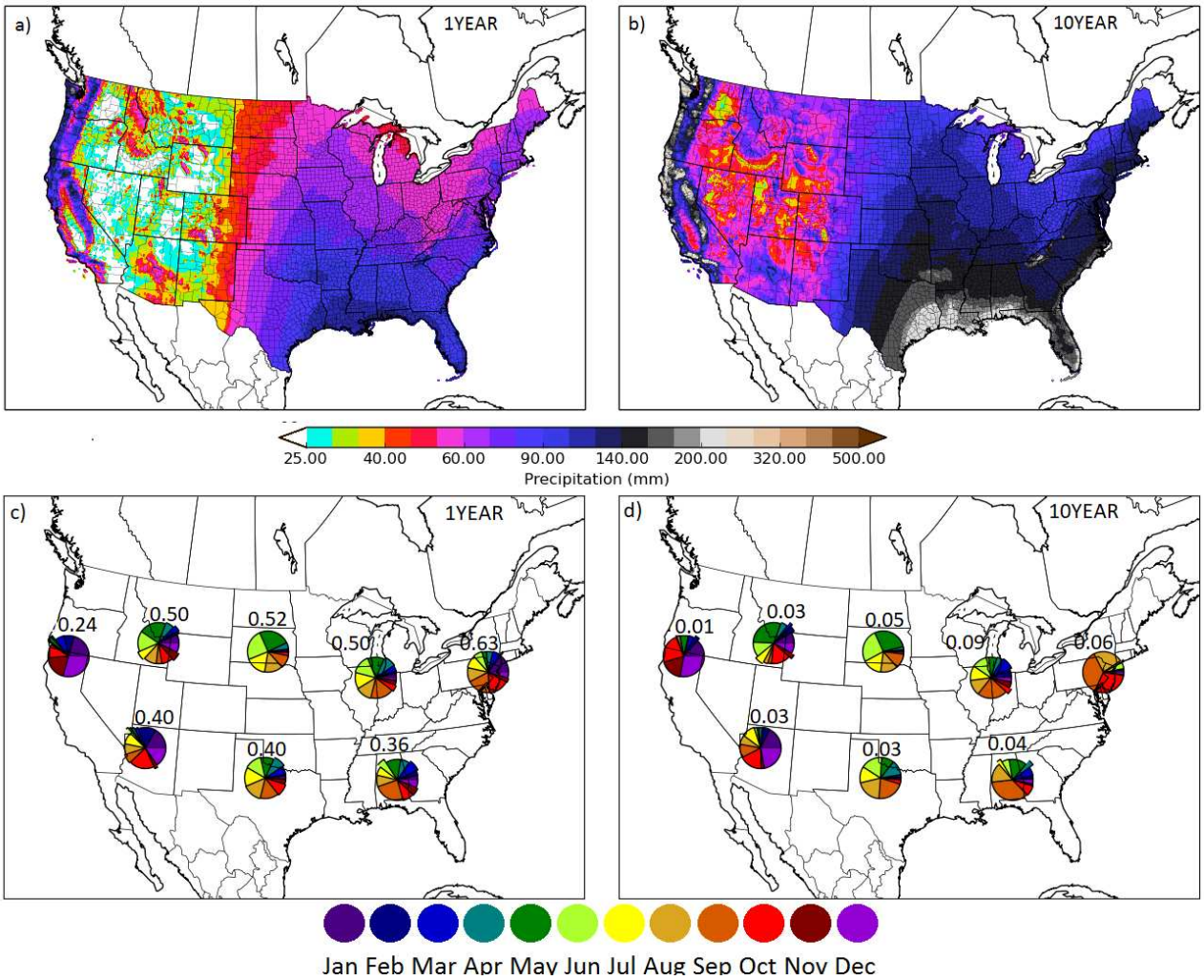


FIG. 3.2. ARI thresholds at the (a) 1-year and (b) 10-year ARI levels over CONUS for a 24-hour accumulation interval. Climatology of observed exceedances of the (c) 1-year, 24-hour ARI thresholds and (d) 10-year, 24-hour ARI thresholds between January 2003 and August 2013 based on Stage IV Precipitation Analysis. Pie charts indicate the monthly distribution of event occurrence within each study region as shown in Figure 3. Numbers above the pie charts indicate the mean number of exceedances per point per year within the region (a priori 1 and 0.1 for 1-year and 10-year ARIs, respectively).

exceedance per point per year over the period of record for the 1-year events (Fig. 3.2c) and 0.1 exceedances for 10-year events (Fig. 3.2d); in reality, event counts are only approximately half of that. This is consistent with previous findings (e.g. Herman and Schumacher 2016a), and likely in part attributable to limitations in the Stage IV product to capture extremes (e.g. Nelson et al. 2016). There is also quite a bit of region-to-region variability in event counts, particularly for 10-year exceedances, much of which is attributable to statistical variability from having a short data record in relation to the event frequency.

TABLE 3.1. Summary of dynamical model fields examined in this study, including the abbreviated symbol to which each variable is referred throughout the paper, a description of each variable, the predictor group with which the field is associated in the chapter text, and the highest resolution for which the field can be obtained from the GEFS/R.

Symbol	Description	Predictor Group	Grid
APCP	Precipitation accumulation in past (3) 6 hours	Core	Native Gaussian
CAPE	Surface-based convective available potential energy	Core	Native Gaussian
CIN	Suface-based convective inhibition	Core	Native Gaussian
MSLP	Mean sea level pressure	Core	Native Gaussian
PWAT	Total precipitable water	Core	Native Gaussian
Q2M	Specific humidity two meters above ground	Core	Native Gaussian
T2M	Air temperature two meters above ground	Core	Native Gaussian
U10	Zonal-component of 10-meter wind	Core	Native Gaussian
V10	Meridional-component of 10-meter wind	Core	Native Gaussian
Q300	Specific humidity at 300 hPa	Upper-Air Extra	$1° \times 1°$
Q500	Specific humidity at 500 hPa	Upper-Air Core	$1° \times 1°$
Q700	Specific humidity at 700 hPa	Upper-Air Extra	$1° \times 1°$
Q850	Specific humidity at 850 hPa	Upper-Air Core	$1° \times 1°$
T250	Temperature at 250 hPa	Upper-Air Extra	$1° \times 1°$
T500	Temperature at 500 hPa	Upper-Air Core	$1° \times 1°$
T700	Temperature at 700 hPa	Upper-Air Extra	$1° \times 1°$
T850	Temperature at 850 hPa	Upper-Air Core	$1° \times 1°$
U250	Zonal-component of 250 hPa wind	Upper-Air Extra	$1° \times 1°$
U500	Zonal-component of 500 hPa wind	Upper-Air Core	$1° \times 1°$
U700	Zonal-component of 700 hPa wind	Upper-Air Extra	$1° \times 1°$
U850	Zonal-component of 850 hPa wind	Upper-Air Core	$1° \times 1°$
V250	Meridional-component of 250 hPa wind	Upper-Air Extra	1°x1°
V500	Meridional-component of 500 hPa wind	Upper-Air Core	1°x1°
V700	Meridional-component of 700 hPa wind	Upper-Air Extra	1°x1°
V850	Meridional-component of 850 hPa wind	Upper-Air Core	1°x1°
W850	Vertical velocity (omega) at 850 hPa	Upper-Air Core	1°x1°

3.2.2 Predictor Assembly

Input predictors, or features, to the random forests can be partitioned into two categories: model predictors and background predictors; the former constitute the vast majority of inputs. Model predictors come from atmospheric fields forecast in the GEFS/R which bear a known physical relationship with extreme precipitation. A core set of f=9 fields used in this study are: accumulated precipitation (APCP), convective available potential energy (CAPE), convective inhibition (CIN), precipitable water (PWAT), surface temperature (T2M) and specific humidity (Q2M), surface zonal (U10) and meridional winds (V10), and mean sea level pressure (MSLP). Sensitivity experiments explore the use of additional upper-air atmospheric fields; a full list of fields used in this study, their associated symbols used in this chapter, and the grids on which they are each archived is included in Table 3.1. The spatiotemporal

TABLE 3.2. List of background predictors used in this study, and their associated symbols and descriptions.

Symbol	Description
ARI1_LOCAL_MEDIAN	Median of 1-year ARIs whose closest GEFS/R grid point is the forecast point.
ARI1_LOCAL_MIN	Minimum of 1-year ARIs whose closest GEFS/R grid point is the forecast point.
ARI1_LOCAL_MAX	Maximum of 1-year ARIs whose closest GEFS/R grid point is the forecast point.
ARI10_LOCAL_MEDIAN	Median of 10-year ARIs whose closest GEFS/R grid point is the forecast point.
ARI10_LOCAL_MIN	Minimum of 10-year ARIs whose closest GEFS/R grid point is the forecast point.
ARI10_LOCAL_MAX	Maximum of 10-year ARIs whose closest GEFS/R grid point is the forecast point.
ARI1_REGIONAL_MEDIAN	Median of 1-year ARIs that lie within the domain from which model predictors are drawn.
ARI1_REGIONAL_MIN	Minimum of 1-year ARIs that lie within the domain from which model predictors are drawn.
ARI1_REGIONAL_MAX	Maximum of 1-year ARIs that lie within the domain from which model predictors are drawn.
ARI10_REGIONAL_MEDIAN	Median of 10-year ARIs that lie within the domain from which model predictors are drawn.
ARI10_REGIONAL_MIN	Minimum of 10-year ARIs that lie within the domain from which model predictors are drawn.
ARI10_REGIONAL_MAX	Maximum of 10-year ARIs that lie within the domain from which model predictors are drawn.
LAT	Latitude of forecast point.
LON	Longitude of forecast point.

variations in these fields are considered as well. Spatially, predictors are structured in a forecast-point relative sense. In the control model, GEFS/R forecast values up to r=4 grid boxes ($\sim 2°$) latitudinally or longitudinally displaced in any direction relative to the forecast point are considered. Temporally, simulated fields are considered at each archive time during the forecast interval, which corresponds to every three hours during the Day 2 period and every six hours during the Day 3 period, for a total of t=9 and t=5 forecast periods for the Day 2 and 3 periods, respectively. All told, this yields $tf(2r+1)^2$ model predictors, which yields respectively M=6,561 and M=3,645 model predictors for the Day Two and Day Three control models. The other category of predictors, background predictors (Table 3.2), are those which are solely associated with the forecast point, and have no relation to the present meteorology. These include the location of the point, as well as the ARI characteristics of the point and in the surrounding area.

3.2.3 Dimensionality Reduction

There are a large number of model predictors, and they are also highly correlated—spatially, temporally, and across variables. With millions of training examples and thousands of features, the forecast problem can become computationally intractable. Further, having many highly correlated features can readily result in model overfitting—making predictions based on noise affecting an individual native feature rather than the underlying signal—a phenomenon commonly termed the "curse of dimensionality" (e.g. Friedman 1997). There are numerous ways these concerns can be addressed; broadly speaking, the most common approaches are either feature selection or feature extraction. In feature selection, a subset of initial predictors are chosen that collectively bear the strongest predictive relationship with the predictand, whereas in feature extraction, a smaller set of new predictors are derived from the

original set. Both of these procedures can be performed subjectively through manual means or objectively through automated means. In this case, all of the input predictors are believed to have a physical relationship with extreme precipitation, and choosing only the most predictive fields (e.g. model QPF) and discarding the rest risks removing valuable predictive information not contained in the retained predictor set. The primary issue with the input predictors in this case is not that many may not have any physical bearing on the predictand, but rather that each predictor represents a value at a different point of a continuous field, or a different property at the same point, and are thus necessarily highly correlated to one another. Furthermore, while one could conceivably extract features using field averages or some other pre-determined method, this may not be optimal. For example, it may be better to weight values closer to the forecast point more heavily, while still retaining some information from the far-field predictors. Given the uncertainty in optimally constructing features by manual means, it is more convenient and repeatable to instead extract features objectively. Though it has some limitations (e.g. Shlens 2014), principal components analysis (PCA; Ross et al. 2008; Pedregosa et al. 2011) is a robust and frequently utilized approach for dimensionality reduction. This creates a small set of uncorrelated predictors that explain the signal in the forecast data and gives insight into the regional modes of atmospheric variability as depicted in the GEFS/R model (explored in more depth in Chapter 4), while leaving the noise in lower-order principal components (PCs), acting in principle to both alleviate overfitting and manage computational requirements.

3.2.4 Machine Learning Algorithms and Sensitivity Experiments

The primary statistical algorithm used in this study is random forests (RFs; Breiman 2001). RFs are in essence an ensemble of *decision trees*, where traditionally each tree individually makes a deterministic prediction about the outcome of the predictand; the relative frequencies of each possible predictand outcome in the ensemble of trees are then used to make a probabilistic forecast. Much further detail on tree and RF construction and mechanics can be found in Chapter 3.2.5 as well as McGovern et al. (2017) and other sources. There are also several parameters which can be tuned to the particular forecast problem in order to maximize model performance. Four-fold cross-validation is used for model development in this study, whereby each model configuration examined is trained four times, once each on three-quarters of the training data, and then evaluated on the final withheld quarter. To avoid issues of sample independence and approximately mimic information that would be available in an

operational context, 974 *consecutive* initializations are used for each quarter of training data. All parameter settings and sensitivity experiments are evaluated in this framework. The set of RF parameters tuned is described in Chapter 3.2.5, and the results presented in Chapter 3.2.6.

In this study, there are a great deal of dynamical model data considered as input information on which the RF can base a prediction. A suite of sensitivity experiments are conducted, as summarized in Table 3.3, in order to investigate which aspects of forecast information contribute most to forecast skill. Experiments include exploring:

- Sensitivity to the inclusion of horizontal variations in atmospheric fields by varying the previously described predictor radius parameter R from 0 to 4.
- Sensitivity to the inclusion of additional upper-air atmospheric fields by comparing the inclusion and exclusion of two sets of fields as noted explicitly in Table 3.1. The first incorporates temperature, specific humidity, zonal and meridional winds at 850 and 500 hPa, and 850 hPa vertical velocity in the so-called Upper-Air Core predictor group, while an additional experiment further includes those same fields at 700 and 250 hPa.
- Sensitivity of predictor temporal resolution. Predictor density is three-hourly for Day 2 guidance and six-hourly for Day 3 guidance; models are additionally trained with predictors at twelve-hourly temporal density for both lead times and six-hourly temporal density for the Day 2 forecast model and compared against the control versions.
- Sensitivity to and type the extent of use of ensemble information, a question which has implications for how operational centers allocate their computational resources. Using forecast information from only the GEFS/R's control member in model training (CTRL) is compared with using the ensemble median from the full ensemble (MEDIAN), and then further with the use of the ensemble second-lowest and second-highest values for each atmospheric field in conjunction with the median (CNFDB) to evaluate the impact of this dimension of forecast information, following the findings of Herman and Schumacher (2016b), which found relatively little sensitivity in performance with respect to how ensemble information is used, but using the near-minimum, median, and near-maximum values outperformed using the mean and spread.
- Sensitivity to predictor pre-processing methodology. Models are trained with and without the aforementioned PCA pre-processing step, and an assessment of the effect of this pre-processing step on model skill is made by comparing the two.

TABLE 3.3. Summary of the models trained in this study, and the corresponding names designated to the models. 'X' indicates the process is performed or the information is used; a lack of one indicates the opposite. MEDIAN corresponds to the ensemble median, CTRL corresponds to the ensemble control member's fields, and CNFDB uses the median in addition to the second-from-lowest and second-from-highest member values for each field. Horizontal radius is listed in grid boxes from forecast point; timestep denotes the number of hours between GEFS/R forecast field predictors. Slashes indicate the first number applies to the Day 2 version of the model, while the latter number applies to the Day 3 version. Letters enclosed by parentheses indicate sub-versions of models, with one parameter changed to the value adjacent to the letter. Asterisks indicate a model applies only to Day 2, and not Day 3. Otherwise, models apply to all eight forecast regions and have both Day 2 and Day 3 versions. Those models with bolded names are incorporated into the weighted blend of the final model configuration.

Model Name	**CTL_NPCA**	**CTL_PCA**	UAC_PCA	UAF_PCA	CORE_CNFDB	CORE_CTRL	CORE_LSPACE	CORE_LTIME	**CTL_LR**
Algorithm	RF	RF	RF	RF	RF	RF	RF	RF	LR
PCA Pre-Processed		X	X	X	X				X
Uses Core Fields	X	X	X	X	X	X	X	X	X
Uses UAC Fields			X	X					
Uses UAE Fields				X					
Ensemble Information	MEDIAN	MEDIAN	MEDIAN	MEDIAN	CNFDB	CTRL	MEDIAN	MEDIAN	MEDIAN
Horizontal Radius	4	4	4	4	4	4	0 (a), 1 (b), 2 (c), 3 (d)	4	4
Timestep	3/6	3/6	3/6	3/6	3/6	3/6	3/6	12 (a), 6 (b*)	3/6

- The effect of region size on forecast skill, hypothesizing models trained for larger regions may exhibit higher skill due to more available training data. This is performed by aggregating the ROCK and SW regions into a new WEST one, combining the Southern Great Plains (SGP), Northern Great Plains (NGP), and Midwest (MDWST) regions into a CENTRAL region, and collecting SE and NE regions into a single EAST region, while leaving PCST—with its unique extreme precipitation climatology—unperturbed.
- Sensitivity of model performance as a function of model algorithm, specifically by comparing with logistic regression (LR), a common and comparatively simpler alternative to statistically deriving forecast probabilities. Further discussion of LR and other machine learning alternatives to the RF algorithm is included in the next section.

3.2.5 Algorithm Descriptions

3.2.5.1 RANDOM FORESTS

As noted in the main text, RFs are simply an ensemble of decision trees. Decision trees consist of a network of two types of nodes: decision nodes and leaf nodes. Decision nodes each have exactly two children, which may be either decision nodes or leaf nodes, with a binary split based on the numeric value of a single input predictor determining whether to traverse to the left or right child. A leaf node has no children and instead makes a categorical prediction of the outcome of the input example based on the leaf's relationship to its ancestor nodes. For a given forecast, one begins at a decision tree's root,

traversing through its children based on the relative value of the forecast's predictors to each decision node's threshold critical value for the predictor associated with the node. This process is repeated until a leaf node is reached; its value corresponding to the leaf becomes the tree's deterministic prediction.

Decision trees can be a powerful approach for a wide array of applications, but they also have several significant drawbacks. In particular, they are very prone to overfitting (e.g. Brodley and Utgoff 1995), fitting to the noise of the training data rather than just the underlying relationships. They also don't convey any information about forecast uncertainty, as would be the case in a probabilistic framework. RFs are used instead to alleviate these concerns by producing a probabilistic forecast in a way that can significantly decrease error from overfitting the supplied training error with only a slight increase to error from oversimplistic model assumptions, provided the trees are sufficiently uncorrelated. The difficulty then revolves around generating a large set (forest) of skillful decision trees that are not strongly correlated. The decision tree generating procedure described above is deterministic: a given set of training data will always produce the same decision tree. A forest of identical decision trees, of course, adds no value over using a single decision tree. Two additional proccesses—tree bagging and feature bagging—are employed to produce unique trees. Tree bagging produces unique trees through a straightforward bootstrapping procedure. Specifically, a forest of size B is formed from the n training examples by creating B samples of size n, with replacement, from the original training data, and running the decision tree algorithm on each sample. Overfitting due to correlated trees can still occur under this approach, particularly if a small subset of the original features are much more robust predictors of the verifying category than the rest (Breiman 2001; Murphy 2012). To overcome this problem, feature bagging is also employed, whereby only a random subset of the m original input predictors are considered at each decision node; the size of the random subset is denoted here as S; $1 \leq S \leq m$. This combination can result in a set of B largely uncorrelated trees, each of which is individually fairly skillful.

With any machine learning algorithm, there are numerous considerations in the actual model construction, which manifest themselves in tunable parameters. Compared with other machine learning algorithms, such as gradient boosting or support vector machines, RFs are often praised for their relative insensitivity to their parameters with respect to model performance, but it is nevertheless important to explore the parameter space in order to realize the full utility of the algorithm. The forest size B is perhaps the most obvious parameter. The general relationship between model performance and B is well known and consistent across all prediction problems; it starts quite low at very low B, initially

increases rapidly with increasing B, and then slowly asymptotes to some threshold performance limit as the relationships between input features have been fully explored by the forest and the inclusion of new trees becomes redundant. Larger forest sizes require more computational expense, so the goal is to select B such that it is small enough to be computationally tractable but large enough to be near the performance limit. Another parameter noted above is S, the number of features to consider at each node split. If this number is too small, model performance may suffer from only considering irrelevant or otherwise unpredictive features in the context of the node; if S is too large, performance will also suffer because of underdispersive trees producing an overfit forest solution. Another frequently explored parameter is the splitting criterion evaluation function. Most commonly used are either the Gini impurity or the information gain; past studies have shown that this choice is not important for many forecast problems. Information gain is used in this study; it can be expressed for a training set T, candidate splitting feature x_a and candidate split value v_a as:

$$IG(T, x_a, v_a) = H(T) - H(T|x_a < v_a) \tag{3.1}$$

where H(T) is the so-called entropy of a tree, defined for each of the K verifying categories, with each category i having forecast probability p_i, as:

$$H(T) = -\sum_{i=1}^{K} p_i \log_2 p_i \tag{3.2}$$

The chosen splitting feature and split value are selected among those considered which maximize Equation 3.1 (e.g. Quinlan 1986; Murphy 2012). However, there are two other parameters that have the most substantial influence on model performance. The first, denoted Z, is the minimum number of training examples required to split a node. Traditionally, RFs create a leaf only once a node is 'pure', that is, all the remaining training examples associated with that node have the same labels (event outcomes). In this way, each tree makes a categorical prediction of the predictand outcome, and probabilities are generated only in counting the proportion of trees in the forest making a particular forecast. However, this can make predictions from an individual tree very susceptible to the outcome of a particular historical case, and in some cases result in substantial overfitting. Instead, by increasing Z, an RF can be allowed to make 'impure' leaves; at these nodes, an individual tree makes a probabilistic prediction based on the proportion of remaining training examples exhibiting each event class rather than continuing to split based on the remaining training data. Making S too large, however, can result in underfitting—lumping data as indistinguishable when there are in fact underlying discernible

distinctions between remaining training examples with different labels. The last parameter, denoted P, is not actually an RF algorithm parameter at all. When PCA is performed, there is always a question about the number of components to retain. Though there are some heuristics (e.g. North et al. 1982), there is no definitive method to know *a priori* how many retained components P will produce the most skillful forecasts (Wilks 2011). If P is too small, valuable forecast data is discarded and predictive performance consequently suffers. However, if it is too large, the retained PCs eventually become essentially just noise, and the RF, by fitting to these predictor values in the training data, will yield an overfit model that does not generalize to unseen data. Experiments that will not be discussed herein revealed that using information gain to determine splits and letting B=1000 produced skill near that of an infinitely large forest, and skill was insensitive to modifications of these settings, including modest increases in the forest size beyond this point. However, the Z-S-P parameter space are explored for the models trained and those results are presented in the next section.

One final consideration concerns the handling of rare event scenarios. For rare event problems, one necessarily has many more examples of the common event class in comparison to the rare class, leaving the rare class somewhat underrepresented in the learning problem, and model fitting that is done with respect to the rare class is often too dependent on a small number of examples. An approach that has been applied with some success in past studies (e.g. Ahijevych et al. 2016) is to sample training data disproportionately from the rarer classes, so that the number of training example associated with each event class are approximately equal. A comparison between this so called "balanced" sampling and unmodified "unbalanced" sampling is also made and the results presented in Chapter 3.2.6.

3.2.5.2 Logistic Regression

One sensitivity experiment compares model performance as a function of the model algorithm by comparing skill of forecasts produced by RFs with those produced with logistic regression (LR). LR is in many senses a simpler model than an RF, since the structural form of the relationship between the predictors and the predictand is predefined before training. RFs, in contrast, make few assumptions about the relationships between the predictors and the predictand, allowing more diverse diagnoses of underlying relationships. However, this lack of assumptions can result in overfitting. As an application of the generalized linear model, LR assumes a linear predictors-predictand relationship via the logit function. In LR, a single regression equation, or K equations for a multicategory problem with K

categories, is computed to represent the probability of the outcome being category k given the set of input predictors $\mathbf{x}$. In particular, verifying probabilities are computed using the softmax function:

$$P(y=k|\mathbf{x}) = \frac{e^{\mathbf{x}^\mathbf{T}\mathbf{w}_k}}{\sum_{j=1}^{K} e^{\mathbf{x}^\mathbf{T}\mathbf{w}_k}} \tag{3.3}$$

In training a LR model, the goal is to determine the optimal weights $\mathbf{w}_k$ associated with each predictor in order to yield the most accurate predictions for each event class. As with RF models, LR can be prone to overfitting if unconstrained. For RFs, one aforementioned approach to alleviate this problem is to increase the above-termed Z parameter, which stops node splitting earlier on and makes the model less tailored to the specific training data supplied to it. Complexity in LR can be thought of as being analogously represented by large weights, or regression coefficients. In order to ensure better generalizability of the trained regression equations, it is often good practice to penalize large weights through a process known as regularization. When this is done, the computation of optimal weights can be represented as a minimization problem with two terms. For 1) a matrix Y with binary elements that are non-zero if and only if training example i has associated verifying category k and 2) a model outputting a probability matrix P for each training example and category, the multinomial loss J to be minimized can be computed as:

$$J(\mathbf{Y},\mathbf{P}(\mathbf{w})) = \frac{1}{2}\mathbf{w}^\mathbf{T}\mathbf{w} - \frac{1}{CN}\sum_{i=1}^{N}\sum_{k=1}^{K}\mathbf{Y}_{i,k}\,log(\mathbf{P}_{i,k}) \tag{3.4}$$

where C represents the extent of regularization, with smaller values indicating that large weights are penalized more than with larger values of C. Alternative approaches to regularization exist (e.g. Pedregosa et al. 2011; Murphy 2012), and are explored to some degree in sensitivity experiments of Chapter 3.2.6.

3.2.5.3 Computational Considerations

Other machine learning algorithms do not scale well to the high dimensionality of the forecast problem explored here. While time to train a model is not of primary concern for operational forecasting since it is performed only once (or periodically) offline, there are nevertheless some practical considerations; models that take months or longer to train would be unlikely to be realistic choices, for example. The "online" forecasting component, that is, the time required to take a new forecast, input it into a trained model and receive a forecast, *is* of operational concern, but all of the forecast techniques considered here can produce forecasts in a matter of minutes, and the small differences are not considered to be of practical concern. Using the random forest classification heuristic of considering

the square root of the total number of features at each node split (Geurts et al. 2006), the computational complexity of training an RF of size B from N training examples with F features (N > F) may be expressed as $O(B\sqrt{F}N\log(N))$, and may be readily parallelized across trees or within trees. Some algorithms are quadratic or even cubic (e.g. Cortes and Vapnik 1995) in the number of training examples, and do not parallelize as readily. LR is linear in the number of training examples, but requires matrix multiplication, a process that yields a computational complexity of $O(NF^2)$. PCA pre-processing, and dimensionality reduction more generally, acts both to make learning algorithms more computationally tractable and also reducing overfitting by alleviating the so-called "curse of dimensionality".

3.2.6 Parameter Tuning

RF model parameters were tuned for each region and lead time separately through the 4-fold cross-validation procedure employed throughout the study. Overall, the optimal parameters were found not to vary with the two different lead times, but did vary for two of the parameters as a function of forecast region, at least to an extent; the full results appear in Table 3.4. For the S parameter—the number of predictors considered for each node split, the default heuristic of the square root of the total number of features was found to maximize RPSS for all regions and lead times. In all instances where both were tested, unbalanced sampling from the event classes in proportion to their true observed frequencies outperformed balanced equal sampling from each event class, in contrast to Ahijevych et al. (2016) and others; the finding appeared to be attributable to biased probabilities produced from the balanced sampling technique. For the Z parameter, the minimum number of remaining training examples in an impure parameter subspace required to perform a further node split, was generally found to be around 120. Lesser values maximized skill in the western regions, with values of 30 maximizing skill in the SW and ROCK regions, and Z=4 producing the best skill over PCST. A couple of the larger regions of the east, SE and MDWST, maximized RPSS with a value of 240, although the sensitivity between Z=120 and Z=240 was small for all regions. For P in the CTL_PCA models, skill was generally maximized with P=30, that is, retaining the 30 PCs which explain the most variance of the entire GEFS/R predictor set. For most regions, there was very limited sensitivity in the P=30–40 interval—although there was larger sensitivity outside this interval—and P=40 was found to produce slightly better skill in the NGP region. The PCST region was again the main exception, where P=60 was found to maximize cross-validation RPSS.

LR model parameters were tuned using an identical framework to ascertain the type of regularization, either based on a L1 norm which penalizes non-zero weights, or L2 norm—described in Chapter

TABLE 3.4. Optimal RF parameters obtained in cross-validation for the Z-S-P parameter space. SQRT indicates the square root of the total number of predictors; symbols are otherwise as described in the chapter text. Evaluated values were 1, 2, 4, 8, 16, 30, 60, 120, 240, 480 for Z, and 20, 25, 30, 40, 50, 60, 70, 80, 90, 100 for P.

Region	S Parameter	Z Parameter	P Parameter
ROCK	SQRT	30	30
NGP	SQRT	120	40
MDWST	SQRT	240	30
NE	SQRT	120	30
PCST	SQRT	4	60
SW	SQRT	30	30
SGP	SQRT	120	30
SE	SQRT	240	30

3.2.5—which penalizes large magnitude weights. L2 regularization was consistently found to produce superior results, perhaps because the number of retained PCs was already taken from the P parameter in the RF experiments, acting to nullify many potential non-zero weights of higher numbered PCs. Unlike the RF experiments, there were occasionally some large differences in the obtained optimal regularization parameter value C between lead times within the same region. Generally, models performed better with more regularized solutions, but there were some notable exceptions, with the Day 2 NGP model and Day 3 NE model obtaining optimal C parameter values on the other end of the spectrum.

3.2.7 Model Evaluation

Based on the parameter tuning and sensitivity experiment results, final model configurations are selected. The final model is run over a completely withheld 4-year evaluation period spanning September 2013–August 2017. The forecasts gerenated from the final model are compared with those from the full ensemble of raw GEFS/R QPFs, as well as the full 50-member ECMWF global ensemble, accessed from TIGGE (Molteni et al. 1996; Bougeault et al. 2010). The comparison with the former provides an assessment of what improvement, if any, these models yield compared with the raw guidance from which their forecasts are derived when evaluated in a real-time setting. The latter, meanwhile, provides

TABLE 3.5. Optimal LR parameters obtained in cross-validation for the C parameter and regularization type for all lead times and regions. Evaluated for C were 0.0001, 0.0008, 0.0060, 0.0464, 0.359, 2.78, 21.54, 167.8, 1291, and 10000.

Region	Regularization	C Parameter, Day 2	C Parameter, Day 3
ROCK	L2	0.0001	0.0001
NGP	L2	10000	0.0008
MDWST	L2	0.0001	0.0464
NE	L2	2.78	10000
PCST	L2	0.0001	0.0001
SW	L2	0.359	0.0001
SGP	L2	0.0008	0.0464
SE	L2	21.54	0.0008

an assessment for how these forecasts compare with state-of-the-science operational ensemble guidance available at these lead times. To make these comparisons, the QPF from each ensemble member of the two ensembles is regridded onto the ~4.75km Stage IV HRAP grid on which the Atlas thresholds lie using a first-order conservative scheme (Ramshaw 1985). These regridded QPFs are then compared with the 1-year and 10-year ARI thresholds to create deterministic exceedance forecasts with respect to the two thresholds for each ensemble member. These binary grids are then upscaled to the GEFS/R grid using the same procedure as the verification upscaling: any exceedance in the downscaled grid corresponds to an exceedance at the nearest GEFS/R point in the upscaled grid. Since the predictand categories are necessarily mutually exclusive, the 1-year ARI exceedance grids are modified so that any member forecasting a 10-year ARI exceedance at a point is not forecasting a between 1-and-10-year exceedance at that same point and time period. The prevailing operational method of generating forecast probabilities from a dynamical ensemble—democratic voting, whereby the fraction of ensemble members forecasting the event is used as the forecast probability (e.g. Buizza et al. 1999; Eckel 2003)—is applied to each ensemble to generate the exceedance probabilities for the reference forecasts.

Skill, both in the final assessment of model performance as well as in all aforementioned sensitivity experiments, is quantified by means of the Rank Probability Skill Score (RPSS) with a climatological reference:

$$RPSS = 1.0 - \frac{\Sigma_{d=1}^{D}(\Sigma_{p=1}^{P}(\Sigma_{m=1}^{K}(\Sigma_{j=1}^{m}P_{jpd} - O_{jpd})^2))}{\Sigma_{d=1}^{D}(\Sigma_{p=1}^{P}(\Sigma_{m=1}^{K}(\Sigma_{j=1}^{m}P_{clim_j} - O_{jpd})^2))} \tag{3.5}$$

with D forecast days; P forecast points; K predictand categories; P_{jpd} and O_{jpd} corresponding respectively to the forecast probability and observance of predictand category j on day d and at point p; and P_{clim} corresponding to the climatological frequency of occurrence, as defined by the respective ARIs of the predictand. A score of 1.0 indicates a perfect forecast, and a score of 0.0 indicates model performance equivalent to forecasting climatology. Final assessment also includes analysis of reliability, both subjectively through reliability diagrams, and quantitatively via the Murphy (1973) decomposition of the Brier score (BS) for category j^*:

$$BS_{j^*} = \Sigma_{n=1}^{N}(P_{N\,j^*} - O_{N\,j^*})^2 = \frac{1}{N}\Sigma_{c=1}^{C} N_{c\,j^*}(P_{c\,j^*} - \overline{O_{c\,j^*}})^2 - \frac{1}{N}\Sigma_{c=1}^{C} N_{c\,j^*}(\overline{O_{j^*}} - \overline{O_{c\,j^*}})^2 + \overline{O_{j^*}}(1-\overline{O_{j^*}}) \quad (3.6)$$

where there are $N = DP$ total forecasts, broken into C discrete probability bins with N_c forecasts being issued for each bin c. $\overline{O_{j^*}}$ denotes the climatological (based on the period of record) frequency of observing event category j^* and $\overline{O_{c\,j^*}}$ denotes the proportion of forecasts in probability bin c observing event category j^*, where j^* is the aggregation of event categories of at least j in the RPSS framework. $\overline{O_{j^*}}(1-\overline{O_{j^*}})$, the so-called "uncertainty" term, also represents the BS of a climatological forecast. Converting to a Brier skill score (BSS) framework by dividing out by this term:

$$BSS_{j^*} = 1.0 - \frac{BS_{j^*}}{BS_{clim_{j^*}}} = \underbrace{\frac{\frac{1}{N}\sum_{c=1}^{C} N_{c\,j^*}(\overline{O_{j^*}} - \overline{O_{c\,j^*}})^2}{\overline{O_{j^*}}(1-\overline{O_{j^*}})}}_{\text{"Resolution"}} - \underbrace{\frac{\frac{1}{N}\sum_{c=1}^{C} N_{c\,j^*}(P_{c\,j^*} - \overline{O_{c\,j^*}})^2}{\overline{O_{j^*}}(1-\overline{O_{j^*}})}}_{\text{"Reliability"}} \quad (3.7)$$

This analysis is conducted for both the 1- and 10-year thresholds.

Skill calculations and comparisons are made for the host of sensitivity experiments and for each region, lead time, and model configuration. For each comparison, statistical significance is assessed by bootstrapping to obtain identical sets of cases for each of the two forecast sets being compared. Skill scores are derived from the subsample of each forecast set, and a skill difference is computed. This process is repeated 1000 times to generate a distribution of skill differences, and statistical significance is ascertained with respect to whether the 0.5^{th} and 99.5^{th} percentile skill score difference values from the bootstrap trials overlap zero. This 99% confidence bound is used in contrast to 90% or 95% bounds to compensate for concerns arising from conducting statistical significance analysis on numerous different comparisons. While some uncertainty analysis is included in the figures presented, much of the statistical significance difference results discussed in-text are omitted for the sake of concision.

3.3 Results: Sensitivity Experiments

Examining forecast skill as a function of time step between atmospheric field predictors (i.e. the CORE_LTIME models of Table 3.3; Fig. 3.3a), two striking findings concern 1) the large variations in forecast skill across regions and 2) the evidently low sensitivity of forecast skill to time step length within any given region. For the 3-hour time step, predictors are gathered from a total of 9 forecast times; with the 6-hour step, 5 forecast times are used; and with the 12-hour time step, a total of 3 forecast times are used. The 12-hour time step therefore has one-third the total number of predictors as the model with the 3-hour time step, but still yields nearly identical forecast skill results. In most regions and forecast periods, there is a slight degradation in performance going from the 6- to 12-hour time step, but the difference is not generally statistically significant by a 99% bootstrap skill score difference test (not shown). The one exception to this is in the PCST region, which has much higher skill overall than the other regions for both forecast periods, and exhibits somewhat higher sensitivity to the predictor time step than the other regions, particularly in going from 6-hours to 12-hours, with RPSS differences of approximately 6%.

Similar to the temporal resolution findings, there is a general lack of sensitivity as a function of predictor spatial extent (Fig. 3.3b). This finding comes in stark contrast to that of Herman and Schumacher (2016b), which found great sensitivity of predictor spatial extent in forecasting airport flight rule conditions. Albeit weak, a slight improvement in skill for most forecast period, region combinations can be noted with increasing predictor radius, often to the extent that the skill difference between 0 and 4 grid box radii is statistically significant (not shown). Two regions in particular, the NE and PCST, exhibit by far the most sensitivity to predictor spatial extent, with differences of roughly 0.02 observed over the evaluated interval. Also of note is that a radius of 4 grid boxes—the highest number evaluated—did not always yield the best performance results; most notably, the Day 2 model for the NE region maximized skill at a radius of 2, with a slight deterioration of forecast skill with increasing radius thereafter. In those regions where the GEFS/R cannot explicitly resolve the processes responsible for producing extreme precipitation, the RF is ultimately making forecasts more on environmental factors; these do not vary drastically in time or space, and thus a single number or small set of numbers at or immediately surrounding the forecast point are sufficient to characterize the basic properties of the environment. This is all that the RF is really using for much of its predictions (discussed more in Chapter 4). However, in regions impacted more readily by larger scale systems where the dynamical model can more directly

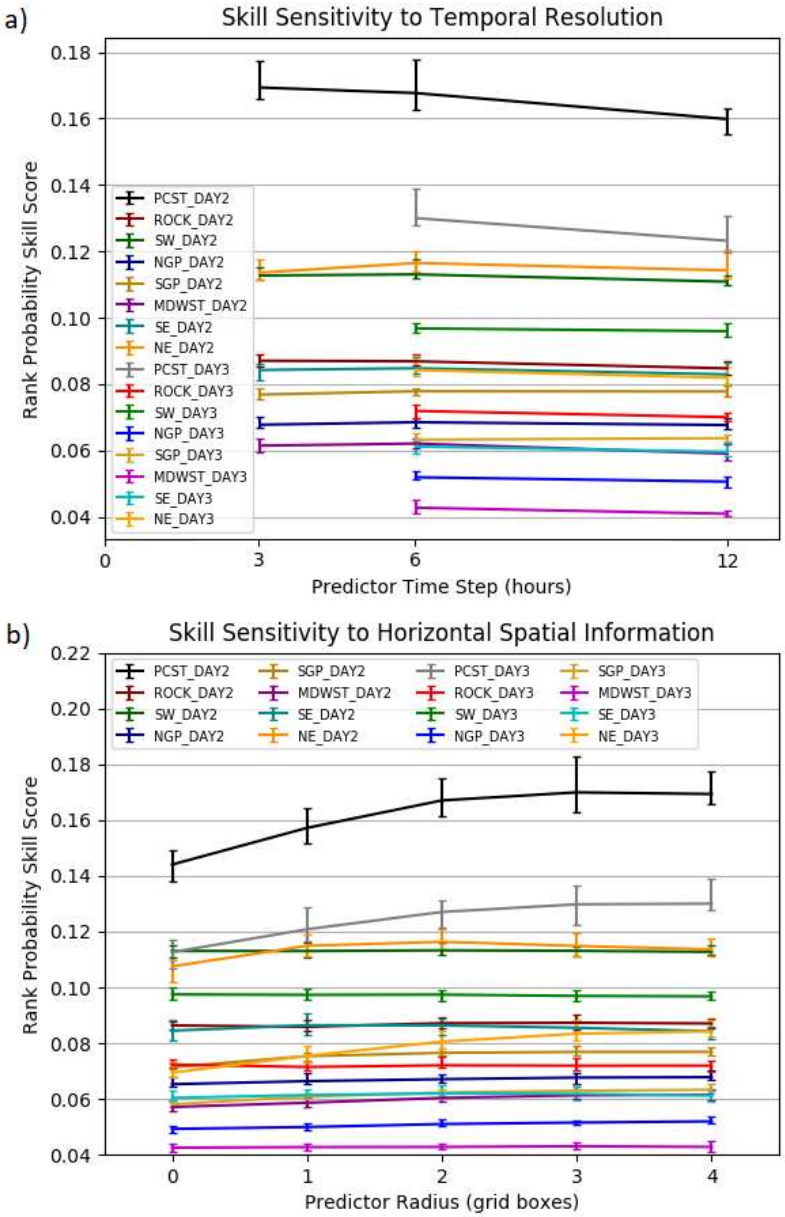


FIG. 3.3. Sensitivity experiment RPSS results for (a) the CORE_LTIME models, as a function of the timestep between incorporation of new atmospheric field forecast values, and (b) the CORE_LSPACE models, as a function of the radius of predictor information incorporated, each including both Day 2 and Day 3 versions of the model and for each region studied. Lines correspond to a particular day, region pair as indicated in the respective panel legends. Error bars in both panels correspond to 90% confidence bounds obtained by bootstrapping.

simulate the precipitation processes such as PCST and the NE, the spatial variations in atmospheric fields carry more signal rather than noise and thus contribute more predictive value.

Like varying spatial and temporal density, there is relatively little sensitivity to the inclusion of more atmospheric fields (Fig. 3.4a). Slight but consistent improvement is observed in adding the core upper-air fields as predictors, but adding further levels beyond the core group was found to not improve predictive skill, and actually resulted in a decrease in skill for the PCST, NE, ROCK, and SE regions—those which are most affected by larger scale precipitation systems. Though still rather small, somewhat more distinct sensitivity to type of ensemble information included (Fig. 3.4b) can be seen here across all regions, with improvements seen using predictor information from the GEFS/R ensemble median versus using only the control member, and slight further improvement using the ensemble second-from-minimum and second-from-maximum in addition to the ensemble median. The largest differences in magnitude are again for the PCST region, but in this experiment, clear and statistically significant (not shown) improvements are also seen for low skill, convectively active regions such as MDWST.

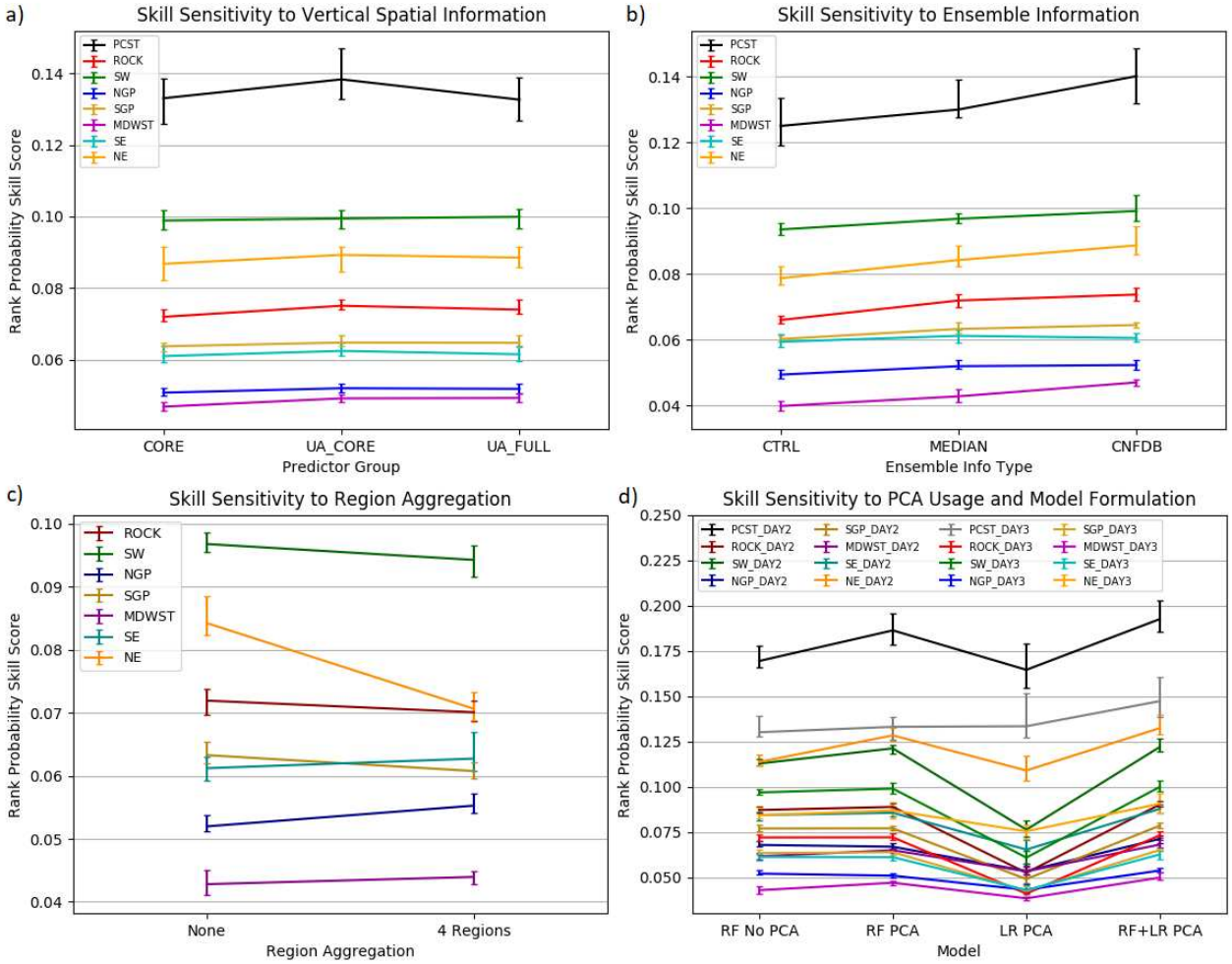


FIG. 3.4. Sensitivity experiment RPSS results. Panel (a) as a function of the atmospheric fields included as input to the RF algorithm, for Day 3 forecast and broken out by region. From left to right, the columns correspond to results using: 1) just the 'Core' atmospheric field group, 2) both the 'Core' and 'Upper Air Core' groups, 3) the 'Core', 'Upper Air Core', and 'Upper Air Extra' groups. For more information on which fields are included in each predictor group, consult Table 3.1. Panel (b) as a function of the type of GEFS/R information used as input predictors to the RF algorithm, for Day 3 forecasts and broken out by region. From left to right, the columns correspond to results using: 1) just the forecast fields from the GEFS/R control member, 2) the ensemble median forecast values from the full ensemble, 3) the ensemble median, 2nd-from-minimum, and 2nd-from-maximum forecast values from the full ensemble. Panel (c) as a function of region aggregation, with the left column using the eight regions depicted in Figure 2.11, and the right column using training data which aggregates data from seven of the eight original regions into three regions, as described in the text. Panel (d) as a function of model algorithm for different forecast days and regions as indicated in the figure legend. From left to right, columns correspond to results of the CTL_NPCA model, CTL_PCA model, CTL_LR model, and a weighted combination of models as described in the paper text. For all panels, error bars correspond to 90% confidence bounds obtained by bootstrapping.

Aggregating regions (Fig. 3.4c) results in a slight degradation in forecast skill. In principle, it is possible for a decision tree to automatically forecast for specific regions by splitting first on the latitude and longitude predictors, and then further partitioning based on meteorological variables thereafter. However, these findings demonstrate that there is some—albeit limited—utility in manually partitioning training data with distinct hydrometeorological relationships, rather than relying on the machine learning algorithm to discern the distinction automatically. Comparing the impact of applying PCA pre-processing to the RF (Fig. 3.4d, leftmost two columns), performing PCA tends to either improve performance, as is the case for the PCST, NE, SW, and MDWST regions, or make little difference, as seen in the ROCK, NGP, SGP, and SE regions. The positive differences tend to be larger in magnitude, both in relative and absolute senses, for Day 2 model versions compared with Day 3. Forecasts produced through LR tend to be substantially worse than those generated by RFs (Fig. 3.4d, center columns). However, the exact magnitude to which this is the case varies by region; substantial differences in skill are seen between RF and LR forecasts for the SW, ROCK, and SGP regions, while there is almost no skill difference between the Day 3 forecasts in the PCST region. This may suggest the linear assumptions inherent to the LR algorithm perform better in larger scale systems than in the more convectively active ones in which the responsible processes are highly nonlinear, but this causality is not entirely clear. Finally, a weighted average of RF and LR forecasts outperforms its component members for all regions and forecast periods. The extent of overperformance is strongly tied to the skill difference between the

RF and LR models; when the skill difference is small, the value of the weighted average is comparatively large to when the RF performs much better than LR (cf. Fig. 3.4d PCST and SW lines). Since these weighted averages performed the best in cross-validation, a weighted average using each of the CTL_NPCA, CTL_PCA, and CTL_LR models was chosen for the final model configuration.

3.4 Results: Final Model Performance

For both the final ML models and the forecasts from the raw QPFs of both the GEFS/R and ECMWF (Fig. 3.5), a usually statistically significant deterioration in forecast skill from Day 2 to Day 3 is evident in each CONUS region over the four year test period. Forecast skill is significantly higher in regions with extreme precipitation associated partially or primarily with synoptic scale precipitation episodes, such as PCST, SW, and ROCK, rather than smaller-scale convective systems that characterize extreme precipitation as in the NGP, SGP, and MDWST regions. At an extreme, the NGP and SGP GEFS/R raw QPFs exhibit no skill in predicting ARI exceedances at these lead times. Especially for the ML models, the bigger Day 2 vs. Day 3 skill differences are also seen where the skill is higher, again suggesting the direct forecasting of the precipitation as opposed to forecasts more reflecting the forecast environment, either dynamically via parameterized convection in the case of raw QPFs, or directly in the case of the ML model forecasts. Furthermore, the ML models exhibit a larger skill deterioration between Days 2 and 3 than either of the raw ensemble forecast sets.

Comparing the forecast systems, the ECMWF forecasts consistently and statistically significantly outperform the GEFS/R forecasts at all lead times except in the SE region (Fig. 3.5). Encouragingly, the ML model forecasts are statistically significantly more skillful for all eight regions and both lead times compared with the GEFS/R forecasts from which they are based. The post-processing is thus clearly accomplishing its purpose of improving forecast skill. But it is also apparent that the GEFS/R is not a state of the science model for extreme QPF prediction given its lower skill compared with the ECMWF. The real test of the ML model then is how it compares with current best operational guidance for these lead times, represented here with the ECMWF ensemble. The comparison (Fig. 3.5) is generally quite favorable, with the Day 3 ML forecasts outperforming even the Day 2 ECMWF forecasts across all regions except ROCK and PCST. In the non-western regions, the extent of overperformance is quite considerable when comparing equal lead times, with skill score improvements of factors of two to three seen in many comparisons. In the ROCK and PCST regions, the ML and ECMWF forecasts performed about equally at Day 2, and ECMWF performed slightly better at Day 3. Overall, the ML models demonstrated

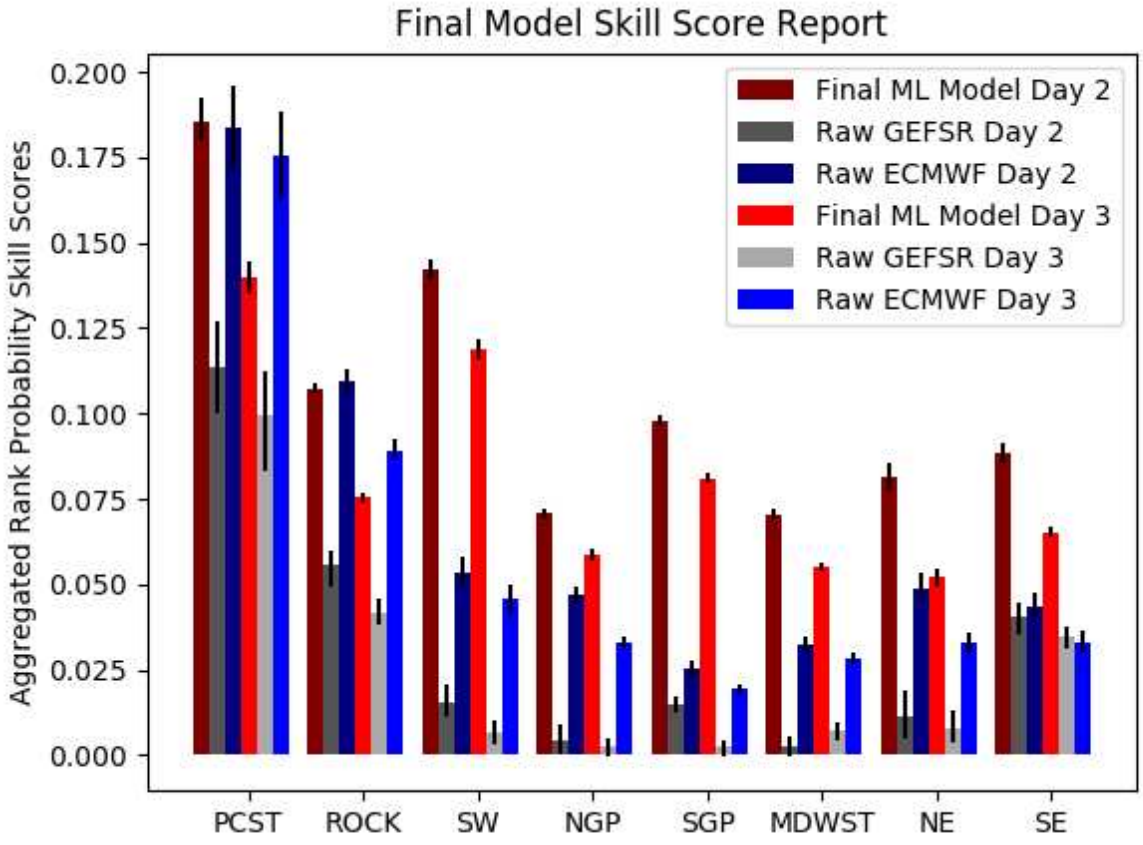


FIG. 3.5. Final RPSS results obtained over the four year test period spanning September 2013–August 2017, broken out by region. Red bars correspond to the results of the final forecast models trained in this study, while gray bars depict results from the raw GEFS/R QPF probabilities derived from the full ensemble. Dark bars illustrate Day 2 performance results, while lighter colors show results for Day 3. Error bars correspond to 90% confidence bounds obtained by bootstrapping.

ability to consistently outperform current operational model guidance, especially in convectively active regions where there is no operational guidance that can dynamically resolve the physical processes producing extreme precipitation at these lead times.

Reliability diagrams of Day 2 raw GEFS/R and ECMWF forecasts (Fig. 3.6) reveal highly overconfident probabilistic exceedance forecasts for all regions, both severity levels, and both ensembles as evidenced by the shallow slope relative to the one-to-one line in each panel. The raw GEFS/R forecasts (Fig. 3.6a,b) are relatively sharp, with more than 0.01% of forecasts falling into each probability bin above 10%, and a vast majority of zero probability forecasts (not shown). For all regions, there are cases where every ensemble member has simultaneously predicted a 1-year exceedance (Fig. 3.6a),

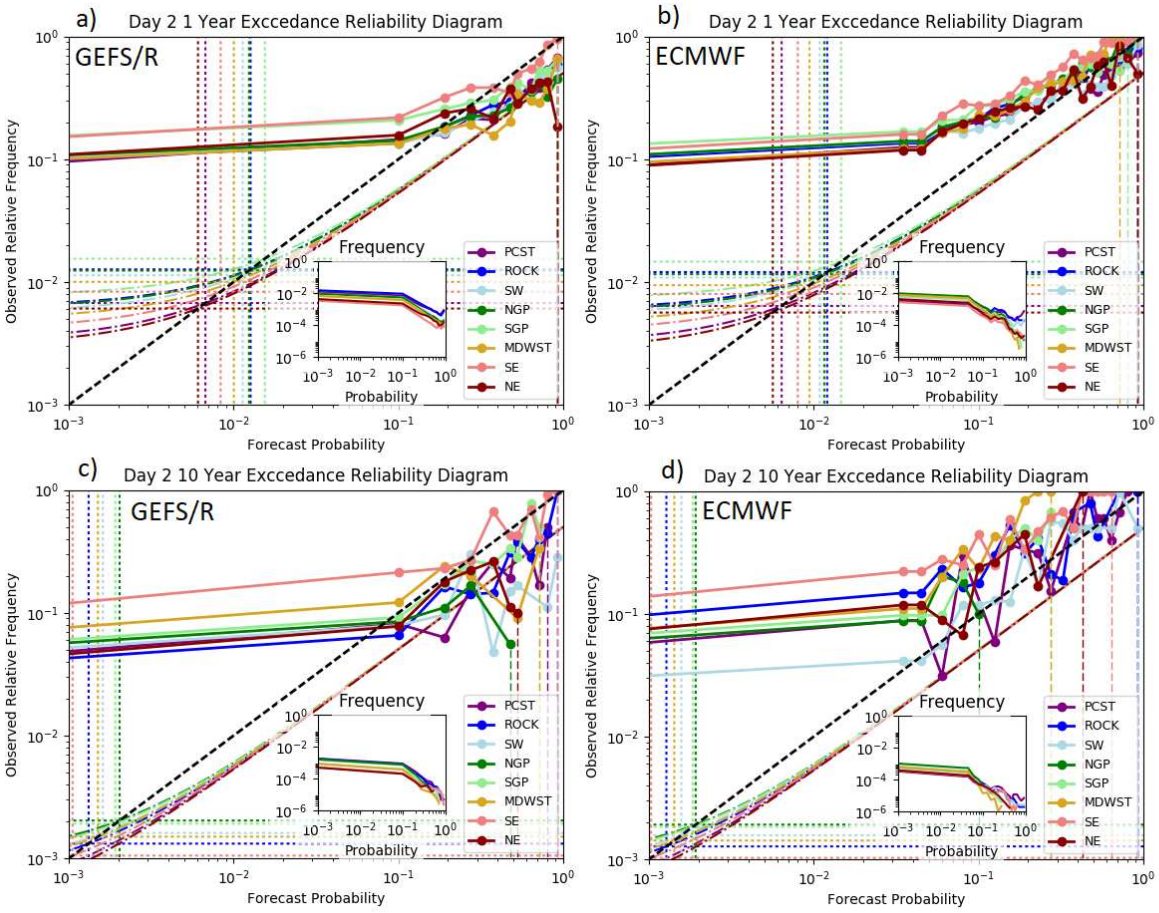


FIG. 3.6. Reliability diagrams for Day 2 forecasts generated from raw QPFs of the full GEFS/R and ECMWF ensembles. Colored opaque lines with circular points indicate observed relative frequency as a function of forecast probability; the dashed black line is the one-to-one line, indicating perfect reliability. Colors correspond to the performance of the forecasts over different regions, as indicated in the legend in the lower-right of each panel. Inset panels indicate the total proportion of forecasts falling in each forecast probability bin, using the logarithmic scale on the left hand side of each panel; lines are again colored by region in accordance with the legend. Panel (a) shows 1-year exceedance forecast from GEFS/R, (b) to 1-year exceedance forecasts from ECMWF, (c) to 10-year from GEFS/R, and panel (d) to 10-year from ECMWF. All axes are logarithmic as labeled. Colored dotted lines indicate the climatological event probability for each region for the ARI level of the corresponding panel, while the dash-dotted lines indicate no skill lines for the color-corresponding region. The curves continue off the left end of each panel towards the ORF of forecasts in the zero forecast probability bin.

but the same is not true for 10-year exceedance predictions in the northeastern regions: NE, NGP, and MDWST (Fig. 3.6b). The ECMWF (Fig. 3.6b,d) is also overconfident, but we see that it is also negatively biased for all cases. Its degree of overconfidence is dampened compared with the GEFS/R, and it is not as sharp, with fewer very occurrences of very high forecast probabilities except in the westernmost

regions of ROCK and PCST (Fig. 3.6b, inset panels). With 50 members rather than 11, there is also substantially more resolution across the probability spectrum in the ECMWF forecasts. By the very nature of how these forecasts are generated, quite a bit of sharpness is inherent at the cost of reliability, since it is not possible for probabilities near the climatological event frequency to be issued for either raw ensemble, but particularly for the GEFS/R.

The Day 2 reliability diagrams for 1-year exceedance forecasts from the different components of the final model—CTL_NPCA, CTL_PCA, and CTL_LR—are shown in Figure 3.7. The CTL_NPCA (Fig. 3.7a) shows markedly different characteristics than either of the raw ensembles. In particular, all of the regions exhibit an underconfidence signal, with low probability events below about 2% for 1-year events (Fig. 3.7a) occurring with observed relative frequencies below the forecast probabilities. The relative event frequencies are conversely appreciably higher than the forecast probabilities would indicate for probabilities above 5%. Among the regions, the PCST probabilities are the most negatively biased, while NE probabilities are the most positively biased. Overall, reliability is much better than for either raw ensemble, but this comes at the expense of sharpness. Less than 1 in 10,000 forecasts are above about 20% for (Fig. 3.7a, inset panels), and maximum probabilities are in the 30–80% range depending on the lead time and region, compared with 100% for all lead times and regions in the raw ensembles. The CTL_PCA model (Fig. 3.7b) exhibits very similar reliability characteristics to the CTL_NPCA model, including the underconfidence, reduced sharpness compared with the raw ensembles, and different regional probability bias characteristics. It tends to be more negatively biased than CTL_NPCA at low and high probabilities (cf. Fig. 3.7a,b), correctly so at high probabilities and undesirably so at low ones. The CTL_LR model (Fig. 3.7c) exhibits some similarities and some differences with the RF-based models. PCST forecasts are consistently the most negatively biased, followed by ROCK and the SE, with NE region forecasts being the least negatively biased. However, unlike the RF-based forecasts, the LR model issues more high probabilities; for example, forecasts in the highest probability bin were issued for most regions (Fig. 3.7c). At the highest probabilities, the forecasts revert to being positively biased, as they are for events with probabilities issued in the 0.01–1% range. At very low probabilities , LR-based forecasts are substantially more negatively biased than for RF-based forecasts, leading to considerable overconfidence overall when considering that the vast majority of forecasts issued occur on this low probability end of the spectrum. While LR (and regression in general) is effective at removing bias in

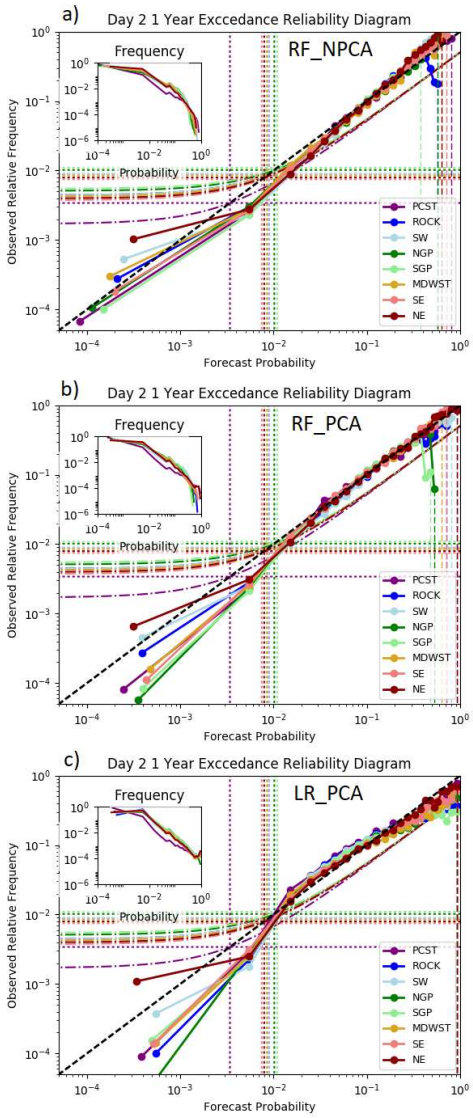


FIG. 3.7. Reliability diagrams for Day 2 forecasts of 1-year ARI exceedances for different statistical algorithms. Panel characteristics as in Figure 7, except note that axes have been modified to include more of the low probability tail due to increased resolution in the plotted forecast sets. Panel (a) corresponds to forecasts from the CTL_NPCA model, panel (b) to the CTL_PCA model, and panel (c) to the CTL_LR model. Bin right edges correspond to forecast probabilities of 0, 1e-10, 1e-7, 1e-4, 1e-3, 0.01, 0.02, 0.03, 0.04, 0.05, 0.07, 0.09, 0.11, 0.14, 0.17, 0.21, 0.25, 0.30, 0.35, 0.40, 0.45, 0.50, 0.55, 0.60, 0.675, 0.75, 0.85, and 1.0, except that first five probability bins have been aggregated into a single frequency-weighted probability bin for plotting on the figure.

a global sense, since a single regression equation must necessarily apply globally to all forecasts, it inherently cannot perform more localized, context-depend forms of bias correction, leading to forecast probability-dependent model biases.

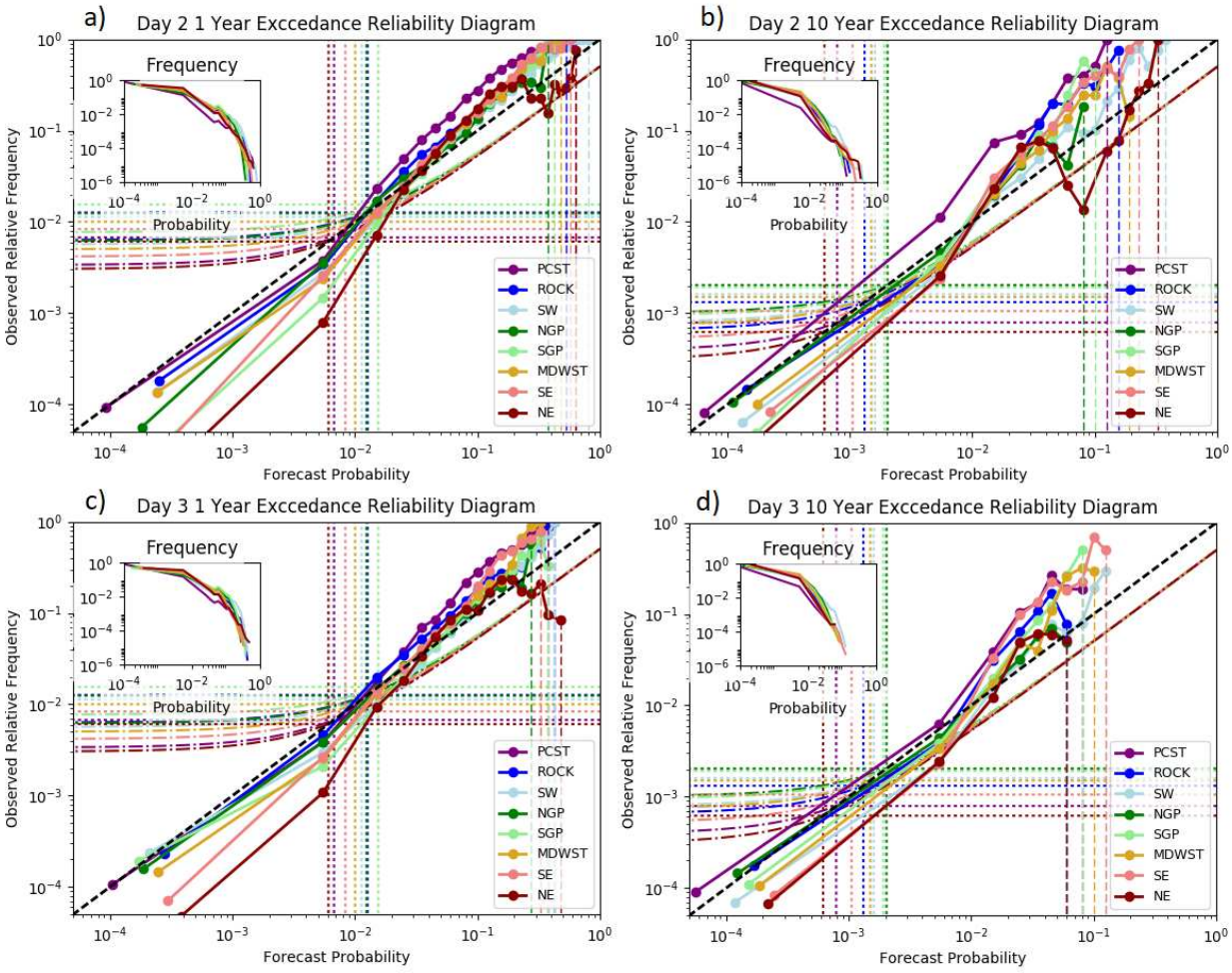


FIG. 3.8. Reliability diagrams for the final forecast model, with panel attributes as in Figure 8. Panel (a) shows Day 2 forecast results for 1-year ARI exceedance forecasts, (b) to Day 2 10-year ARI exceedance forecasts, (c) to Day 3 1-year exceedance forecasts, and panel (d) to Day 3 10-year ARI exceedance forecasts.

The final ML model reliability (Fig. 3.8) unsurprisingly reflects a blend of the component members, retaining some of the underconfidence of the RF-based models while adding a bit of sharpness from the CTL_LR model in regions where it verified skillfully enough in cross-validation (e.g. PCST, Fig. 3.4d) to garner much weight. The probability distribution for 1-year exceedance events is not markedly different between the Day 2 and Day 3 forecasts (cf. Fig. 3.8a,c), but the relatively higher probabilities issued for 10-year exceedances in Day 2 do not occur at the Day 3 lead times (cf. Fig. 3.8b,d). This is consistent with increasing confidence in very extreme events with decreasing lead time—something seen very pronounced in the final model, but to a much lesser extent in the raw ensemble forecasts.

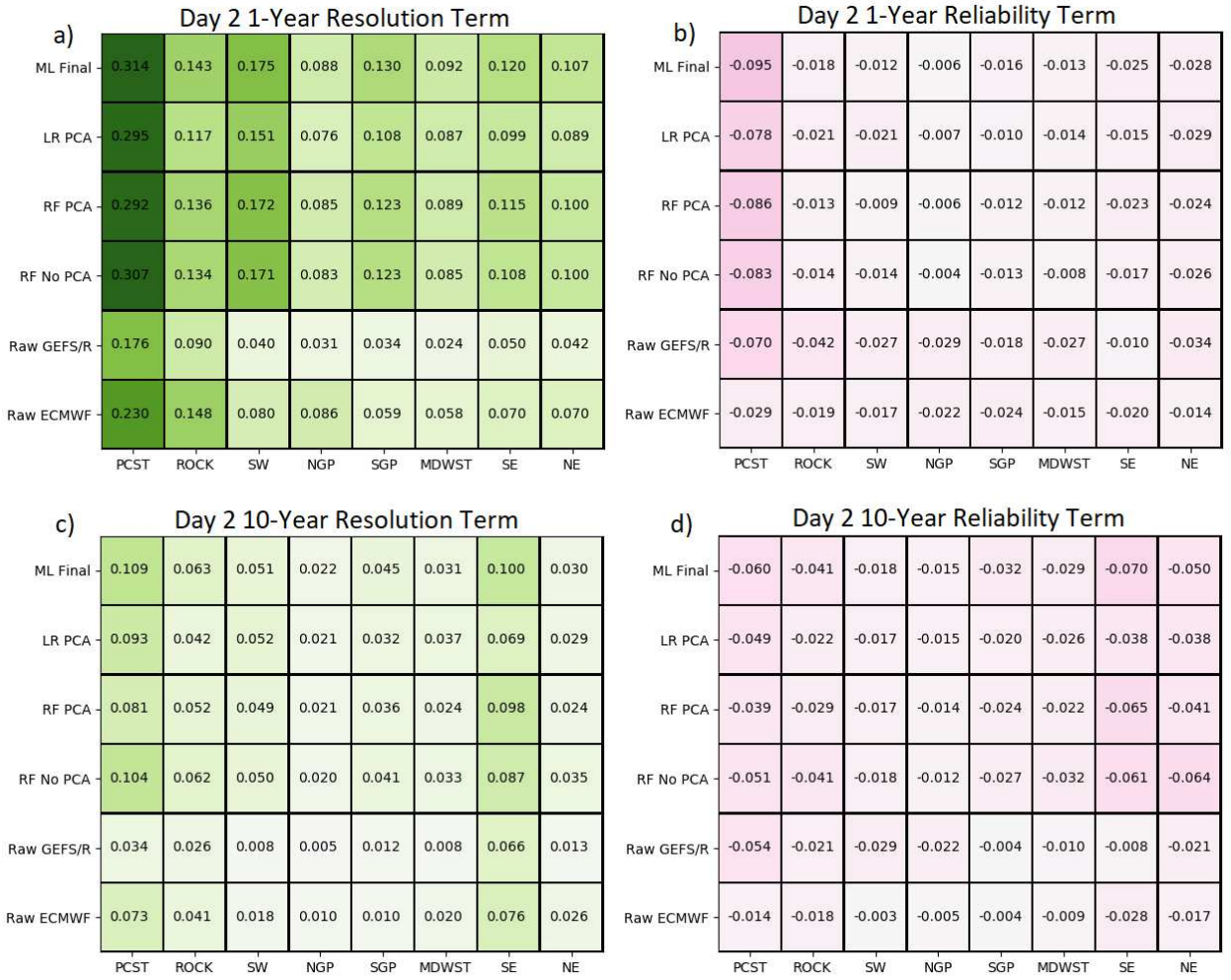


FIG. 3.9. Modified Murphy (1973) decomposition results, following equation 3.7 in text. Panel (a) depicts the equation 3.7 "resolution" term for all models and regions for Day 2 forecasts at the 1-year severity level, panel (b) depicts the "reliability" term results for the same forecasts and severity level. Panels (c) and (d) are analogous to panels (a) and (b), but for 10-year ARI exceedance forecasts. Numeric values indicate the value of the corresponding term of the table, as indicated by the model label (row) and region (column).

The relationship between the reliability analysis and skill via the Brier score decomposition (Murphy 1973) quantitatively solidifies many of the general observations discerned by inspection of the reliability diagrams. Though sharper than competing forecasts, the raw GEFS/R forecasts consistently exhibit the worst resolution component contribution to forecast skill for all regions and severity levels, both for Day 2 forecasts (Fig. 3.9a,c) and Day 3 forecasts (Fig. 3.10a,c) due to an inability to actually distinguish events from non-events by resolving the responsible physical mechanisms. The final ML models exhibit better resolution term skill contributions than the ECMWF ensemble forecasts, with the

exception of the ROCK and NGP regions for 1-year events (Fig. 3.9a, 3.10a). Between the component models, resolution term skill tended to best for CTL_NPCA forecasts over the test period, particularly at the 10-year severity level (e.g. Fig. 3.9c) but the extent of the difference tended to be relatively small and there were numerous instances where PCA-based models exhibited more resolution. The weighted average consistently exhibited higher resolution than any of the component members. With respect to the reliability contribution to skill (Fig. 3.9b,d Day 2; Fig. 3.10b,d Day 3), ECMWF forecasts were, perhaps surprisingly given the lack of explicit calibration, the most reliable forecast set for all regions and lead times, while in many cases the ML models had a more negative contribution to the total skill than the raw GEFS/R, likely resulting from the underconfidence. The resolution term is at largest one and at least zero in this decomposition, while the reliability term is at most zero. The magnitude of the resolution terms is consistently several factors larger than the reliability term for all forecast sets, and the differences in that term generally have a larger absolute impact on the overall Brier skill scores.

Lastly, while by no means a comprehensive characterization of the system, a sample of real cases over the test period are presented to illustrate some of the strengths and weaknesses of the system. On the evening of 19 May 2015 and morning of 20 May 2015, a vigorous mesoscale convective system developed over southern Oklahoma and northern Texas, producing very heavy rainfall that contributed to historic flooding in the region during May 2015 (e.g. Wolter et al. 2016). Stage IV analysis (Fig. 3.11a) reveals that the 24-hour precipitation totals exceeded 1-year ARI thresholds within much of an E/W band encompassing the region, with embedded areas of 10-year exceedances along the state border region (Fig. 3.11b). While the ECMWF ensemble forecasts indicate some possibility of extreme precipitation in that region during this time frame at Day 3 (Fig. 3.11d), the probabilities are displaced too far to the south and west, and the probabilities of 10-year exceedances are very low. There is some improvement in positioning with the Day 2 forecast (Fig. 3.11c), but it remains too far west and with probabilities still quite low, particularly at the 10-year ARI level. Raw GEFS/R forecasts at Day 3 (Fig. 3.11f) indicate quite high risk for a 1-year exceedance over a fairly narrow area, better positioned than the ECMWF ensemble at the same lead time but still too far to the west. Outside of this area, the GEFS/R indicates almost no risk of an extreme rainfall event, and also indicates no risk of a 10-year exceedance anywhere in the domain. The Day 2 forecast (Fig. 3.11e) looks similar to the Day 3 outlook, except that the probabilities are reduced somewhat in the target area, which also has incorrectly displaced further to the south and west. The ML model depicts a much different picture. It exudes much less

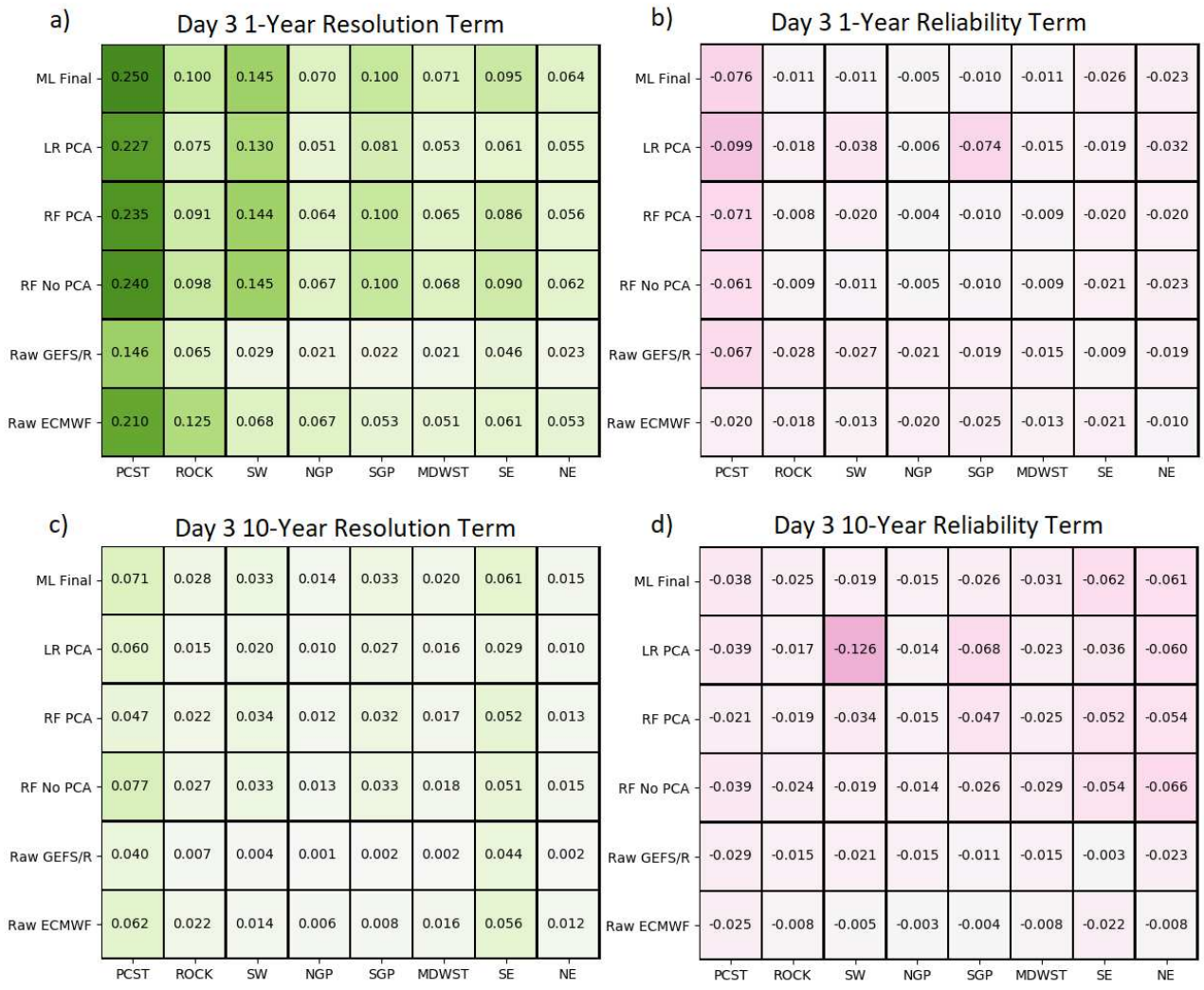
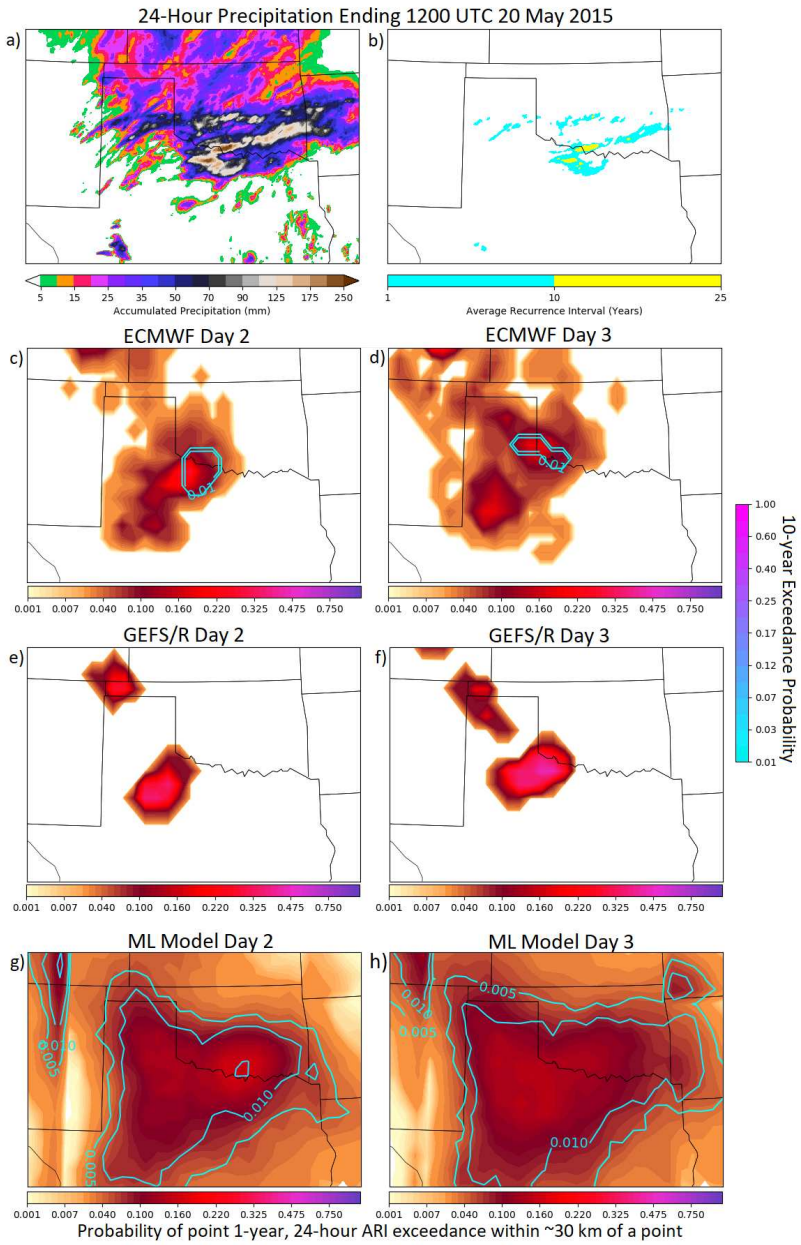


FIG. 3.10. Same as Figure 3.9, but for Day 3 forecasts.

confidence, with lower maximum probabilities compared with either raw ensemble, but non-zero exceedance probabilities of both 1- and 10-year exceedances across much of the domain for both Days 3 (Fig. 3.11h) and Day 2 (Fig. 3.11g). Importantly, the model elevated probabilities compared with the raw guidance in the place that extreme precipitation was actually observed (to the east of where it was forecast in the GEFS/R). In fact, at Day 2 (Fig. 3.11g), the probability maximum is located right where the heaviest precipitation actually occurred, displaced well to the north and east of where it was forecast in the GEFS/R (Fig. 3.11e). Additionally, while still low, the 10-year event probabilities are much higher over the verifying area when compared with either raw ensemble, with maximum Day 2 probabilities of around 30% and 3% for 1-year and 10-year exceedances, respectively. Finally, in contrast to



the raw guidance, the ML model became increasingly confident in an event occurring with decreasing lead time (cf. Fig. 3.11g,h).

A different mesoscale precipitation produced extreme precipitation over southwestern Wisconsin, southeastern Minnesota, and northeastern Iowa during the evening and overnight hours of 21 September and 22 September 2016, respectively. Based on ST4 QPE (Fig. 3.12a), much of the area experienced 1-year ARI exceedances for the 24-hour period ending 1200 UTC 22 September 2016, and within the 1-year exceedance area, there were many embedded cells that produced 10-year ARI exceedances

FIG. 3.11. Case study depicting forecasts from the final ML model and both reference ensembles for the 24-hour period ending 1200 UTC 20 May 2015. (a) 24-hour Stage IV QPE ending at 1200 UTC 20 May 2015 and (b) corresponding ARI exceedances of 1-year and 10-year thresholds. (c) ECMWF ensemble neighborhood ARI exceedance probabilities in the filled (1-year) and unfilled (10-year) contours for the 36–60 hour forecast initialized 0000 UTC 18 May 2015 and (d) for the 60–84 hour forecast initialized 0000 UTC 17 May 2015. Panels (e) and (f) depict analogous fields as panels (c) and (d), respectively, except for forecasts from the raw GEFS/R QPFs. Panels (g) and (h) similarly show respectively 36–60 and 60–84 hour forecasts, except for from the final version of the ML model trained in this study. Contours for 10-year events are 0.005, 0.01, 0.03, 0.05, 0.075, 0.10, 0.125, 0.15, 0.175, 0.20, 0.25, 0.3, 0.4, 0.5, 0.6, 0.8, and 1.0.

(Fig. 3.12b). ECMWF forecasts at Day 3 indicated risk of extreme rainfall, even at the 10-year severity level (Fig. 3.12d), but the location was poor, with exceedance probabilities high in eastern Minnesota and northern Wisconsin where extreme rainfall was not observed, and very low probabilities in northeastern Iowa and southeastern Wisconsin where it was. Both the positioning and risk of very extreme precipitation improved for the Day 2 forecst issuance (Fig. 3.12c), but probabilities still remained too far to the north. The GEFS/R at Day 3 (Fig. 3.12f) indicated very little risk of extreme precipitation in the area, with just one member correctly predicting a 1-year exceedance in southeastern Minnesota. The risk of an event occurring within the domain increased for the Day 2 issuance, but the locations got worse, with maximum risk indicated in eastern Nebraska, western Iowa, and northeastern Wisconsin, with the only 10-year prediction occurring in the latter location. Somewhat like the raw GEFS/R, the ML model had only some indication of extreme precipitation risk at Day 3 (Fig. 3.12h). However, it both had the higher probabilities (near 10% in both cases) distributed over a much larger area, and indicated some risk of a 10-year event, with probability maxima near 1.5%. Additionally, it had the maximum probability axis nearly collocated with where heaveist precipitation occurred, well to the south of the ECMWF probabilities, albeit still slightly too far to the north. The Day 2 forecast issuance (Fig. 3.12g) was largely similar. The two main changes are a correctly increased risk in the area where the event actually verified, and an incorrectly increased risk of heavy precipitation in eastern Nebraska where the raw GEFS/R had heavy precipitation on Day 2 (Fig. 3.12e).

3.5 DISCUSSION AND CONCLUSIONS

An ML model based on RFs and LR is used to generate CONUS-wide probabilistic forecasts for the exceedance of 1- and 10-year ARI thresholds for 24-hour precipitation accumulations during the Day 2 and Day 3 periods. Approximately eleven years of GEFS/R forecasts, in particular the ensemble median, are used to train these models, and forecasts are made using numerous simulated atmospheric

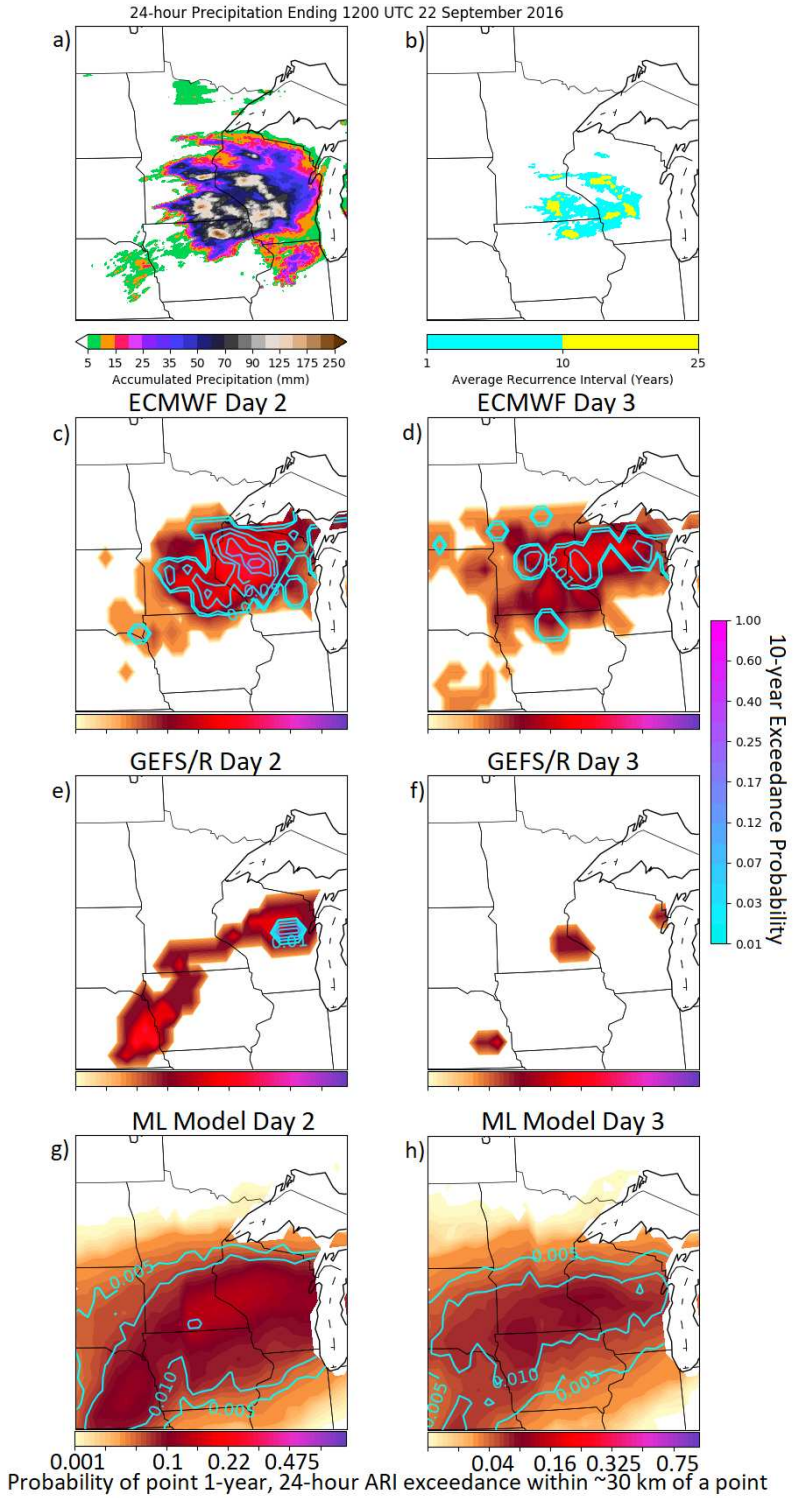


FIG. 3.12. Same as Figure 3.11, but for the 24-hour period ending 1200 UTC 22 September 2016.

fields (Table 3.1) varying in both space and time, in addition to a variety of geographic and climatological forecast predictors (Table 3.2). Separate models are trained for each of the two 24-hour periods and for each of eight different regions of CONUS, as depicted in Figure 2.11. A variety of sensitivity experiments are performed, as outlined in Table 3.3, to ascertain the utility of different aspects of forecast information in predicting locally extreme precipitation. Finally, the final forecast models were evaluated, and compared with forecasts based only on the ensemble of raw QPFs from the GEFS/R and ECMWF. The ML models trained in this study demonstrably outperformed the raw GEFS/R forecasts for all regions and forecast lead times (Fig. 3.5), often more than doubling the forecast skill and adding substantially more than 24-hours lead time improvement in forecast skill. With the exception of the PCST and ROCK regions, the same held for comparison of the ML model forecasts with ECMWF ensemble forecasts as well. Both raw ensembles tended to be negatively biased and higly overconfident in predicting extreme QPFs (Fig. 3.6), particularly at the 10-year ARI for central CONUS regions; this was reversed in the final ML model forecasts, which were more reliable at higher probabilities, but generally underconfident (Fig. 3.8).

In general, unlike past studies (e.g. Herman and Schumacher 2016b), in most regions, the temporal resolution and extent of spatially displaced predictors from the forecast point considered had little to no impact on forecast skill (Fig. 3.3), in addition to the use of upper-level information and additional ensemble information (Fig. 3.4). These results are suggestive of two findings. First, most of the relevant information about predictors displaced spatiotemporally from the forecast point, other atmospheric fields, or other ensemble member information, can be derived with at least moderate accuracy using just the information from the ensemble median from a group of core set of fields collocated and concurrent with the forecast; that is, these additional predictors contain only limited independent forecast information, at least for this coarse dynamical model and this underdispersive ensemble configuration. It also suggests that, for the most part, the predictive ability is coming primarily through a characterization of the overall environment, which can be reasonably summarized with only a subset of predictors, rather than the simulated spatiotemporal variability and full 3-D characterization of the atmospheric evolution in the underlying dynamical model. This finding comes in contrast to similar studies of other forecast problems using the GEFS/R, such as the Herman and Schumacher (2016b) study which investigated using the GEFS/R to create ML-based probabilistic forecasts of cloud ceiling and visibility at different airports and found considerable value in the inclusion of spatially displaced predictors. However, there is at least one major exception; none of this really held for the PCST region;

here, more complex models with more predictors did notably improve forecast skill. This is perhaps in part because the physical processes associated with extreme precipitation are much better resolved in the GEFS/R in this region compared with the others, and so the added information adds usable forecast utility beyond simply duplicatively characterizing the atmospheric environment for the forecast. The largest skill difference of the sensitivity experiments came for most regions in changing algorithmic assumptions and processes (Fig. 3.4d); the simpler linear assumptions of LR tended to degrade forecast skill compared with the more limited assumptions underlying the RF models.

The results of this study reveal that the application of more sophisticated statistical methods and ML algorithms such as RFs can demonstrably improve forecasts of extreme precipitation and potentially other rare, high-impact weather events in the medium range when compared with the methods and techniques that are most prevalent in forecast operations today. One unique aspect here is the scope of this model; while most past studies which employed these techniques for numerical weather prediction have focused on a small domain, or just a sampling of points, the models trained here demonstrate an ability to generate skillful, reliable forecasts year-round for all of CONUS and a range of lead times. There are many forecast problems that remain to be explored, but the results of this study and others strongly suggest that further development and application of these data-intensive statistical techniques could substantially improve our forecasts over the current state of the art, even compared with using more sophisticated dynamical models. To that end, implementation of this methodology for operational use to assist Weather Prediction Center forecasters with the development of their excessive rainfall outlooks is currently underway.

This forecast technique presents some advantages over purely dynamical approaches, as dynamical models are inherently limited by two factors by which these statistical techniques are not. First, dynamical models require ever increasing computational resources for increasing model resolution; constraints on computing power prevent sufficient resolution to directly resolve many small-scale processes, many of which are observed in the highest impact weather phenomena. Second, dynamical models are limited by our physical understanding of the processes we are attempting to simulate or forecast. Machine learning algorithms, in contrast, can detect predictive patterns in the available information even in places where we do not know or understand the physical connection between the information and the phenomenon which we wish to predict. While they are also limited in complexity by computational and data resources, the strict limits on resolvability are not there: physical resolution can often be gained through post-processing of larger scale information. There is thus ample reason

to believe that further investigation of these techniques for NWP is a worthwhile venture, and eventual implementation into forecast operations could help forecasters with their tasks by skillfully synthesizing many different sources of forecast information to help alleviate their often time-pressed schedules. This in turn can aid end-user preparedness and, in the case of high-impact events, hopefully help to protect lives and property.

One of the main advantages of the methods explored in this study compared with other popular machine learning methods, in addition to their computational tractability, is the ability to visualize their output and gain insights into detecting and quantifying specific biases in the underlying GEFS/R model, and physical insights into the most valuable forecast information for predicting locally extreme precipitation. This is the focus of the following chapter.

Some limitations of this work are worthy of note. Stage IV precipitation is used as truth for this study; though there is not a clearly better verification source available, it does have its drawbacks. It does have some spurious quality control issues, and often stuggles in areas of complex terrain due to radar beam blockage, interference, and limited gauge coverage (Herman and Schumacher 2016a; Nelson et al. 2016). Since the model is trained to forecast Stage IV QPE exceedances, this can lead to some idiosyncrasies and other anomalies associated with the biases observed in the Stage IV product. One such anomaly is the persistent presence of very small areas of exceedances in some regions of complex terrain during times of favorable convective conditions. This can be removed by quality control procedures to some extent, but some artifacts do remain. This happens most prominently in the terrain of western New Mexico; a small region there has many more instances of ARI exceedances over both the training and test periods than any other part of CONUS. The ML-based models recognize this, and for the SW region consistently issue much higher probabilities in this region. In one sense, this is correct—it is correctly predicting what it was trained to predict—but is still undesriable behavior due to a disparity between "truth" in the study and the true extreme rainfall risk. Solutions to this issue and related issues in other parts of the country must be explored in order to maximize operational utility. Additionally, while the choice of using the ARI framework was an intentional decision and provides numerous benefits, it is not an end-all for predicting heavy precipitation impacts. While ARIs often have *better* correspondence with impacts than a fixed threshold, there are still regional discrepancies in which ARIs have optimal association with impacts, and the framework employed here does not account for antecedent conditions, which can be critical for assessing flash flood risk. More investigation

into the relationship between QPE exceedances and rainfall impacts should be performed to maximize the practical significance of the model predictand.

Additionally, the predictors for this study come from a very coarse and otherwise rather antiquated global model. The GEFS/R was used for this study because, unlike almost any other dynamical model, it has been nearly static for a very long period of record and has nearly stationary bias characteristics—an essential property for performing this kind of analysis. However, the models trained herein are not working off of the 'state of the art' of flash flood predictors. The longer range Day 2 and Day 3 lead times were chosen for this study in part because the discrepancy between GEFS/R forecast quality and 'state of the art' is smaller at these longer lead times due to less convection-allowing guidance being available, and higher-resolution models degrading in utility with increasing forecast lead time (e.g. Zhang et al. 2003, 2007).

There are also some complications that must be considered for real-time implementation. As one example, the regional models are trained completely independently of one another, with different training data and different solutions. Consequently, they can occasionally give rather different predictions on nearly identical inputs, resulting in undesirable probability discontinuities across region boundaries. Appropriate methods for removing probability discontinuities in space must be further explored.

Future work will seek to alleviate these limitations in a variety of ways. Exploration of using different predictands, likely combining hydrometeorological information from a variety of sources, will be made for more explicit flash flood prediction. This may involve a regionally varying predictand definition, with some ARI thresholds better corresponding to flash flood impacts in some regions compared with others. Additionally, although a large number of predictors were explored in this study, there are many additional choices for predictors that could ostensibly further improve forecast skill. While atmospheric fields are represented here in absolute terms, it may be beneficial to instead represent some fields relative to the local climatology of the forecast point in terms of standardized anomalies. This is particularly true for fields like PWAT, where standardized anomalies have often shown better correspondence with precipitation impacts across varied regions than absolute values (e.g. Junker et al. 2009; Graham and Grumm 2010; Nielsen et al. 2015). More exploration of derived fields of physical relevance to extreme precipitation processes should also be explored. Some possible examples include upslope flow to gauge forcing for ascent by the horizontal wind, column mean wind to ascertain potential for slow-moving storms, and deep-layer shear as a metric for supercell potential.

This study also focused on a rather specific time interval and took all dynamical predictors from a single, somewhat antiquated ensemble system. Future expansion both to the 12–36 hour Day 1 period and beyond the Day 3 period will be explored, including predictors from more contemporary CAM guidance and potentially including observations as well for the shorter lead time forecasts. Operational models also tend to undergo periodic upgrades and thus do not remain static like the ensemble system used here. The sensitivity of ML model performance to changes in dynamical model bias characterics that result from these upgrades is a question of considerable operational relevance and an additional factor worthy of future investigation. It was also seen that the ML models suffered to varying degrees from underconfidence and, in some instances, negative bias. Methods of probability calibration of the ML model probabilities as a final post-processing step (e.g. Hagedorn et al. 2008; Hamill et al. 2008; Bentzien and Friederichs 2012; Herman and Schumacher 2016b) should be explored in future work, and parameter choices reconsidered in light of this additional calibration. Finally, this study only explored a subset of available machine learning algorithms. Other choices, including adaptive learning algorithms, may be able to better exploit predictor-predictand relationships, appropriately update to reflect changes in an underlying dynamical model, and produce superior forecasts for the locally extreme precipitation and flash flood forecast problem (e.g. Liu et al. 2001; Roebber 2015; Pelosi et al. 2017).

CHAPTER 4

"DENDROLOGY" IN NUMERICAL WEATHER PREDICTION: WHAT RANDOM FORESTS AND LOGISTIC REGRESSION TELL US ABOUT FORECASTING EXTREME PRECIPITATION

4.1 INTRODUCTION

Machine learning algorithms have demonstrated considerable utility in many scientific disciplines, including computer vision (e.g. Rosten and Drummond 2006), natural language processing (e.g. Collobert et al. 2011), and bioinformatics (e.g. Larranaga et al. 2006). Machine learning has also been used with considerable success in a wide range of future prediction scenarios, from financial market analysis (e.g. Cao and Tay 2003), to election forecasting (e.g. Bermingham and Smeaton 2011), to numerical weather prediction (NWP; e.g. Hall et al. 1999; Roebber 2013; Rozas-Larraondo et al. 2014; McGovern et al. 2017). Recently, these techniques have been receiving increasing attention and application in NWP; many of these preliminary forays have demonstrated considerable utility of these techniques over historical competitors (e.g. Herman and Schumacher 2016b), with occasional exception (e.g. Applequist et al. 2002).

One frequently noted criticism of machine learning forecast models is their lack of interpretability and neglect of underlying physics and dynamics of the forecast problem, rendering additional interpretation and analysis of their output difficult or impossible. These critiques did not first appear with the emergence of machine learning; in fact, these qualms with statistical forecast models have been expressed since early days of NWP (e.g. Lorenz 1956). And there is legitimate reason for these concerns; given the chaotic nature of the atmosphere system, any model—statistical or dynamical—will necessarily have formulaic limitations, systematic biases, and failure modes regardless of the level of care exercised during model construction. When the model's processes are opaque, it can be difficult to rationally diagnose these circumstances, and the ability of the forecaster to add value over the raw guidance is inhibited. Thus, even when, for example, a statistical model exhibits better objective performance compared with a competing dynamical model, if a human forecaster understands the underpinnings and characteristics of the dynamical model but not the statistical model, he or she may still be able to provide better final forecasts using the dynamical guidance over the statistical guidance. The "understanding" referenced here does not require a complete and comprehensive mathematical understanding sufficient to exactly reproduce the result by hand; even using a very simple dynamical

model, it is extraordinarily difficult to reproduce an accurate forecast by manual means (e.g. Richardson 2007), and seldom are interpreters of model guidance familiar with numerical specifics, dynamical core particulars, or parameterization details. Rather, there is a well-understood overarching process of using data assimilation to produce an analysis and initialize a model which embodies the primitive equations governing the atmosphere in some capacity, and then integrating the model forward in time to produce a forecast. Additionally, the intermediate steps—output from hours after initialization but before forecast valid time—are fully inspectable and comprehensible. In contrast, to many, statistical models and especially those employing machine learning seem comparatively opaque; a host of predictors are ingested, and a forecast(s) is produced, with little if any information provided on how the model got from the predictors it used to the answer it generated. While a small part of this is perhaps inherent to statistical forecasting, with improved visualization of statistical models developed for NWP, physical insights into how the predictors used relate to the forecasted phenomenon may be gained, and ability to deduce likely biases based on the present meteorology may be acquired.

Among statistical forecast algorithms, regression models have the longest and most extensive use in operational NWP (e.g. Glahn and Lowry 1972) and are perhaps the most easily and directly interpretable through their regression coefficients. Using the regression coefficients, operational regression models such as the Statistical Hurricane Intensity Prediction Scheme (SHIPS; DeMaria and Kaplan 1994) can display the individual effect of each element of the present meteorology on the final prediction. With care, this also allows interpretation of the relative utility of different pieces of meteorological information in predicting the forecast phenomenon of interest, in this case tropical cyclone intensity (e.g. Jones et al. 2006). Direct inspection of the parameters is equally insightful for other types of regression, such as in multivariate logistic regression (LR) for probabilistic forecasts (e.g. Bremnes 2004). While direct interpretability is an attractive quality of regression models, the parametric nature of them and like algorithms imposes assumptions on the relationship between the predictors and the predictand or between predictors themselves when such relationships may not be accurate or even known or physically understood (Wilks 2011). Linear and logistic regression, for example, both impose a fundamentally linear predictor-predictand relationship and treat predictors independently, not directly accounting for the covariance between multiple predictors and their joint relationship with the predictand. While imposing these restrictions can actually be helpful when they are physically valid, predictive performance degrades when these imposed assumptions are invalid.

Especially when the physical relationships are not known or well quantified, it is often attractive to employ an algorithm which does not impose such assumptions. One such example is the random forest technique (RF; Breiman 2001). RFs have been used for many different applications in NWP including but not limited to prediction of storm-type classification, turbulence, cloud ceiling and visibility, convective initiation, and hail size (e.g. Williams 2014; Herman and Schumacher 2016b; Ahijevych et al. 2016; Gagne et al. 2017; McGovern et al. 2017). Though the algorithm is more general, the innerworkings of an RF may be diagnosed, like with regression coefficients for LR, primarily by means of feature importances (FIs) to be used and discussed in more detail in this study. While these have already been used to assess RF NWP models in some past studies (e.g. Gagne et al. 2014; Herman and Schumacher 2016b), in using locally extreme precipitation forecasting as an example, we will demonstrate here that they can be used to understand spatiotemporal relationships as well as relationships across atmospheric fields in predicting the phenomenon of interest, even when the number of predictors grows large, the event becomes rare, and algorithmic steps that complicate the relationship between the predictor inputs and reality are performed.

Chapter 3 expanded upon these prior studies using machine learning for NWP in a variety of ways. While there have been limited prior studies using machine learning to explicitly investigate very rare events (e.g. Marzban and Stumpf 1996; Marzban and Witt 2001), and some prior studies constructing statistical models for QPF (e.g. Hall et al. 1999; Sloughter et al. 2007; Whan and Schmeits 2018), there has been little published work to date combining both facets. The work in Chapter 3 was among the first to do so, training statistical models to forecast locally extreme precipitation across CONUS in the medium-range. The CONUS-wide gridded scope of the models trained therein is also uncommon among machine learning models, which are often trained for points (e.g. Herman and Schumacher 2016b) or over a limited domain (e.g. Gneiting et al. 2005). Furthermore, the scope of predictors was very large, with thousands of predictors capturing the spatiotemporal environmental characteristics of the forecast point during the accumulation period. Many different sensitivity experiments were performed, and the performance of the model forecasts was evaluated in detail from both the perspective of forecast skill and reliability. Overall, forecasts were found to add both considerable skill and reliability across all of CONUS compared with both climatology and the raw forecasts of the global ensemble from which the model predictors were derived. However, the study did not investigate the internals of these models: how to visualize what they're doing to get from their input to their output, and what these algorithms and models reveal about the prediction of locally extreme precipitation events overall.

Using the regression and tree-based models of Chapter 3, this "dendrological" study investigates the details of the fitted trees, as well as the regression models. We illustrate how models based on seemingly abstract and complex algorithms and techniques can, with modest effort, be readily interpreted and understood. It is shown that, not only can these models yield more skillfully verifying forecasts than raw dynamical model output or forecasts derived from simpler, more traditional post-processing approaches, but they can also provide both statistical and physical insights into why they behave as they do, as well as insight into the deficiencies, errors, and limitations of the dynamical model predictors on which they are based. In this study, examination of the Chapter 3 models sheds insights onto how a global, convection-parameterized dynamical ensemble behaves in forecasting extreme precipitation events across the hydrometeorologically diverse regions of the contiguous United States (CONUS), and on what statistical corrections can be made to improve forecasts thereof. Section 2 briefly summarizes the methods of Chapter 3 to describe the underpinnings of the models evaluated in this study and how they were derived. Section 3 describes how the models will be visualized and interpreted in this study. Sections 4, 5, and 6 present results respectively for PCA diagnostics, RF models, and LR models. Section 7 concludes with a synthesis of the findings and a discussion of their implications.

4.2 Data and Methods Summary

What follows is an abbreviated description of the full data and methods of Chapter 3, highlighting the aspects that are critical for proper interpretation of the results presented herein. The interested reader is encouraged to review the full methods of that study for a more complete discussion of the mathematical underpinnings of the algorithms, justification of choices made, and the sensitivity experiments performed therein.

The models evaluated in this study are trained to forecast locally extreme precipitation across CONUS for 24-hour precipitation accumulations, quantified with respect to average recurrence interval (ARI) exceedances. In particular, models are trained to issue probabilistic forecasts for exceedances of 1-year and 10-year ARIs within a $\sim 0.5^\circ \times 0.5^\circ$ spatial domain during a 24-hour 1200 UTC–1200 UTC accumulation interval. Forecasts are made for two different forecast lead times comprising the 36–60 and 60–84 hour periods—denoted respectively Day Two and Day Three—with separate models trained for each period. Unique models are also trained for each of eight different geographic regions of CONUS, as depicted in Figure 2.11. Here, CONUS has been partitioned to produce cohesive regions with some hydrometeorological homogeneity with particular regard to similar magnitudes of extreme

precipitation, similar diurnal and seasonal precipitation climatologies, and similar storm types and precipitation processes associated with extreme precipitation.

Dynamical model data used for training the statistical models in this study comes from NOAA's Second-Generation Global Ensemble Forecast System Reforecast (GEFS/R; Hamill et al. 2013) dataset. The GEFS/R is an 11-member ensemble with T254L42 resolution—which corresponds to an effective horizontal grid spacing of ∼55 km at 40° latitude— initialized once daily at 0000 UTC back to December 1984. Forecast fields evaluated in this study are archived on a grid with ∼ 0.5° horizontal spacing. For Day 2 models, forecast fields use 3-hour temporal resolution, while 6-hour resolution is used for Day 3 models. Trained models discussed in this study are based on the ensemble median of a core set of nine atmospheric fields: accumulated precipitation (APCP), surface-based convective available potential energy (CAPE) and convective inhibition (CIN), precipitable water (PWAT), surface temperature (T2M) and specific humidity (Q2M), surface zonal (U10) and meridional winds (V10), and mean sea level pressure (MSLP). Models are trained using daily forecasts spanning from January 2003 through August 2013.

The National Centers for Environmental Prediction (NCEP) Stage IV Precipitation Analysis product (Lin and Mitchell 2005) has been created daily in an operational capacity since December 2001. Stage IV provides 24-hour analyses over CONUS on a ∼4.75 km grid. It uses both rain gauge observations and radar-derived rainfall estimates to generate an analysis, and is further quality controlled via NWS River Forecast Centers (RFCs) to ensure stray radar artifacts and other spurious anomalies do not appear in the final product. Despite some limitations (Herman and Schumacher 2016a), its analysis quality; resolution, allowing relatively accurate quantification of very heavy precipitation; and data record length make it preferable to other precipitation analysis products, and is therefore used as the precipitation 'truth' for this study.

The thresholds associated with the 1- and 10-year ARIs are generated using the same methodology of Herman and Schumacher (2016a), where CONUS-wide thresholds are produced by stitching thresholds from several sources. NOAA's Atlas 14 thresholds (Bonnin et al. 2004, 2006; Perica et al. 2011, 2013), an update from older work and currently under development, are used wherever they were available at the time this research began. For five northwestern states—Washington, Oregon, Idaho, Montana, and Wyoming—updated thresholds are not available, and derived Atlas 2 threshold estimates are used instead (Miller et al. 1973; Herman and Schumacher 2016a). In the Northeast—New York, Vermont, New Hampshire, Maine, Massachusetts, Connecticut, and Rhode Island—and Texas, both of which did not

have Atlas 14 threshold estimates at the time research commenced but have either since received an update or have an update in progress, Technical Paper 40 (TP-40; Hershfield 1961) estimates are used. Everywhere else uses the Atlas 14 threshold estimates.

Generating predictors by taking GEFS/R forecast values from 9 different fields every 3 or 6 hours over a 24-hour forecast period at every grid point within $\sim 2^\circ$ of the forecast point yields thousands of model predictors. In addition to the large quantity, they are also highly correlated—spatially, temporally, and across variables. With millions of training examples and thousands of predictors, the forecast problem can become computationally intractable and the correlated variables can result in overfitting. To address these issues, use of a pre-processing step whereby the model predictors undergo dimensionality reduction via principal components analysis (PCA) is explored. This creates a small set of uncorrelated predictors that explain the signal in the forecast data and gives insight into the regional modes of atmospheric variability as depicted in the GEFS/R model, while leaving the noise in withheld lower-order principal components (PCs). While PCA has been most applied in the atmospheric sciences for identifying spatial patterns at the largest scales (e.g. Thompson and Wallace 1998; Wheeler and Hendon 2004), flavors of PCA have been successfully applied to identify smaller synoptic and mesoscale features as well (e.g. Mercer et al. 2012; Peters and Schumacher 2014).

Chapter 3 performed a wide array of sensitivity experiments, exploring model predictive performance as a function of predictor temporal resolution, spatial extent, inclusion or exclusion of different atmospheric fields, use of ensemble information, algorithmic parameters, and choice of model algorithm. To manage the scope of this study's analysis, only results as a function of the last of these is presented. Much like the skill results presented in Chapter 3, general physical findings are found not to vary appreciably as a function of any of these unshown dimensions of variability. Three specific Chapter 3 models are evaluated in depth in this study: 1) the CTL_NPCA model using random forests and no PCA dimensionality reduction; 2) the CTL_PCA model using random forests with pre-processing using PCA dimensionality reduction; and 3) the CTL_LR model using logistic regression and also using PCA pre-processing. Table 4.1 provides a summary comparison of these three models for reference.

Random forests (Breiman 2001) are in essence an ensemble of *decision trees*, whereby each tree of the forest makes an individual prediction of the predictand outcome; the relative frequencies of each possible outcome in the ensemble of trees are then used to make a single probabilistic forecast. Decision trees are explained in mathematical depth in Chapter 3; an alternative way to conceptualize them begins with a many dimensional predictor phase space, where each predictor has a unique dimension.

TABLE 4.1. Summary of the models trained in this study, and the corresponding names designated to the models. 'X' indicates the process is performed or the information is used; a lack of one indicates the opposite. MEDIAN corresponds to the ensemble median. Horizontal radius is listed in grid boxes from forecast point; timestep denotes the number of hours between GEFS/R forecast field predictors. Slashes indicate the first number applies to the Day 2 version of the model, while the latter number applies to the Day 3 version. Models apply to all eight forecast regions and have both Day 2 and Day 3 versions.

Model Name	CTL_NPCA	CTL_PCA	CTL_LR
Algorithm	RF	RF	LR
PCA Pre-Processed		X	X
Ensemble Information	MEDIAN	MEDIAN	MEDIAN
Horizontal Radius	4	4	4
Timestep	3/6	3/6	3/6

Beginning with an unpartitioned phase space (tree root), a decision tree makes successive splits along axes of this space partitioning it into increasingly many smaller subspaces (splits), and then assigning predictions to each subspace (leaves). An RF creates many different similarly plausible partitions of the subspace, and a forecast is determined by the subspace labels associated with the given point in predictor space.

Logistic regression is an implementation of the generalized linear model, designed for binary predictions and classification more generally where the predictand is constrained to be either one outcome or another, rather than over a continuous space as with linear regression (Wilks 2011). Like with linear regression, logistic regression uses as its input a linear combination of the predictors. The difference arises in the use of the *link function*. For linear regression, the link is the identity function; that is, the prediction *is* the aforementioned linear combination of the predictors. In the case of logistic regression, the predictor-predictand link is made through use of the logit function instead (Wilks 2011). In particular, the model output in multinomial logistic regression—the probability of each event class—is given by use of a generalization of the logistic function:

$$P(y=k|\mathbf{x}) = \frac{e^{\mathbf{x}^{\mathrm{T}}\mathbf{w}_k}}{\sum_{j=1}^{K} e^{\mathbf{x}^{\mathrm{T}}\mathbf{w}_k}} \tag{4.1}$$

where k is the event class, x is the predictor vector, and w is the vector of regression coefficients. Note that separate coefficient vectors are computed for each event class.

4.3 Methods: Model Properties and Assessment

One of the most powerful aspects of machine learning algorithms—and RFs in particular—is finding patterns in the supplied training data. Because of the extent and diversity of the data supplied in these experiments, the RFs trained for this study have the theoretical capability of diagnosing and automatically correcting for various kinds of GEFS/R model biases. In particular, context-dependent quantitative biases, such as GEFS/R QPF being systematically too high or too low, may be diagnosed; spatial displacement biases in the placement of extreme precipitation features may be diagnosed; and temporal biases in the initiation or progression of extreme precipitation features may also be diagnosed to some extent. These can be at least partially visualized through RF Feature Importances (FIs). The most intuitive way to conceptualize their quantitative significance is by the number of splits based on the feature summed over the forest, with each split weighted in proportion to the number of training samples encountering the split so that a split at the root node is considered much more important than a split deep into a tree (Friedman 2001). Values are normalized so that the sum of all importances is one; an importance of one then indicates that all decision nodes in every tree of the forest split on the corresponding feature, while an importance of zero indicates that no decision node splits based on that feature. Importances are produced for each input feature; without PCA pre-processing, this means that an individual importance value is produced for each forecast point-relative location, forecast time, atmospheric field combination. In many cases, it is convenient to present importances summed over one or more of these dimensions for a summary perspective of the model output. When PCA pre-processing is performed, the model output is instead importances of individual PCs in predicting ARI exceedances. FIs calculated in this way—often termed the "Gini importance"—are only one method of providing a summary representation of an RF (Strobl et al. 2007). In the leading alternative method, the so-called "permutation accuracy importance" approach (Strobl et al. 2008), for each predictive feature, the feature value for each sample used to construct a given tree is permuted to a different sample's value. Importance is calculated as the decline in predictive performance between the model when the permuted values of the feature are supplied from when the true values are used. This is calculated individually for each tree and then averaged over the entire forest. While this approach has some advantages over other approaches (e.g. Breiman 2001; Strobl et al. 2007, 2008), the "Gini importance" measure is used for this study for consistency with past studies in the field (e.g. Herman and Schumacher 2016b; Gagne et al. 2017; Whan and Schmeits 2018) and computational simplicity (Pedregosa et al. 2011).

One of the main advantages of LR is its interpretability; through the regression coefficients, there is a direct, concrete connection between the predictors and the forecast predictand. And although the regression in the CTL_LR model is performed on the principal components and not the native atmospheric variables, the relationship to the native features may be readily backed out through the PCA loadings matrix L:

$$\begin{aligned}\mathbf{x^T w}_k &= w_{k,1}PC1 + w_{k,2}PC2 + w_{k,3}PC3 + ... + w_{k,R}PCR \\ &= w_{k,1}(\sum_{m=1}^{M}\mathbf{L}_{1,m}F_m) + w_{k,2}(\sum_{m=1}^{M}\mathbf{L}_{2,m}F_m) + w_{k,3}(\sum_{m=1}^{M}\mathbf{L}_{3,m}F_m) + ... + w_{k,R}(\sum_{m=1}^{M}\mathbf{L}_{R,m}F_m) \quad (4.2)\end{aligned}$$

which yields

$$\begin{bmatrix} F_1 & F_2 & F_3 & \dots & F_M \end{bmatrix} = \begin{bmatrix} L_{1,1} & L_{1,2} & L_{1,3} & \dots & L_{1,R} \\ L_{2,1} & L_{2,2} & L_{2,3} & \dots & L_{2,R} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ L_{M,1} & L_{M,2} & L_{M,3} & \dots & L_{M,R} \end{bmatrix} \times \begin{bmatrix} w_{k,1} \\ w_{k,2} \\ \vdots \\ w_{k,R} \end{bmatrix} \quad (4.3)$$

where R is the number of retained PCs, M is the number of native features, k is the event class, F is the vector of native features, and w is the vector of regression coefficients.

Both algorithms have their advantages, disadvantages, and caveats in interpretation. As noted above, LR has the advantage of a direct quantitative link between any given predictor of interest and the predictand. RF FIs, in contrast, give only an "importance" number, which gives no indication of the sign or magnitude of the predictor in order to correspond with event observance. The task of manually inspecting the value of every node split based on the predictor, while it can be executed, is cumbersome and it's difficult to draw general conclusions due to the deeply layered subspaces involved. However, RFs do have major advantages over LR in interpretation as well. As a linear model, LR coefficients are constrained to apply globally, but this is often not an appropriate constraint. Some predictors may only become important when other conditions are satisfied—for example, CAPE might only be important when there is a lifting mechanism to release the instability—rendering them insignificant in most cases but very important under select circumstances. In LR, where the coefficient applies uniformly regardless of the circumstances, the coefficient would necessarily be small, while the RF FI for the same predictor could be relatively large by taking advantage of the importance of a the variable in a particular subspace(s) of the larger predictor space. RFs also handle correlated predictors better than LR. In

regression problems, when one predictor is highly correlated with another, one is liable to have a situation whereby the "weight" is disproportionately allocated to one predictor over the other, giving the false appearance that one variable is highly predictive while the other is not. In RFs, with two highly correlated predictors that are thus approximately equally predictive, node splits will occur essentially randomly between one or the other, and the RF FIs thus have a tendency to balance approximately evenly (Gagne 2016). This problem of LR is greatly alleviated in the CTL_LR model by using PCs as the predictors, which are necessarily constructed to be orthogonal to one another. However, analyzing these different algorithmic formulations in tandem enables capturing a more complete picture of the extreme precipitation forecast problem.

4.4 Results: GEFS/R Principal Components Analysis

Inspection of the leading mode of atmospheric variability—PC1, the component that explains the most variance between different days, or model initializations—in Figures 4.1 and 4.2 for the NGP and PCST regions, respectively, and for the remaining regions in an online supplement reveals that the leading mode in each region quite apparently relates to the seasonal cycle. However, the precise nature of that seasonal cycle varies by region. Like colors across subpanels in these figures indicate that atmoshperic fields covary together for the region's displayed PC, while contrasting colors indicate one variable is anomalously high while the other low. Deeper reds associate positively with the PC, with blues associating negatively; lighter colors indicate that the given predictor does not relate as strongly with the PC. Spatial color inhomogeneities within a subpanel suggest the PC is associated with a spatial gradient in the field, while loadings changing throughout the forecast period—shown via comparison of the unfilled contours—is indicative of some degree of regime change. By happenstance, positive values of PC1 in all regions is compared with a summer signal, while negative values are associated with a winter signal. In all regions, the summer signal is associated at all times of day with high surface temperature and moisture (e.g. Fig. 4.1d,e), higher PWAT and CAPE (e.g. Fig. 4.1g,h), and lower MSLP and CIN (e.g. Fig. 4.1b,i). In almost every region, the warm-season signal (positive PC1) is weakly associated with anomalous precipitation (APCP) for the region (e.g. Fig. 4.1a); in other words, this states that the warm-season is also the wet-season in most regions of CONUS. However, in the Pacific Coast (PCST) region, precipitation is predominantly received during the cool-season (Herman and Schumacher 2016a), and this is reflected by negative loadings for the APCP field seen in Figure 4.2a. The primary regional differences between the seasonal cycle, and reflected in the PC1 loadings, is seen in

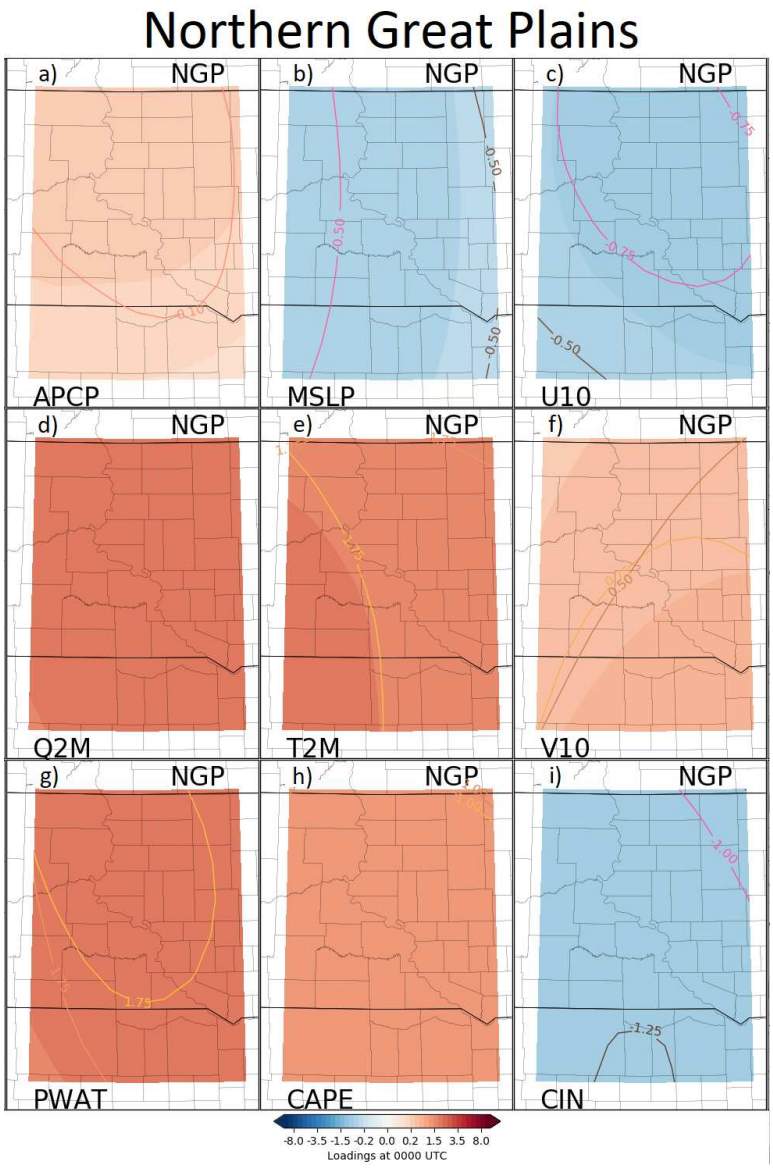


FIG. 4.1. PC1 loadings for the NGP region. Panels (a)–(i) correspond respectively for loadings associated with the APCP, MSLP, U10, Q2M, T2M, V10, PWAT, CAPE, and CIN fields. Filled contours depicts loadings for forecast values at 0000 UTC during the forecast period (forecast hour 48) with reds indicating positive loadings and blues negative loadings; magenta and yellow contours indicate negative and positive loadings, respectively, for 1500 UTC during the period (forecast hour 39), while brown and beige contours depict negative and positive loadings for 0900 UTC during the forecast period (forecast hour 57). Darker colors indicate larger values, and accordingly a stronger relationship with the principal component as indicated in the figure colorbar.

the wind fields (Fig. 4.1 and 4.2, panels c,f). In most regions, including NGP, the warm-season is assoicated with anomalous southeasterly flow at low levels, as evidenced by positive V10 loadings (Fig. 4.1f) and negative U10 loadings (Fig. 4.1c). However, this is not true of the western regions; PCST, like with APCP, exhibits the opposite behavior to the eastern regions in the wind fields, with a warm-season

characterized by anomalous northwesterly flow (Fig. 4.2c,f). The strength of association with PC1 also varies between atmospheric fields. The seasonal cycle, at least as reflected in PC1, is predominantly a thermodynamic and moisture signal; this is seen by observing larger loading magnitudes with fields such as Q2M, T2M, and PWAT compared with APCP and other fields (cf. Fig. 4.1d,e,g with 4.1a–c,f,i).

In one sense, the seasonal cycle and thus PC1 is rather trivial—it is already largely known and understood. It would be possible to train these models with deseasonalized predictors and an additional predictor(s) to represent location in the seasonal cycle, and this prospect is worthy of further investigation in future work. But this could appreciably harm predictive performance of the model; in many instances, a certain quantity of a precipitation ingredient such as precipitable water or CAPE (e.g. 35 mm or 1500 J kg^{-1}, respectively) are necessary to generate locally extreme precipitation-producing storms; by instead supplying deseasonalized predictors, these physical thresholds, which may be climatologically much more likely in one season than another, are severed from the numerical values of the model predictors. This forces the model to in essence relearn the seasonal cycle via a combination of the seasonal indicator predictor and deseasonalized atmospheric predictors, in addition to all of the other relationships it must diagnose, placing an extra burden in model training. This would likely sacrifice predictive accuracy of the trained models, perhaps with the gain of a more physically insightful PC1.

PC2—the leading mode of atmospheric variability at a point aside from the seasonal cycle—is depicted for the NGP and PCST regions in Figures 4.3 and 4.4; PC2 loadings for other regions may be found in the online supplement of Herman and Schumacher (2018a). While PC1s were largely similar between the regions, there are substantial regional differences between the PC2 loadings. Generally, while PC1 is predominantly a thermodynamic signal, many PC2s are predominantly a kinematic signal, with the largest loading magnitudes typically seen in U10 and V10. Furthermore, while PC1 loadings had little temporal dependence, for PC2 and beyond, loadings changing sign or magnitude across the forecast period is commonplace (e.g. Fig. 4.3a,b,i). One notable commonality is that in many regions, PC2 shares at least some characteristics one might expect associated with frontal passage, including rapid changes in meridional winds (e.g. SE, SW, NE regions—see online supplement), pressure falls (e.g. Fig. 4.3b), precipitation and moisture changes (e.g. Fig. 4.3d,g), and even instability "advection" (e.g. SGP). In the PCST region (Fig. 4.4), where fronts are thermodynamically weak compared with other regions, they govern a smaller portion of atmospheric variability in the region and are not associated with PC2. The signal looks somewhat atmospheric river-like, with heavy precipitation (Fig. 4.4a), column-integrated moisture advecting in from the southwest with strong low-level southerly flow (Fig.

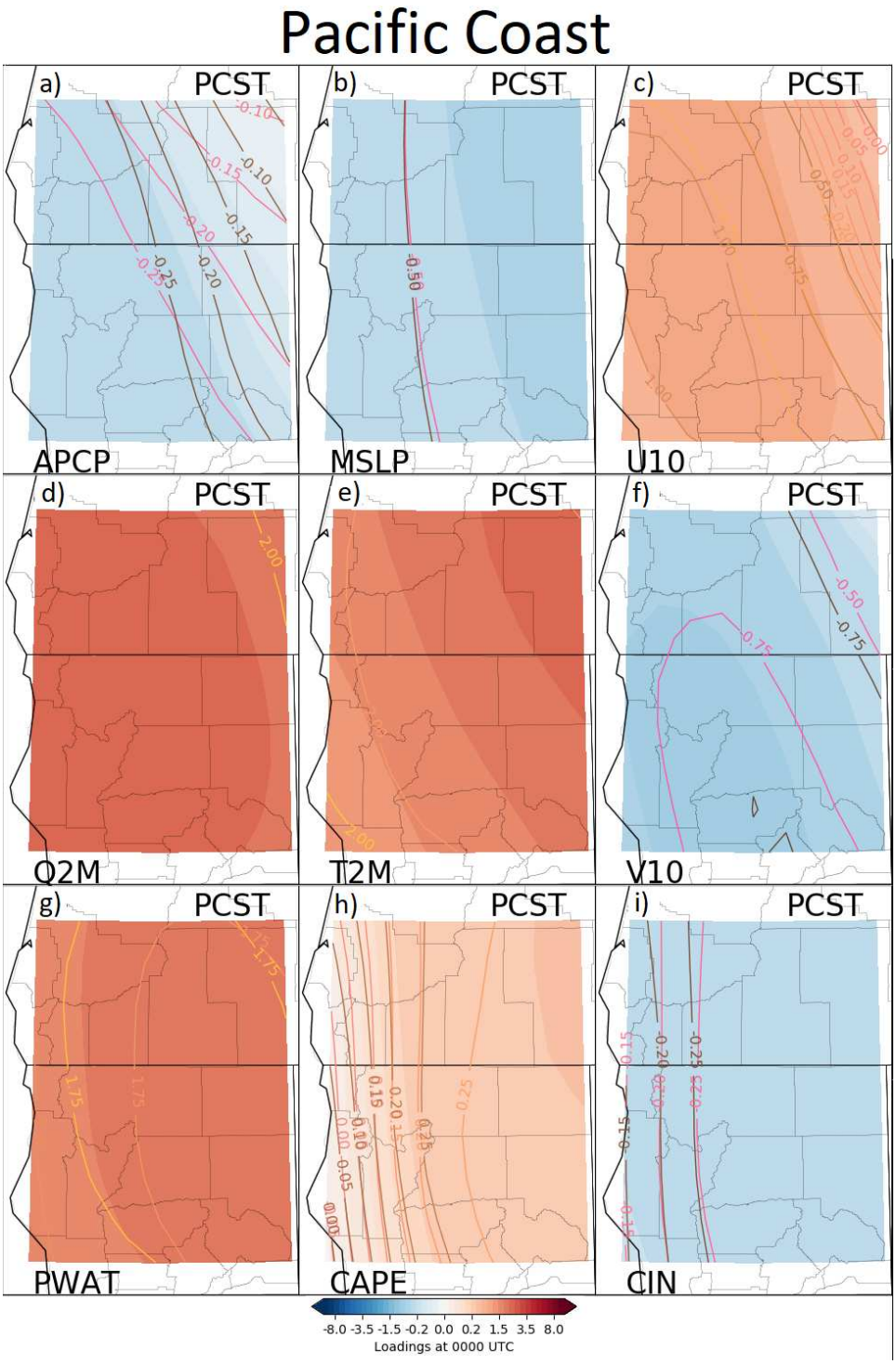


FIG. 4.2. Same as Figure 4.1, except for the PCST region.

4.4f,g), and low pressure and temperature (Fig. 4.4b,e), at least when compared with the warm-season. Again though, some loadings do not appear entirely consistent with this interpretation (e.g. Fig 4.4c). None of the PC2 loadings appear to have a direct physical interpretation that clearly matches with every

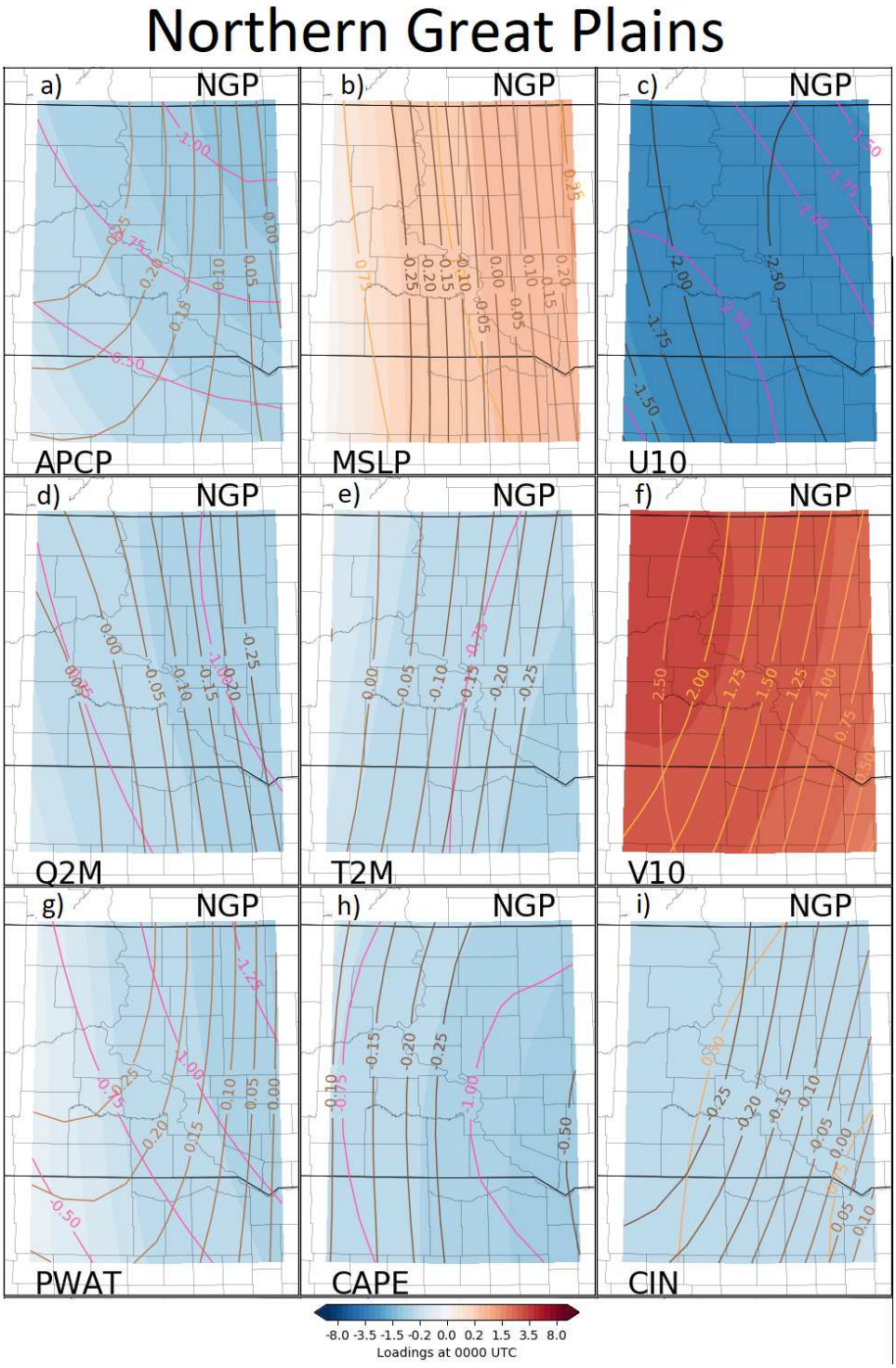


FIG. 4.3. Same as Figure 4.1, except for PC2.

aspect portrayed by the PC, a known drawback imposed by the combined orthogonality and maximum variance limitations imposed in the PCA formulation (e.g. Richman 1986).

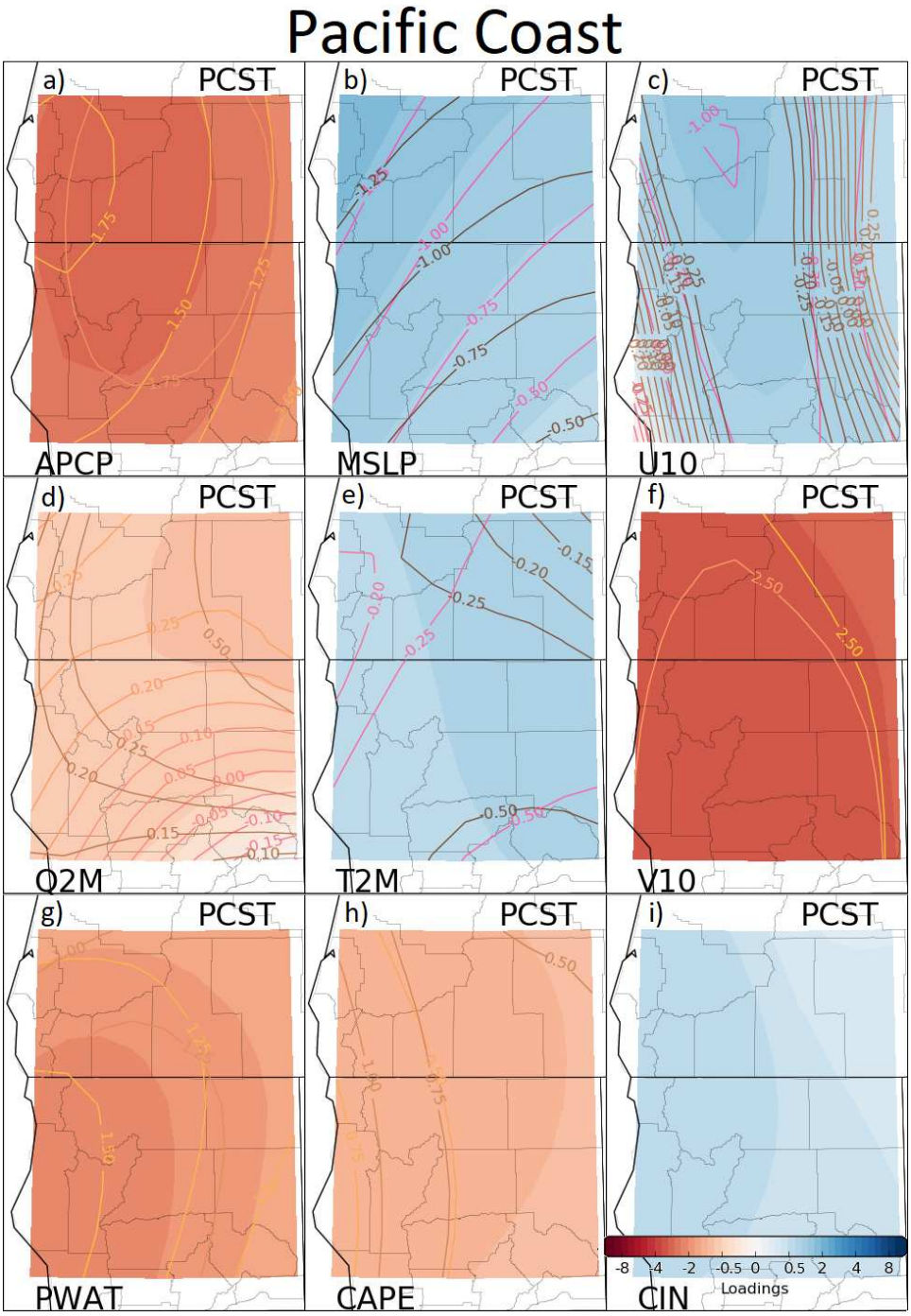


FIG. 4.4. Same as Figure 4.2, except for PC2.

The principal components themselves speak only about general regional atmospheric variability at a point as portrayed by the GEFS/R; they do not themselves relate at all to any predictand of study, in this case extreme precipitation. However, to gain understanding on how the GEFS/R forecast data projected into the rather abstract PC space relates to extreme precipitation, visualization of exceedances

and non-exceedances in various 2-D PC subspaces can be made. For brevity, this is performed in a subset of regions for both forecast periods, beginning with ROCK in Figure 4.5. Since there are tens or hundreds of retained PCs, the number of 2-D subspaces is very large; shown in each of these figures are only those subspaces containing two of the five PCs identified as most predictive of extreme precipitation in that region for the given forecast period, based on RF FIs presented in Section 5. PC1, which as is seen in Figures 4.1 and 4.2 represents the seasonal cycle for both days, has some ability to discriminate events from non-events. The blue, cyan, and magenta pixels, indicating ARI exceedances, are found predominantly on the far positive end of the PC1 axis (the right side of Figures 4.5f,k,p,u; the top of Figures 4.5b-e), indicating that most ARI exceedances occur during the warm-season in the ROCK region. The intermingled yellows, oranges, greens, and pixels of other colors throughout a panel, somewhat akin to that of Figure 4.5l, indicates poor distinction between events and non-events in that PC subspace. ROCK Day 2 PC2, summarized in Figure 4.5f2, is characterized mostly by winds becoming increasingly westerly and decreasingly southerly throughout the period, and appears to have little discriminative ability in the plotted PC subspaces, although events tend to be found mostly at larger PC2 absolute values. Based on Figures 4.5p,q,r,x, it is apparent that ARI exceedances occur predominantly during times of high PC4 values. From Figure 4.5s2, PC4 is characterized by northwesterly flow throughout the period, high CAPE and APCP, and low CIN and T2M. Events also occur predominantly in the high PC5 subregions of the corresponding subspaces, another high APCP PC also characterized by southeasterly flow, high CIN, and low CAPE. It is also of note that, in general, there is better discrimination between events and non-events than somewhat and quite extreme events, as evidence by the abundance of yellow and blue pixels in the subspaces and relative scarcity of cyan and magenta pixels. With regards to the Day 3 subspaces, some interesting apparent correspondences between the leading PCs can be noted; at high PC1 values (warm-season), values of PC2, 3, and 4 tend to also be high, while PC6 values tend to be low. Some distinction between levels of event severity can be seen; for example, in Figure 4.5d, in the high PC1, low PC4 subregion of the subspace, 1-10 year ARI exceedances are common relative to the 10+ year ones, while in the high PC1, high PC4 subregion, more extreme events are more common. Some of the best 2-D subspace distinctions are actually seen in the lower PC subspaces, particularly in the PC3, PC4 subsapce of Figure 4.5n, and the PC4, PC6 subspace shown in Figure 4.5t. For Day 3, non-events dominate the negative areas of PC4, while almost all events are found in the positive PC4 subspace. Physically, comparing Figure 4.5s1 with 4.5s2, one observes that

PC4 has a very similar signature in the native space in the Day 2 and Day 3 versions of the model. In Figure 4.5t, a sharp distinction between exceedance events and non-events is apparent, with non-events seen on the negative side of a diagonal separating the high PC4, high PC 6 subregion from the low PC4, low PC6 part. ROCK Day 3 PC6 (Fig. 4.5y1) is characterized by high precipitation and moisture, but low temperature and CAPE.

For extreme precipitation events in the NGP region (Fig. 4.6), the seasonal cycle as represented in PC1 is even more predictive of preponderance of ARI exceedance events than in the ROCK region, with blue and magenta tiles dominating the positive PC1 subregions of the subspaces and yellow and orange tiling covering the negative PC1 subregions for both the Day 2 and Day 3 models in Figures 4.6f,k,q,u and b-e, respectively. All the remaining PCs appear much less predictive, with many intermingled tile colors seen throughout the subspaces. There is some tendency for exceedance events at Day 3 to be found in negative areas of PC5, which is itself characterized by low APCP throughout, a shift from westerly to easterly winds, and a shift from southerly to northerly wind (Fig. 4.6m1). There are also some

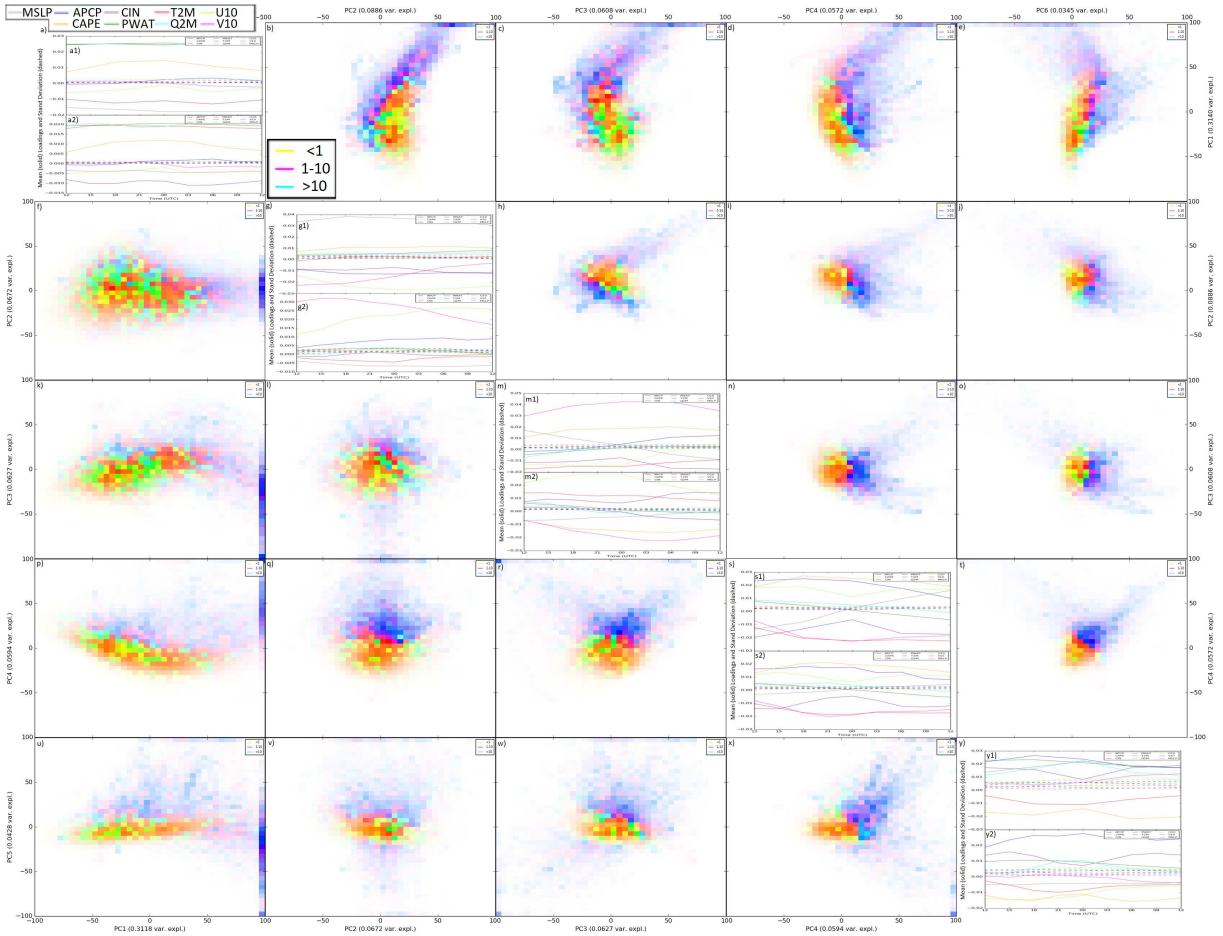


FIG. 4.5. Information about principal components for the CTL_PCA model for the ROCK region, and their joint relationships to ARI exceedances. PCs shown are according to the axis labels, where the number corresponds to the rank, in descending order, in terms of fraction of variance explained between forecast-point relative time series progressions in the atmospheric variables of the CORE predictor group. The fraction of variance explained by each principal component is indicated along the axis labels. PCs displayed are a subset of the total number used; those shown are the PCs identified as most predictive in the RF FIs, shown in Figure 4.15. The panels below the diagonal—panels (f), (k), (l), (p), (q), (r), (u), (v), (w), and (x)—show distributions of ARI exceedance events and non-events in the various 2-D PC subspaces, as indicated by the outer axis labels, for the Day 2 version of the model, while panels above the diagonal—panels (b), (c), (d), (e), (h), (i), (j), (n), (o), and (t)—show distributions in the subspaces for the Day 3 version. Within each of these panels, pixel color is used to indicate the distribution of events within the respective subregion of the subspace. Pure yellow indicates that only non-events (no 1 or more year ARI exceedances) occurred in the pixel's subregion over the period of record; pure cyan indicates that nothing but 10+ year ARI exceedances occur in the pixel's domain, and pure magenta indicates that all forecasts in the subregion over the period of record had associated observed 1-year ARI exceedances, with no forecasts in the subregion lacking a 1-year ARI exceedance or containing a 10-year ARI exceedance. Other colors indicate a blend of event observances, with those primary colors 'mixed' in accordance with the relative proportions; blues, for example, indicate a mix of 1-year exceedances and 10-year exceedances, while red indicates a mix of 1-year exceedances and non-exceedances. Pixels are also darkened according to the number of events within the subregion; dark colors indicate many events within the subregion, while light colors indicate few. Event classes are weighted so that pixel color is determined relative to the proportion of the total occurrences of the event class found in the pixel subdomain, rather than absolute event counts of the different event classes. Panels along the diagonal—panels (a), (g), (m), (s), and (y)—show the spatially averaged PC loadings time series across the forecast period for the various atmospheric fields used in the PCA process. The top of each of these panels—subpanels (a1), (g1), (m1), (s1), and (y1)—show the loadings time series for the Day 3 model, while the bottom half of each panel—(a2), (g2), (m2), (s2), and (y2), show loadings time series for the Day 2 version. For panels (a), (g), (m), and (s), the time series shown correspond with the PCs associated with the column labels and row labels of the corresponding pixel plots for the Day 2 and Day 3 model versions, respectively, while panel (y) displays the loadings of the associated labeled row and column for the Day 2 and Day 3 model versions, respectively. In these panels, solid lines correspond to the spatial mean loadings for the variable associated with the line color, as indicated in the panel legend, at the corresponding time, while the like-colored dashed lines show the spatial standard deviations for the same variable at the given time.

interesting nonlinear relationships in the subspaces; for example, in Figure 5w, exceedance events are found at relatively "extreme" combinations of values of PC9 and PC14, with non-events characterizing the more typical PC value subregion.

There is excellent correspondence between the five most predictive PCs in the PCST region (Fig. 4.7a,g,m,s,y) between the two forecast periods, with the immaterial exception of the PC2 sign flip (Fig.

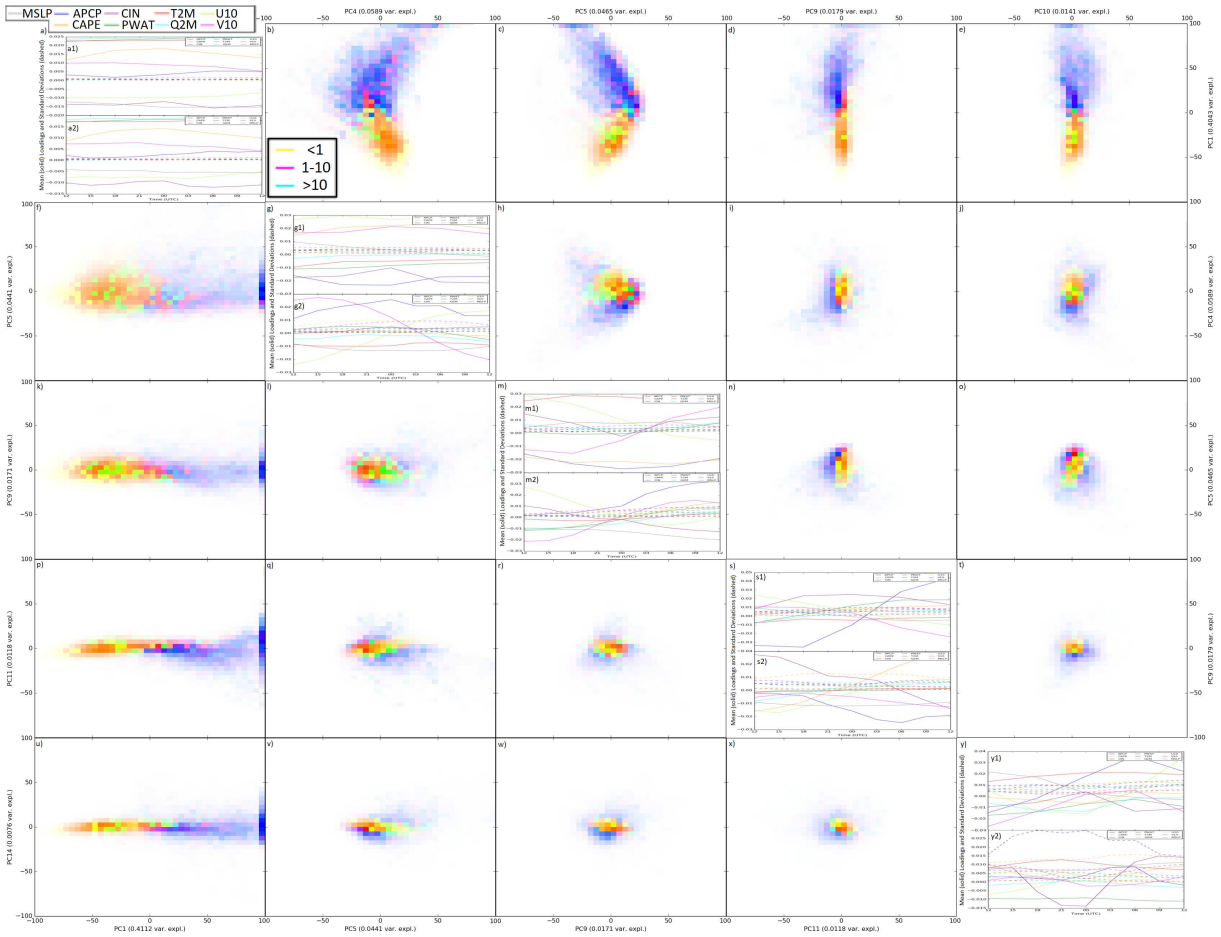


FIG. 4.6. Same as Figure 4.5, except for the NGP region.

4.7a). PC2, seen also in Figure 4.4, shows high APCP throughout the period corresponding with south-easterly surface wind, column moisture, instability, and low pressure and surface temperature. Notably, this PC is able to distinguish between 1- and 10-year events better than any of the subspaces from ROCK or NGP, particularly at Day 2 as evidenced by the large number of magenta pixels in Figure 6f. As suggested also in Chapter 3, this ability to discriminate between extreme precipitation severity levels may partially explain the higher forecast skill for the PCST region compared with the others; there is a smoother transition from non-events to severe extreme events from the low PC2, low PC3 to high PC2, high PC3 region of the subspace (Fig. 4.7f). For both periods, PC3/4 (Fig. 4.7g) is somewhat similar to PC2, with higher APCP going with low pressure and temperature, especially during the day. However, the kinematic fields are reversed, corresponding instead strongly to surface westerly flow. PC4/5 (Fig. 4.7m), somewhat surprisingly considering the region, is essentially just a an instability signal, with high values corresponding to high CAPE and low CIN. Inspecting the relevant panels of Figure

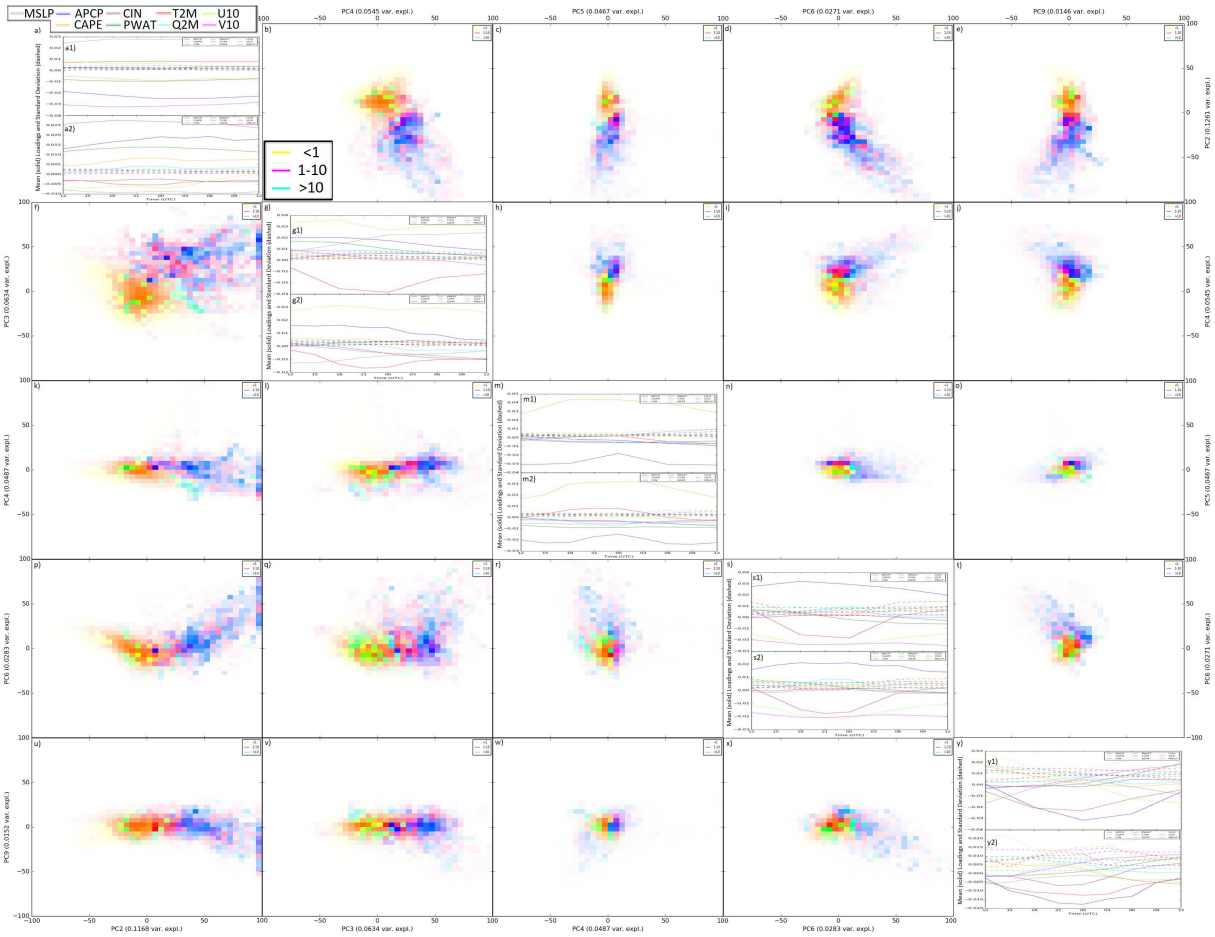


FIG. 4.7. Same as Figure 4.5, except for the PCST region.

6, it does not discriminate especially well in any of the subspaces compare with the prior two PCs, or with PC6 (Fig. 4.7s), which corresponds strongly with precipitation in northwesterly low-level flow, and low daytime temperatures coupled with relatively warmer nighttime surface temperatures. However, unlike the first two leading PCs presented (Fig. 4.7a,g), it is not associated with low pressure. The final PC examined, PC9 (Fig. 4.7y), has large spatial variability as evidence by large standard deviations in many of the fields—particularly surface winds—with positive values associated with a cold and dry signal, especially during the day, coupled with increasing stability and northeasterly low-level flow during the forecast period. Compared with the others, PC9 is not especially discriminative either, though most events occur with negative PC9 values coupled with high PC2 and PC6 values (Figs. 4.7e,t,u,x).

4.5 Results: RF Diagnostics

Associated with an RF is a single FI for each predictor. When no dimensionality reduction is performed in advance, there are thousands of GEFS/R predictors, each predictor associated with a particular atmospheric field, forecast hour, latitude, and longitude. In addition, there is a unique RF for each of the eight regions and each of the two forecast periods. Effectively visualizing and interpreting all of these FIs can be difficult. In order to manage the visualization task, RF FIs are first presented by considering only one dimension of predictor variability at a time. For example, FIs are considered as a function of the atmospheric field associated with the predictor, without regard to the hour or forecast point-relative location of the predictor. FIs are then considered by grouping all predictors with the same forecast hour, and lastly by grouping predictors with the same forecast point-relative location. This allows tractable visualization of a summary of the FI output of the GEFS/R, and helps identify areas for more detailed analysis of a subset of "raw" (single predictor) FIs presented in the second half of this section. For the interested reader, the full set of RF FIs are included in an online supplement to this paper.

GEFS/R QPF, or APCP, is reliably identified as one of the most predictive atmospheric fields for observed extreme precipitation based on RF FIs summed over space and time for each region of the Day 2 version of the CTL_NPCA model (Fig. 4.8). This indicates that the dynamical model, in this case the GEFS/R, has some skill in directly simulating extreme precipitation. However, the extent of model APCP being predictive over other ingredients-based fields varies substantially by region. In the PCST region, where extreme precipitation events are predominantly driven by atmospheric rivers and other large scale systems advecting moisture over orography (e.g. Rutz et al. 2014; Herman and Schumacher 2016a), a convection-parameterized model such as the GEFS/R is able to adequately simulate the largely stratiform precipitation processes. This is reflected in the RF FIs shown in Figure 6e; the model APCP, which adequately captures the processes involved in producing most precipitation events in the region, has a total FI of approximately 50% of the total, more than five times that of any other field. In other regions which feature a mix of synoptic and convective events, such as ROCK, NE, and SE (respectively Fig. 4.8a,d,h), APCP is still by far the most important atmospheric field in predicting observed APCP, but to a much smaller degree than in the PCST region with values in the 0.25–0.4 range. In the regions where extreme precipitation events are most driven by convective scale processes unresolvable by the GEFS/R and which correspondingly have the poorest verifying raw QPFs in predicting extremes (Herman and Schumacher 2016a), such as NGP and MDWST (Fig. 4.8b,c), model APCP is not

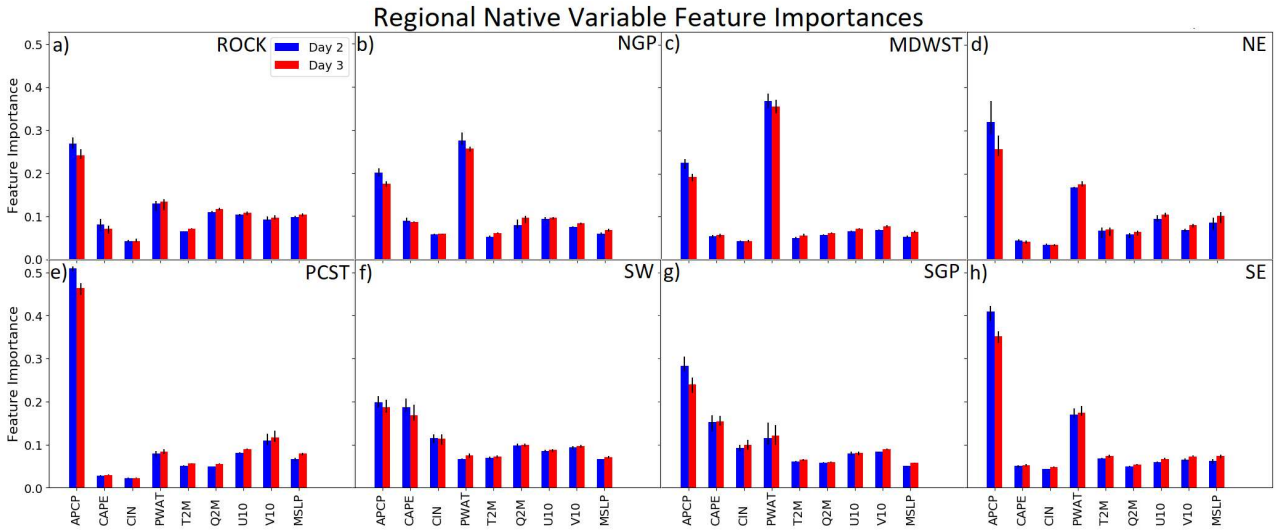


FIG. 4.8. Regional comparison of the summed RF FIs for the different atmospheric fields used in the CTL_NPCA model, summed over the time and two spatial dimensions. The blue bars correspond to the mean summed feature importances of the four models trained via cross-validation for the Day 2 version of the model; red bars correspond to the Day 3 model version. Error bars indicate the minimum and maximum cross-validation summed FIs. Panels (a)–(h) correspond respectively to ROCK, NGP, MDWST, NE, PCST, SW, SGP, and SE regions.

even the most important atmospheric field in predicting ARI exceedances. While still somewhat important, with aggregate RF FIs of approximately 0.18, APCP is identified as less predictive than PWAT in these two regions, with PWAT FIs in the 0.25–0.35 range. One physical explanation is that where the GEFS/R is poor at predicting extreme precipitation events by virtue of an inability to resolve the responsible processes, ingredients such as column-integrated moisture become more useful predictive tools. PWAT remains a valuable predictor in other regions as well, with greater importances also observed in the ROCK, NE, SGP, and SE regions (Fig. 4.8a,d,g,h). In one region, the SW (Fig. 4.8f), surface moisture (Q2M) was considered more predictive than column-integrated moisture (PWAT), but this was not generally the case. In most cases, CAPE and CIN were the least predictive fields among those examined, but the SW region (Fig. 4.8f) was again a considerable anomaly, with CAPE and CIN being respectively the second and third most important fields, and CAPE FIs nearly equal to those of APCP. Regional RF FIs at Day 3 look largely similar to the Day 2 RF FIs, but some minor differences can be discerned. The APCP RF FIs are slightly lower in many of the regions, particularly in the eastern regions (Fig. 4.8d,h). In general, "ingredients"—fields other than the direct APCP from the GEFS/R—are relied on somewhat more at Day 3 compared with Day 2.

Time series of RF FIs shed insight into which times forecast guidance provides the most useful predictive information for the quantity of interest, in this case ARI exceedances, and can help identify systematic biases in the parent model's diurnal climatology of relevant processes such convective initiation. They can also provide insight into when particular information is of value—whether the information is useful as a precursor or concurrent with the actual precipitation. Every region exhibits broadly similar FI time series when aggregated over all variables (Fig. 4.9, red and blue lines), with importance minima at both 1200 UTC times—the beginning and end of the forecast period—and a maximum during the middle of the day. A combination of two reasons likely explain this pattern. First, the middle of the day, in the afternoon and evening hours, is typically the most convectively active and the period in which precipitation and heavy precipitation are most frequently observed (e.g. Stevenson and Schumacher 2014; Herman and Schumacher 2016a). Second, it is also, somewhat coincidentally, the middle of the forecast period, and thus forecast values at this time can be more representative of the period as a whole. In most regions, the difference between the minimum and maximum importance values for a given forecast time spans approximately a factor of two. There is also more variability in the time-dependent FIs, comparing for example the relative width of the red and cyan shaded regions in the panels of Figure 4.9 with the error bars of Figure 4.8. Perhaps the most important finding is that the FI time series partially reflect the diurnal climatology of extreme precipitation events in each region. FIs are higher later in the forecast period in regions where extreme precipitation events tend to occur in the evening and overnight, such as the MDWST and SGP (Fig. 4.9b,c,g) while regions where events tend to be more in the afternoon hours, such as the NE and SE (Fig. 4.9d,h) have a peak at 0000 UTC and a significant drop off in importance by 0600 UTC. While this is seen in the time series with all fields aggregated, it is especially pronounced when considering only the APCP FIs (Fig. 4.9, purple and maroon lines). While the APCP FIs follow the diurnal precipitation climatology specific to the forecast region, PWAT FIs maximize prior to the maximum APCP FIs, particularly in regions where PWAT is found to be predictive (e.g. Fig. 4.9b,c,g), sensibly indicating that the column moisture of the environments in which storms form is an important property for predicting locally extreme precipitation.

Compared with the time series of Figure 4.9, more stark regional contrasts are observed for FIs compared in space (Fig. 4.10). As would be naively assumed, some regions have an importance maximum near the forecast point, with decreasing importance with increasing distance from the forecast location. This is broadly true of the ROCK, PCST, SW, SGP, and SE regions (Fig. 4.10a,e,f,g,h). The other three regions—NGP, MDWST, and NE (Fig. 4.10b,c,d)—have an importance maximum well downstream

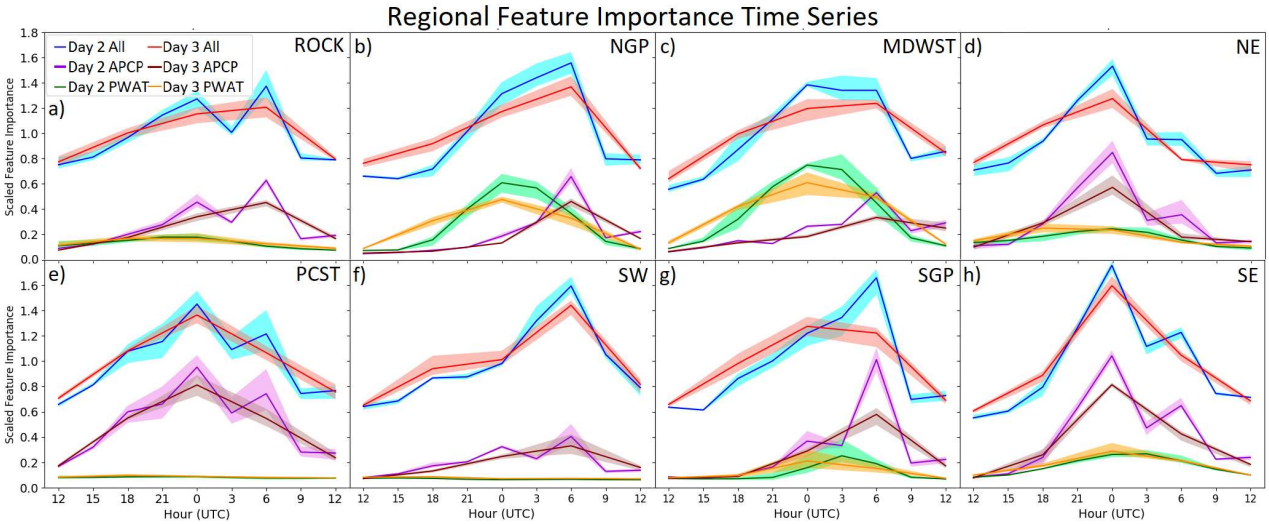


FIG. 4.9. Regional comparison of the summed RF FI time series in the CTL_NPCA model. Blue and red lines depict respectively the Day 2 and 3 versions of the model, summed over both spatial dimensions and all atmospheric fields. Values have been renormalized based on the number of time periods for the version of the model so that the *a priori* expected importance for each time is unity. The purple and maroon lines depict the Day 2 and Day 3 FI time series for only the APCP predictors, summed over the two spatial dimensions. Green and yellow lines are as with the purple and maroon lines, respectively, except for the PWAT FIs. The same normalization is applied to these time series as well, leaving *a priori* expected summed FIs of unity divided by the number of atmospheric fields (9). Shading about each line indicates the range of values obtained through the four folds of cross-validation, with the lines themselves representing mean values of the four folds. Panels (a)–(h) correspond respectively to ROCK, NGP, MDWST, NE, PCST, SW, SGP, and SE regions.

of the forecast point. This summary view does not provide insight into the precise physical reasons why this may be the case; possible causes include a combination of precipitation features moving too quickly, progged systems developing too far downstream, or that the downstream environment is simply better predicted than the environment in which the extreme precipitation events occur, and thus serves as a better predictor than the fields collocated with the forecast point. More investigation into possible reasons will be discussed below. Several other interesting regional differences may be noted. Some regions, such as PCST and SW (Fig. 4.10e,f) have a highly concentrated spatial maximum—with differences in importances between forecast points spanning nearly an order of magnitude—meaning that information from a particular location is much more predictive than surrounding areas. This likely indicates both increased persistence and consistency of model biases in these regions, and enhanced predictability overall as well, consistent with the higher forecast skill in these regions shown in Chapter 3. It also suggests that the RF is likely tracking specific simulated GEFS/R precipitation features in these regions, as opposed to just predicting based on a general characterization of the environment in which

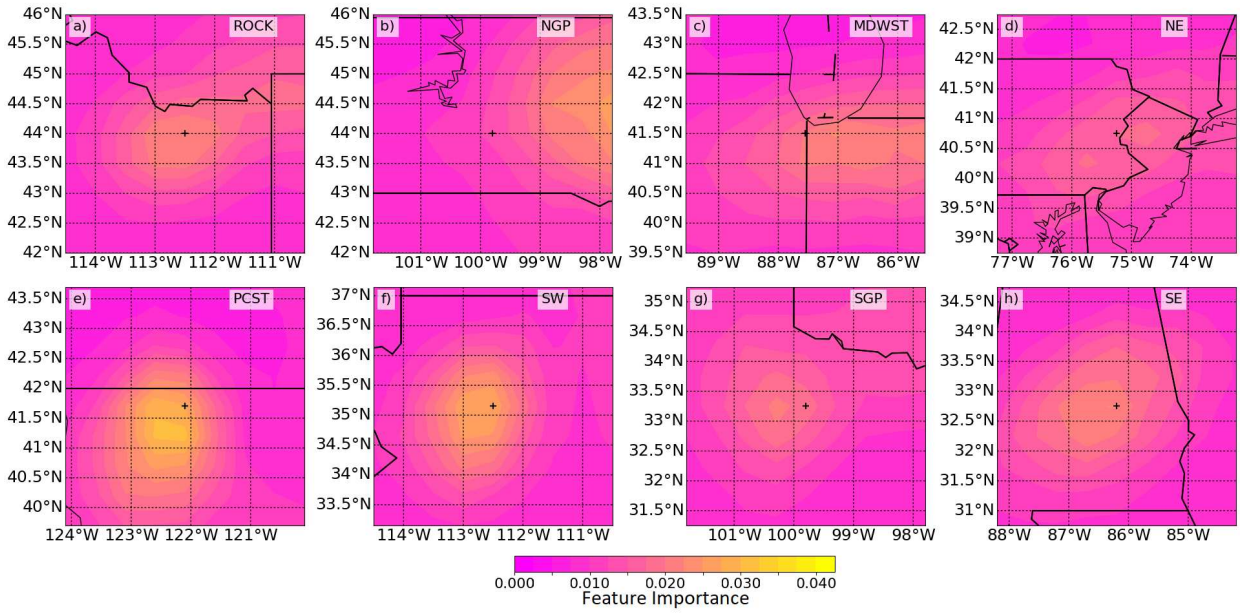


FIG. 4.10. Regional comparison of the summed RF FIs for the Day 2 version of the CTL_NPCA model, summed over variable and time in the filled contours to give importances as a function of forecast point-relative location. Values presented correspond to the mean value obtained through four folds of cross-validation. Panels (a)–(h) correspond respectively to ROCK, NGP, MDWST, NE, PCST, SW, SGP, and SE regions. The intersection of thick black lines indicates the location of the forecast point within each panel; other locations correspond to displaced forecast point-relative locaitons. Maps are drawn with the region centroid at the center of each panel, with state outlines in black underlying the panel to provide quantitative sense of spatial scale. Uniform scales are used for each panel as indicated by the figure colorbars.

storms might form, which would yield more spatially homogeneous FIs. The five aforementioned regions with a maximum FI point near the forecast point also do not all have these two points exactly collocated. In the PCST region (Fig. 4.10e), the point of maximum importance is displaced slightly to the south and west of the forecast point. This is true to some degree in the SW and SE regions as well (Fig. 4.10f,h). Meanwhile, a slight north and particularly west displacement is seen in the SGP region (Fig. 4.10g). These displacements may indicate persistent biases in the portrayal of extreme precipitation elements and/or the ingredients responsible for them. In the ROCK, SGP, and SE regions, a secondary maximum well downstream of the forecast point is observed, in a pattern resembling that of the other northern regions. In the regions that do have a downstream maximum, either primary or secondary, the more western regions—ROCK, NGP, and SGP—have the maximum also displaced well to the north (and east), while the regions farther east such as MDWST and NE have the maximum to the south.

Raw FIs for the APCP field in the Day 2 version of the CTL_NPCA model (Fig. 4.11) reveal that, consistent with Figure 4.9, APCP importances increase to a daytime or evening maximum with importance minima at 1200 UTC, with the strength of the cycle varying by region. Because the accumulation interval lies outside the forecast period for the front end 1200 UTC time, the importance is identified as the lowest there, compared even with the 0600–1200 UTC QPF at the end of the forecast period. Correspondingly, in some regions (e.g. Fig. 4.11b1,c1) there is a lack of a clear, cohesive precipitation feature—as represented by an importance maximum—at the beginning of the forecast period. However, at this time or subsequent to it, a clear importance maximum in the precipitation field emerges in each region and can be seen to track from west to east across the forecast point-relative domain throughout the forecast period, tracking the typical progression of precipitation systems with the mean upper-level flow. At the beginning of the forecast period, FI maxima (Fig. 4.11, column 1) are located 1–2 grid points west of the forecast point, while by the end (column 5), they are located anywhere from 0–3 grid points displaced to the east, with meridional alignment in the PCST region (Fig. 4.11a) and far eastern displacement on the five easternmost regions (e.g. Fig. 4.11b–d). This may be diagnosing regional climatological differences in the progression speed of extreme precipitation-producing systems, which may remain relatively stationary over the complex terrain of PCST, while moving quickly over the flatter terrain farther east. But another important factor that it may be identifying are model biases in the progression of extreme precipitation systems; it may be noting that GEFS/R systematically moves systems in the east too quickly, and systems in PCST perhaps too slowly, resulting in APCP well downstream of the forecast point being predictive of extreme precipitation in the eastern regions in a way that it is not in the western regions. More investigation is required to diagnose the extent to which each of these factors is in play in yielding this end diagnosis. Of further interest is the different progressions of FI maxima across different regions. In the five regions east of the Rocky Mountains—NGP, MDWST, NE, SGP, SE (e.g. Fig. 4.11b–d)—a clear southwest to northeast progression is seen, and is particularly pronounced in the SGP region. The regions meridionally aligned with the Rocky Mountains, ROCK and SW, have little latitudinal variation with time, though a slight southwest to northeast is observed in ROCK and slight northwest to southeast observed in SW (see online supplement). PCST, in contrast to most of the other regions, has a clear northwest to southeast temporal FI progression (cf. Fig. 4.11a1, 4.11a5). These progressions are consistent with the typical synoptic flow of locally extreme precipitation environments of these regions. The southward progression of post-landfall atmospheric rivers warrants further investigation, but is consistent with some previous studies (e.g. Ralph et al. 2010),

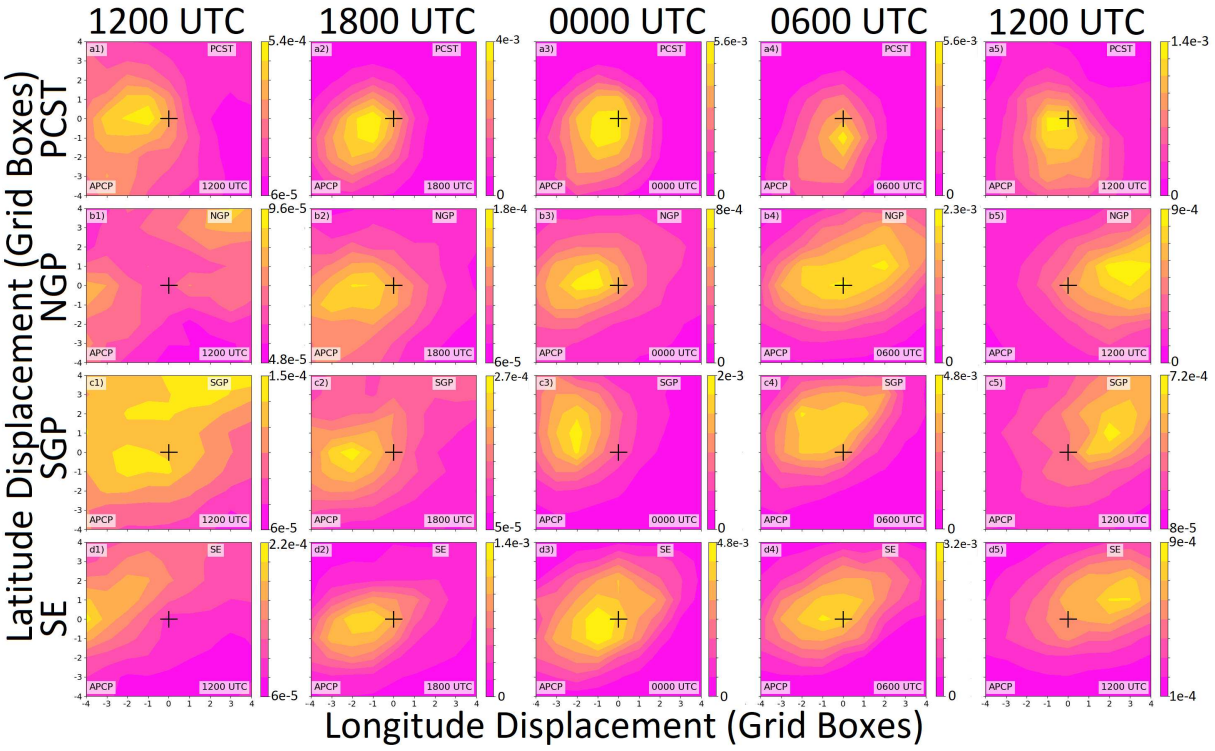


FIG. 4.11. Regional comparison of RF FIs for the APCP field spatially relative to the forecast point at different forecast times in the Day 2 CTL_NPCA model. Rows (a)–(d) correspond respectively to PCST, NGP, SGP, and SE regions. Columns (1)–(5) correspond respectively to forecast integration hours of 36 (1200 UTC), 42 (1800 UTC), 48 (0000 UTC), 54 (0600 UTC), and 60 (1200 UTC). Values depict the mean FIs obtained through the four folds of cross-validation. Note that the scale varies between panels; increments between colors are uniform for each colorbar.

and the southwest-northeast progression in the northeast is consistent with both tropical cyclones, which are almost always progressing poleward after landfall, as well as synoptically-driven mesoscale systems.

Of additional note are the latitudinal displacements of FI maxima. Some regions, such as NGP and particularly SGP (Fig. 4.11b,c), have a persistent northward displacement of FI maxima relative to the forecast point; this is likely associated with the well-documented northward displacement bias of mesoscale convective systems in convection-parameterized models (e.g. Grams et al. 2006; Wang et al. 2009), including the GEFS/R, which are also responsible for many of the ARI threshold exceedance events in these regions. In contrast, a persistent southward FI displacement is seen in the PCST and to a lesser extent in the SW (Fig. 4.11a and Herman and Schumacher (2018a) supplement). This could perhaps be associated with a less documented displacement bias of atmospheric rivers and other agents responsible for extreme precipitation in these regions (e.g. Wick et al. 2013). The FIs for the Day 2 and

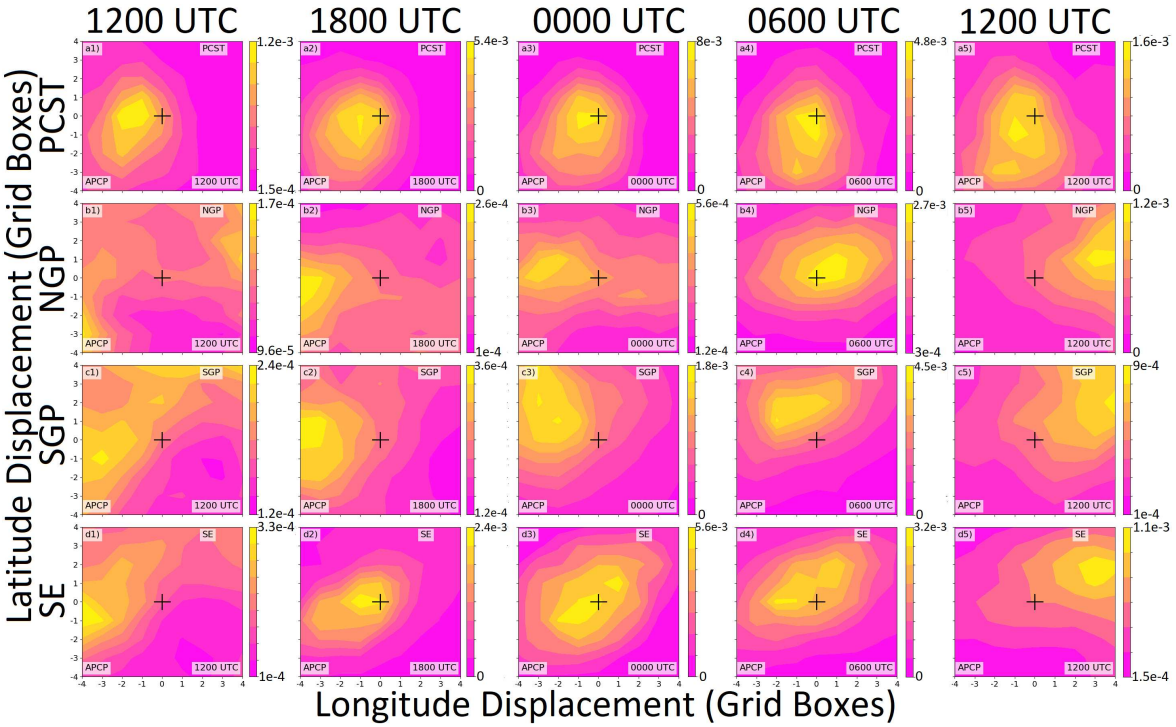


FIG. 4.12. Same as Figure 4.11, except for the Day 3 model version.

Day 3 versions of the model are largely quite similar (cf. Fig. 4.11 and 4.12). Many of the biases and/or displacements noted in the Day 2 RF FIs remain to varying degrees in the Day 3 FIs. Some differences appear to become slightly more pronounced, such as the west-east progression differences among the regions, with the PCST (Fig. 4.12e) shifting slightly farther west and SGP and others farther northeast, particularly at the end of the forecast period (e.g. Fig. 4.12g). The most pronounced difference is the general broadening of FI maxima, likely in association with increasing error and uncertainty associated with larger forecast lead times. This is suggestive of a gradual transition in trained RFs with increasing forecast lead time from bias-correcting a cohesive precipitation system simulated by the GEFS/R to predicting based on a more general characterization of the mean environment. This can be more concretely confirmed in future work by examining a wider spectrum of forecast lead times.

Interestingly and somewhat surprisingly, the PWAT FIs, shown in Figure 4.13, exhibit a much different signature than the APCP FIs. In many regions such as NGP and SGP (Fig. 4.13b,c), the highest PWAT FIs are located well downstream of the forecast point throughout the period. In some of these regions, such as the NE and SW (online supplement), there is an emphasis to the east and southeast of the forecast point, whereas in others, like NGP and SGP (Fig. 4.13b,c), the northeast corner is favored.

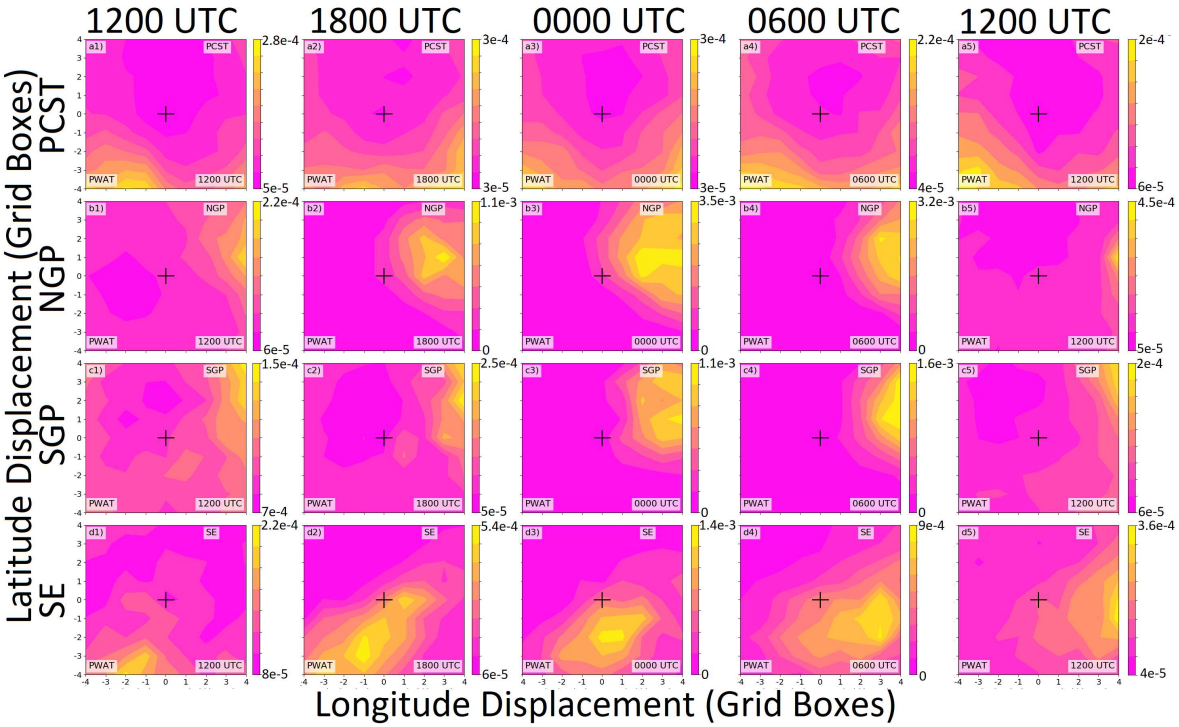


FIG. 4.13. Same as Figure 4.11, except for the PWAT field.

In some cases, the highlighted more important portion of the domain appears to correspond to the favored moisture source for precipitation systems in the region, such as the Atlantic Ocean in NE or the Gulf of Mexico for the SW. This is also the case for PCST (Fig. 4.13a), which has a persistent emphasis of importance well to the south of the forecast point; here, atmospheric rivers advect tropical moisture from the south and west. A couple of regions, in particular the SE (Fig. 4.13d), have a PWAT FI west-east progression like is seen for the APCP FIs in those regions. However, the PWAT FI maxima remain well to the south of the APCP FI maxima (cf. Fig. 4.11d, Fig. 4.13d), again likely capturing the source from which extreme precipitation producing systems develop.

For the CTL_PCA model, this sort of analysis is not possible due to the transformation from feature extraction during pre-processing. However, analogous interpretation can be made through collective diagnosis of the PCs (e.g. Figs. 4.1–4), the relationship between the PCs and the predictand, and the FIs of the PCs themselves (Fig. 4.14). FI tends to decrease with increasing PC number, suggesting a correspondence between the proportion of variance in the native dataset explained by the given PC—which in turn determines its number—and the predictive ability of the PC. However, this is not uniformly the case. Every region, with the partial exception of the NGP region (Fig. 4.14b), has "spikes" in FI whereby a particular PC is identified as considerably more predictive than surrounding PCs that explain similar

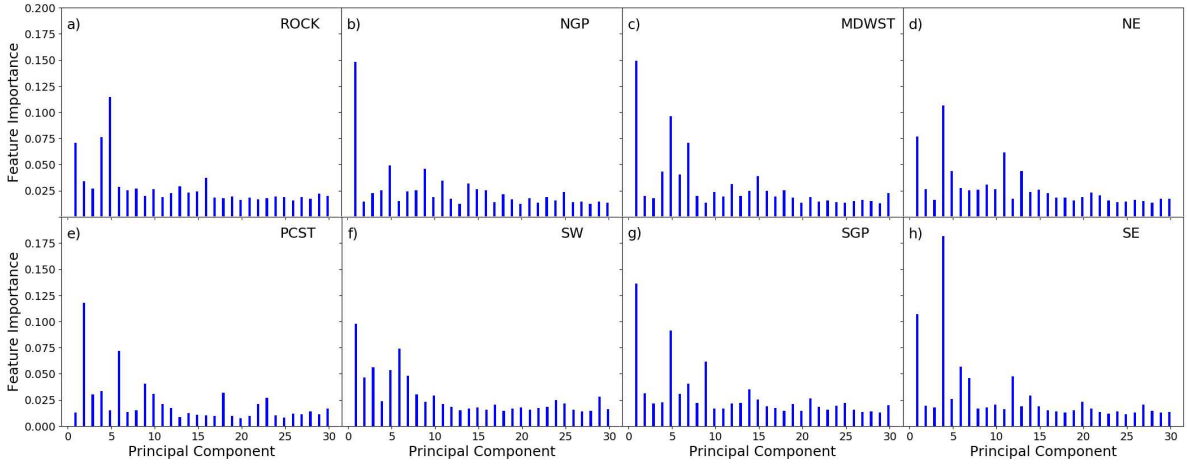


FIG. 4.14. Regional comparison of raw RF FIs for the Day 2 version of the CTL_PCA model, shown in descending order of PC variance explained for the 30 leading PCs. Importances of background predictors exist, but are omitted from this figure. Panels (a)–(h) correspond respectively to ROCK, NGP, MDWST, NE, PCST, SW, SGP, and SE regions. The scale is uniform between panels.

underlying variance. These FI maxima occur at different PC numbers depending on the region, typically somewhere between PC2 and PC15. In some regions, the first PC, which embodies the seasonal cycle (e.g. Fig. 4.1,4.2), is by far the most predictive PC (e.g. Fig. 4.14b,c). On the other side of the specturm, in PCST (Fig. 4.14e), the leading PC is no more predictive than much higher-numbered PCs. In other regions still such as ROCK, NE, and SE (Fig. 4.14a,d,h), PC1 is among the most predictive, but there is at least one other PC that is more predictive despite explaining less variance of the underlying forecast data. One such example is PC4 for the SE region (Fig. 4.14h), depicted in Figure 13. It is associated strongly with precipitation throughout the period (Fig. 4.15a), anomalous moisture, especially aloft (Fig. 4.15g), large CIN throughout the period (Fig. 4.15i), and low temperature and pressure (Fig. 4.15b,e). It is also associated with changing surface winds, from southeasterly winds at the beginning of the period to northwesterly by the end of it, with strong spatial gradients in wind (Fig. 4.15c,f). As with PC2 in some regions, this again exhibits some properties consistent with frontal passage, such as drying and a switch to northerly flow advecting in from the northwest (e.g. Fig. 4.15f,g), and being a cool-season phenomenon (Fig. 4.15e), but other elements seem inconsistent, such as the lack of significant changes in temperature or pressure anomalies over the course of the period (Fig. 4.15b,e). With many different PCs, it can be difficult to consider all the native predictor–PC and PC–predictand relationships comprehensively, but inspection of the FIs of Figure 12 can help target which relationships

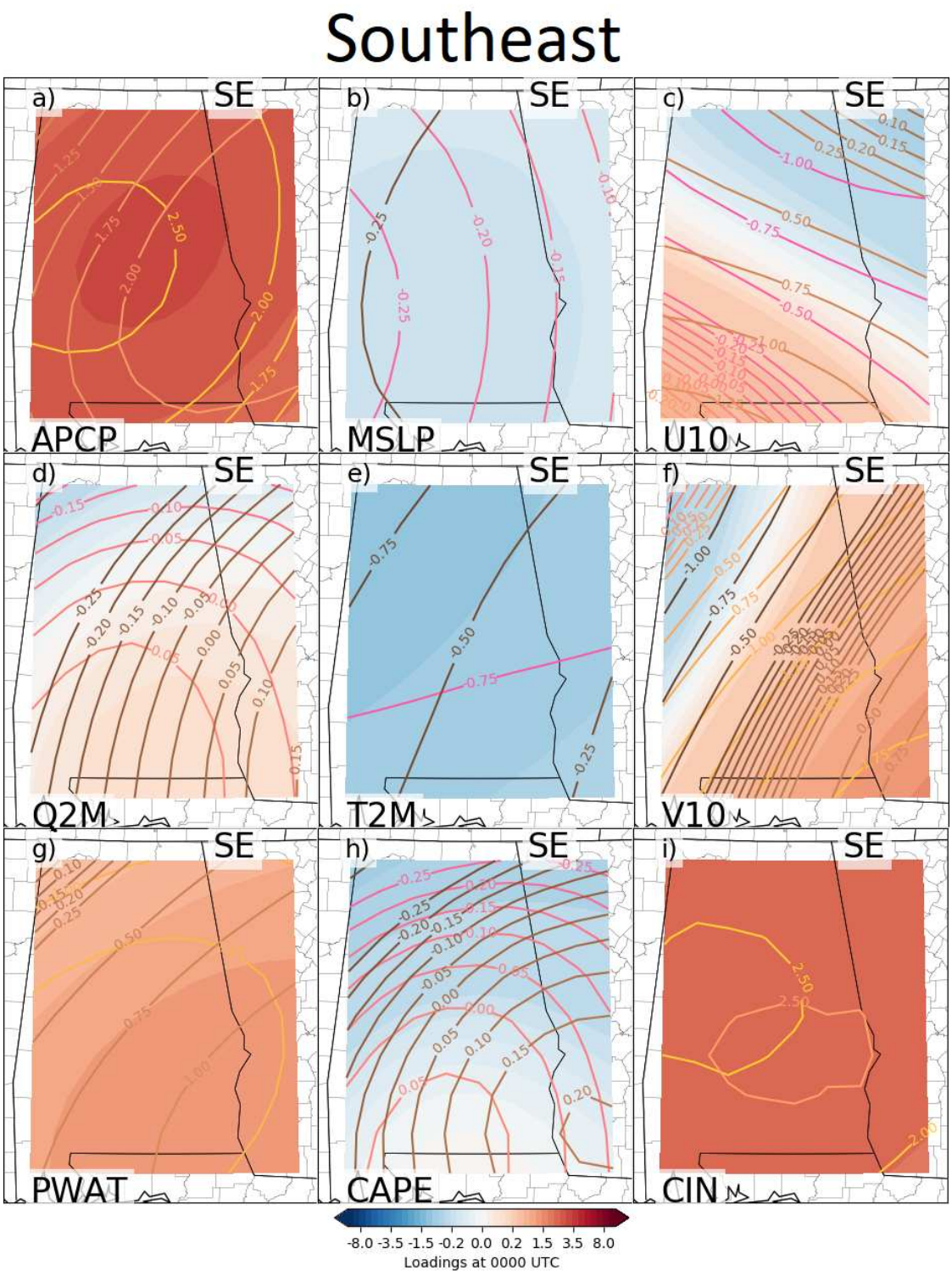


FIG. 4.15. Same as Figure 4.1, but for PC4 of the SE region.

are most useful to investigate. This also allows for improved understanding of how the RF algorithm operates.

4.6 Results: LR Diagnostics

In many cases, the CTL_LR identifies the same general findings as the RF-based models, just in a different capacity. One advantage of LR regression coefficients is that unlike RF FIs, they carry sign information in addition to just magnitude. Further, one can inspect coefficients for different event classes, in this case 1-year vs. 10-year ARI exceedances, separately. Though there are limitations to the quantitative interpretation of the transformed regression equations, such as those for the NGP region in Figure 4.16, they do still identify some important features. For the APCP field (Fig. 4.16a), positive coefficients unsurprisingly dominate throughout both space and time, with the one exception of the upstream side of the domain at the front end 1200 UTC (Fig. 4.16a1), which actually corresponds to the 0600–1200 UTC QPF from before the start of the forecast period. But two other aspects are worthy of note. First, the coefficient maxima track the expected precipitation from the upstream to downstream side during the period, and the most positive coefficients are—like the FIs for the CTL_NPCA model—found displaced to the north of the forecast point; this is particularly evident at 0000 and 0600 UTC (Fig. 4.16a3,a4). Second, the coefficients are largest for the accumulations from 0000-1200 UTC, corresponding to the climatological peak of the diurnal cycle of extreme precipitation events in NGP (e.g. Stevenson and Schumacher 2014). Additionally, the same downstream PWAT FI maximum for the CTL_NPCA model (Fig. 4.13b) is reflected also in the CTL_LR model with positive coefficient maxima downstream of the forecast point throughout the forecast period (Fig. 4.16d); a similar phenomenon is observed with surface moisture (Fig. 4.16f). It is apparent also that anomalous southeasterly flow, particularly around 0000 UTC, increases the probability of extreme precipitation events (Fig. 4.16g3,h3). Anomalous surface easterlies promotes slower storm motion, and anomalous surface southerlies tends to yield continued moisture advection and enhanced storm maintenance (e.g. Doswell et al. 1996). Extreme precipitation event probabilities also increase with low pressure at the beginning of the period (Fig. 4.16i1) increasing to anomalous high pressure by the end of it (Fig. 4.16i5). Many extreme precipitation events in the NGP region are associated with mesoscale convective systems or other training convection. Composites of these scenarios (e.g. Peters and Schumacher 2014) have shown synoptic low pressure, particularly to the south and west of the eventual MCS, in the pre-convective environment that moves out of the area or decays by the post-convective environment; this finding in the LR regression coefficients is consistent with those composites. Lastly, the regression coefficients somewhat counterintuitively indicate that 10-year 24-hour ARI exceedances in the NGP region are more likely with low daytime CAPE and high daytime CIN (Fig. 4.16b2,c2); this trend reverses by the end of the

forecast period (Fig. 4.16b5,c5). This perhaps suggests that highly extreme events can occur best when instability is not exhausted from isolated diurnal convection and is instead maintained for nocturnal mesoscale convective systems that are responsible for the majority of 10-year 24-hour ARI exceedances in NGP (e.g. Schumacher and Johnson 2006).

The coefficients for the SGP region (Fig. 4.17) are very similar, with 10-year exceedances associated with anomalous southeasterly surface flow (Fig. 15g3,h3), low increasing to high MSLP (Fig. 4.17i), and high surface and column moisture especially to the east and southeast of the forecast point (Fig. 4.17d,f). The APCP coefficients (Fig. 4.17a) are more spatially-uniform than in NGP and have their maxima more to the south of the forecast point rather than north. The relationship with CAPE is very weak (Fig. 4.17b), but high CIN (Fig. 4.17c) to the north of the forecast point is found to correspond with SGP extreme precipitation events. These latter three variables collectively tell a similar story to NGP coefficients, but there is a redistribution of coefficient values among the fields.

Some interesting coefficient differences are observed to the east in the SE region (Fig. 4.18). Anomalous easterly surface flow (Fig. 4.18g) over the domain is again found to be conducive to extreme precipitation events; this holds to an extent with anomalous surface southerlies as well, but the coefficient values (Fig. 4.18h) are very small. High moisture across the forecast point domain, both throughout the surface and especially throughout the column (Fig. 4.18d,h) are again found to correspond to extreme precipitation events in the region. Low pressure (Fig. 4.18i) and temperature (Fig. 4.18e) tend to be positive indicators of locally extreme precipitation events. Unlike the Great Plains regions, the CAPE and CIN relationships (Fig. 4.18b,c) are more as expected in association with a more diurnally-tied precipitation climatology in the SE region, with forecasted high CAPE and low CIN during the afternoon increasing the likelihood of an extreme precipitation event. Interestingly, though high APCP corresponds with increased event probability (Fig. 4.18a), there is little temporal continuity of the spatial structure. What does appear to be one or the most significant indicators, as evidenced by the magnitude of the regression coefficients, is APCP to the north of the forecast point the night prior to the start of the forecast period (Fig. 4.18a1), which also leaves high CIN to the north of the forecast point to start the period (Fig. 4.18c1). This may perhaps act to favorably precondition the environment at the forecast point.

The PCST region regression coefficients (Fig. 4.19) yield some unusual and interesting findings that may warrant further investigation. Unlike other regions, many fields exhibit complex coefficient spatial structures, with numerous changes in sign and other smaller features. As the CTL_NPCA model

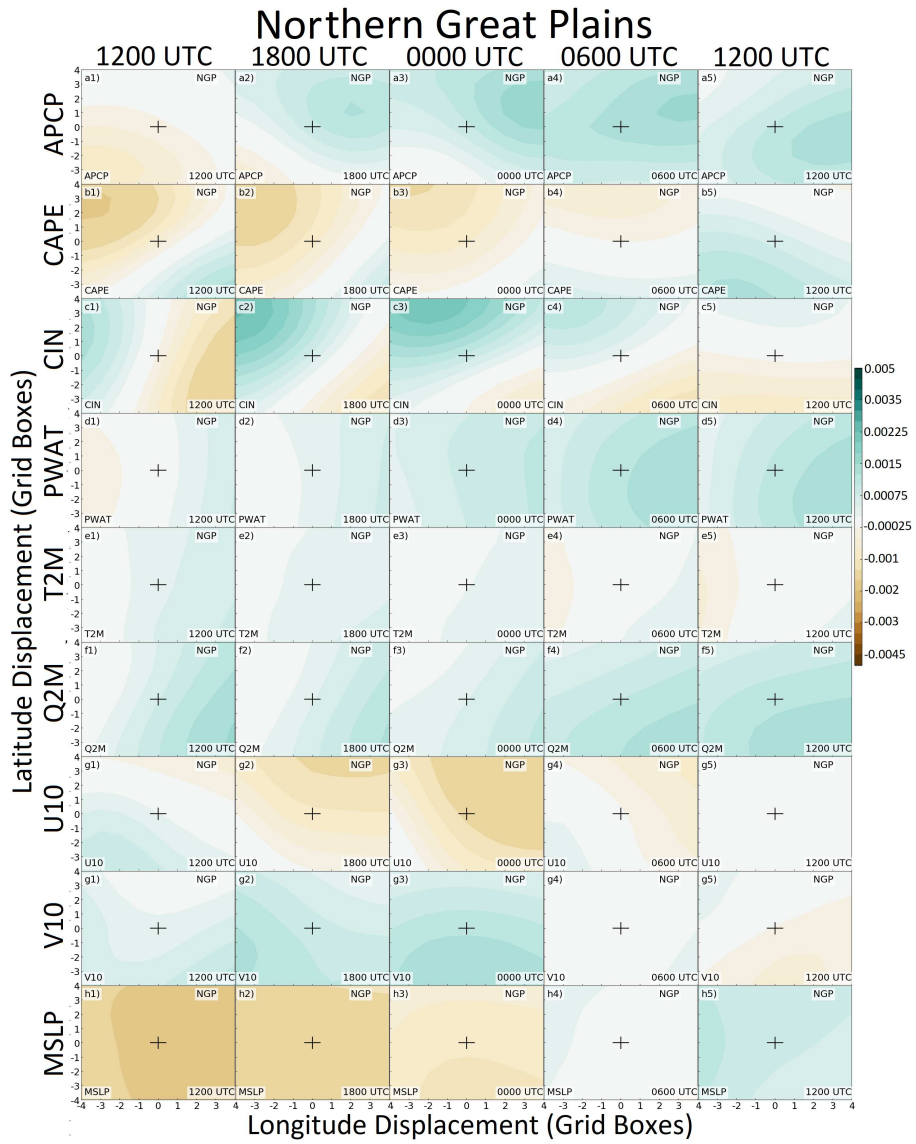


FIG. 4.16. Regression coefficients for the 10-year ARI exceedance equation for the NGP region obtained through logistic regression in the Day 3 version of the CTL_LR model, projected back into native variable space by means of the principal component loadings. Panels (a)–(i) (the rows) correspond respectively to the APCP, CAPE, CIN, PWAT, T2M, Q2M, U10, V10, and MSLP forecast fields. Subpanels (1)–(5) (the columns) correspond to coefficients at the following forecast times: 1200 UTC at the beginning of the forecast period, 1800 UTC during the period, 0000 UTC, 0600 UTC, and 1200 UTC at the conclusion of the forecast period. Green values indicate the anomalously positive values of the indicated field contribute positively to the forecast probability of an ARI exceedance, while browns indicate a negative contribution. The intersection of the thick black lines indicate the location of the forecast point in each panel, with other locations depicting coefficients at spatially displaced locations.

identified (Fig. 4.8e), the CTL_LR model also identifies GEFS/R APCP as by far the most predictive field of PCST extreme precipitation events, as evidenced by the largest regression coefficients in the model

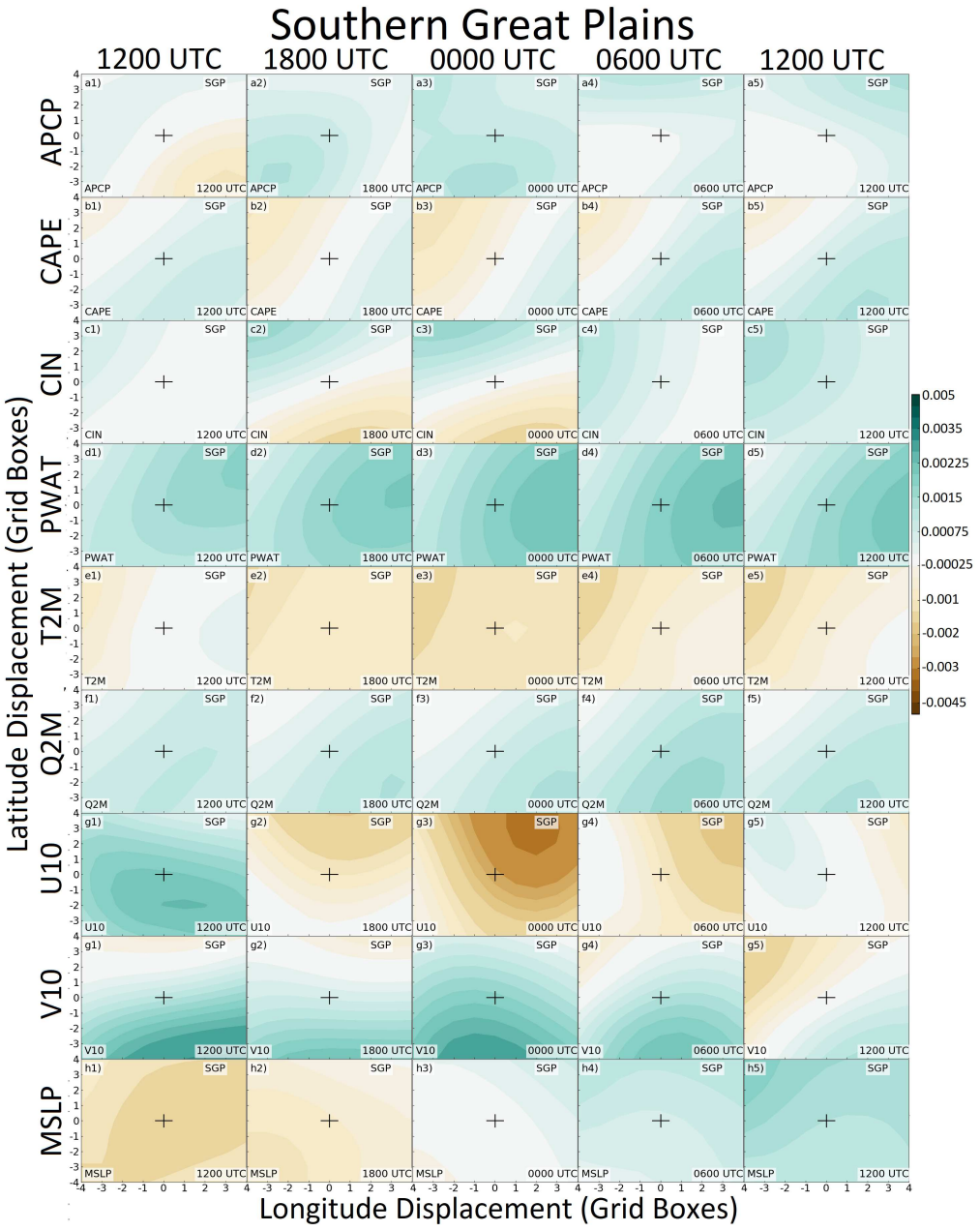


FIG. 4.17. Same as Figure 4.16, except for the SGP region.

occurring in Figures 4.18a2 and a3. Also like the CTL_NPCA model, which found maximum APCP FIs to the south of the forecast point (e.g. Fig. 4.11a4), the same is seen in the CTL_LR coefficients (Fig. 4.19a2,a3). Much of the rest of the signal may be somewhat muddled because events occur most frequently in association with atmospheric river events, and these bring anomalously warm and moist conditions during the cold-season. These tend to offset, leading to weaker coefficients in thermodynamic fields. But for many of these fields (e.g. Fig. 4.19d,e,f), to the extent these coefficients may be

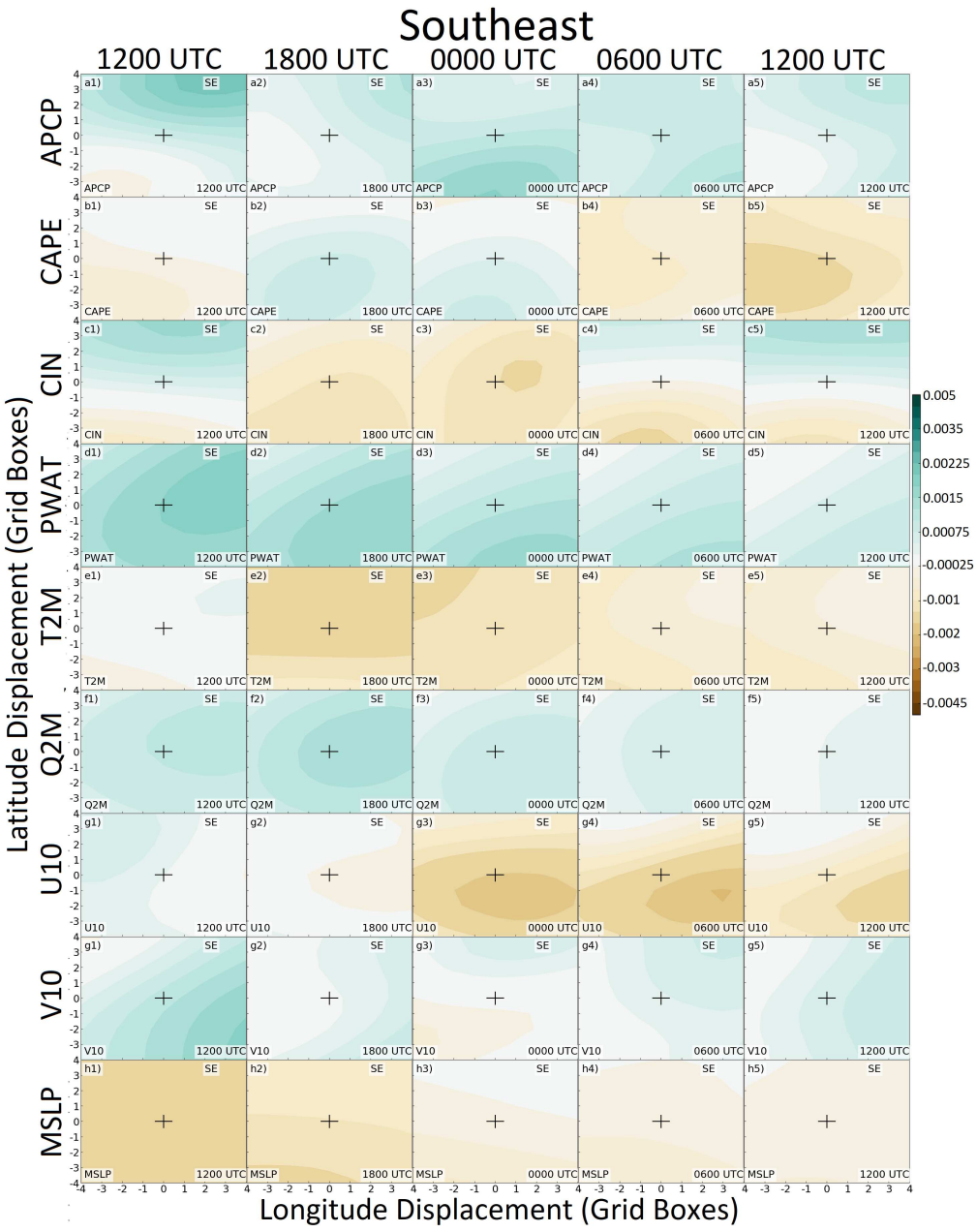


FIG. 4.18. Same as Figure 4.16, except for the SE region.

directly interpreted, low temperature and moisture at the forecast point in a surrounding environment of higher temperature and moisture tend to positively associate with extreme precipitation events in the region. This may seem rather counter-intuitive, but there is some physical basis for these coefficients. In the far-field, the coefficients are consistent with large-scale advection of warm, moist air over the domain, as evidenced by the increasingly positive temperature and moisture coefficients in Figures 4.19d, e, and f, and particularly PWAT (Fig. 4.19d). That column-integrated moisture is most

strongly influenced is consistent with an atmospheric river signature, where moisture is transported at mid and upper levels and not just near the surface. But near the forecast point, where it is precipitating in the model (e.g. Fig. 4.19a3), there is a local minimum in temperature and moisture (Fig. 4.19d3,e3), consistent with column moisture condensing and precipitating out of the column; surface temperatures are likewise inhibited by a lack of radiational heating and perhaps diabatic cooling as well. Unlike the other regions, extreme events are also associated with anomalous westerly surface flow throughout the period (Fig. 4.19g) in this region. In other regions, easterly flow promotes slower storm motions; here, the westerly flow promotes upslope flow. Meridionally (Fig. 4.19h), events are associated with southerly flow transitioning to northerly flow during the forecast period, consistent with cyclone passage. Overall, some of the details of these findings may be somewhat surprising; given that, unlike most regions, the CTL_LR model had almost equal performance to the RF-based models, this may invite deeper investigation into these properties of the coefficients.

For the interested reader, coefficients associated with the Day 2 model, for unshown regions, and also for the 1-year ARI exceedance equations have been included in the online supplement to Herman and Schumacher (2018a).

4.7 Summary and Conclusions

Three models of different formulation from Chapter 3, each trained to forecast locally extreme precipitation across CONUS, are analyzed in depth to assess their internal operations and ascertain what insights, if any, they reveal about forecasting extreme precipitation from the GEFS/R model. One model, the CTL_NPCA model, uses raw GEFS/R fields as input to a random forest algorithm to generate its predictions. The second, CTL_PCA, also uses an RF, but performs dimensionality reduction via principal component analysis on the raw GEFS/R fields and supplies a reduced predictor set consisting of just a subset of retained leading PCs in lieu of the raw fields themselves. The last, CTL_LR, also performs the PCA pre-processing step, but rather than supply the retained PCs to an RF, they are instead supplied to a regularized logistic regression algorithm. It is shown that all of these models, many of which may appear highly abstract, can be readily visualized in different ways in order to understand their internal operations. Both the act of creating derived predictors in pre-processing via PCA and using non-parametric techniques such as RFs add layers of abstraction that make visualization and interpretation more challenging.

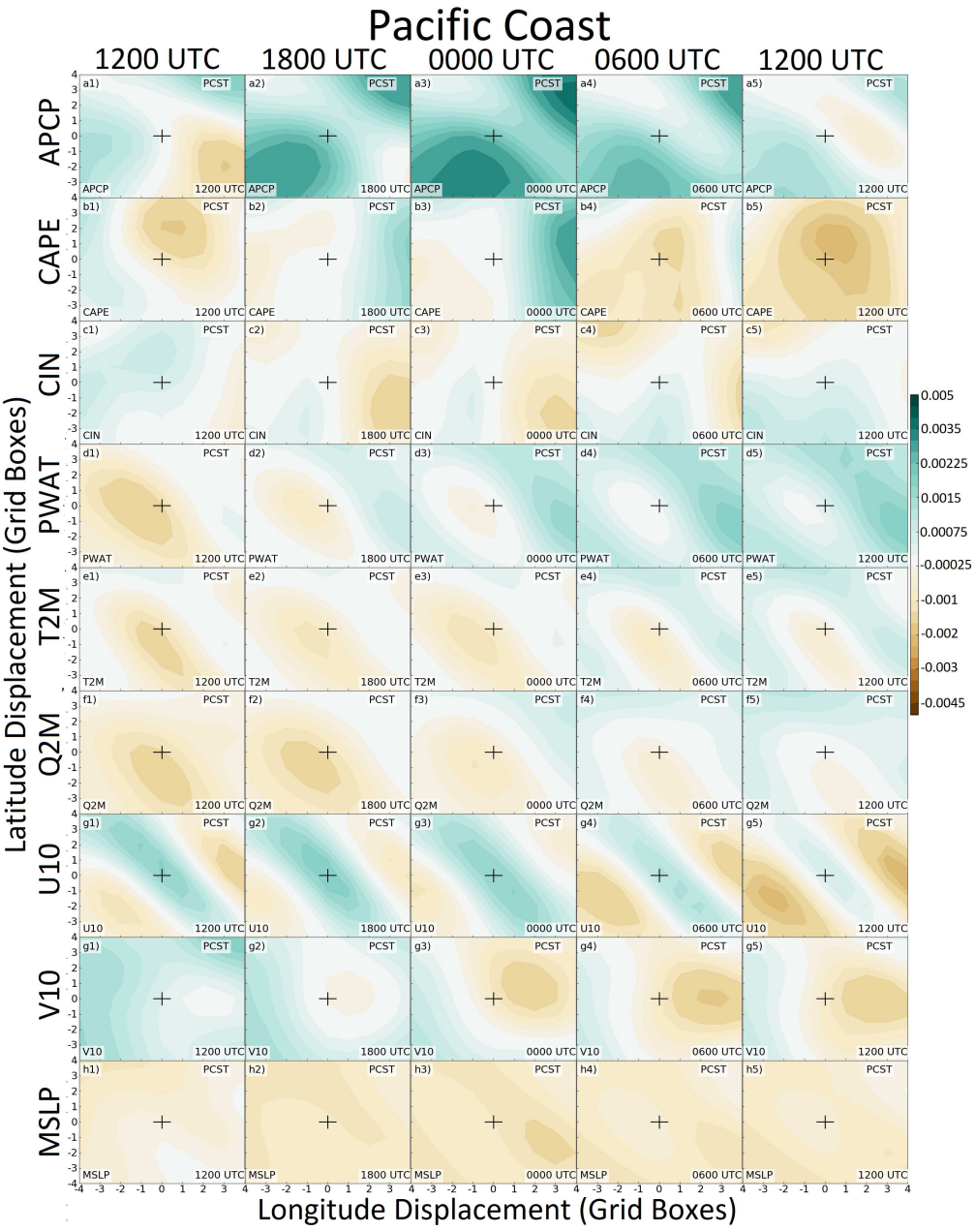


FIG. 4.19. Same as Figure 4.16, except for the PCST region.

Numerous aspects about forecasting locally extreme precipitation with global, convection-parameterized model output have been confirmed, while some new discoveries warrant potential further investigation. Both LR and RFs are able to identify what a human forecaster would expect to be the most predictive variables for extreme precipitation, with the largest regression coefficients and FIs generally identified for model QPFs—the direct prediction of the predictand from the GEFS/R. Moreover, the models further validate the findings of Herman and Schumacher (2016a) and

other studies that found the GEFS/R and like models with parameterized convection and relatively large horizontal grid spacing have better forecasts of extreme precipitation—and in fact better QPFs all-around—over the Pacific Coast of CONUS and the worst performance over the Great Plains and central CONUS. This is seen to an extent in comparing the APCP regression coefficients for the CTL_LR model, but is especially true of the FIs in the CTL_NPCA model, which exhibit by far the highest APCP FIs in the PCST region, and the lowest FIs in the NGP and MDWST regions. In fact, in the regions where extreme precipitation is most dominated by small-scale convective processes, such as NGP and MDWST (e.g. Schumacher and Johnson 2005, 2006), model QPF isn't even identified as the most predictive atmospheric field from the GEFS/R, with PWAT instead exhibiting the highest FIs. Similarly, while in most regions CAPE as portrayed in the GEFS/R is not identified to be a very predictive quantity for predicting locally excessive 24-hour precipitation, in one region, SW, where many extreme events are associated with isolated diurnally and orographically forced precipitation within monsoonal moisture, it was found to be almost equally predictive to the QPF itself. This framework and these models thus act to dynamically discern an appropriate "weighting" based on the hydrometeorology of the given region and the characteristics of the dynamical model from which the predictors are derived.

In time, the models again follow processes and focus examination dynamically depending on the region in ways consistent with how a human forecaster might approach the forecast problem. The APCP FIs follow the diurnal precipitation climatology in each region, with maxima late in the forecast period over the Great Plains and Midwest, and earlier peaks over the coasts. Environmental conditions such as PWAT maximize in importance prior to the APCP FI maxima, diagnosing the relevance of these environmental properties as antecedent storm conditions. In space, the algorithm tracks precipitation features through time and space from the west edge of the predictor domain at the beginning of the period to the east edge at the end of the forecast period. Some persistent displacement biases are also noted, with a northern displacement of the maximum APCP FIs relative to the forecast point in convectively active regions usch as NGP, SGP, and MDWST, in accordance with prior findings of mesoscale convective system displacement biases from convection-parameterized models (e.g. Grams et al. 2006; Wang et al. 2009; Clark et al. 2010), and a southern displacement in the PCST region, suggestive of a systematic southward displacement bias of atmospheric river events which dominate the extreme precipitation signal in the region. In aggregate, FIs are usually highest near the forecast point, though, especially in northern states west of the Rockies (NGP, MDWST, NE), the highest mean FIs are found

downstream of the forecast point. This is particularly true in the PWAT field, and the precise reasons for this identification require further investigation.

In the design of statistical forecast models, it is important to consider not necessarily just the skill of the raw model output, but the potential skill of forecasts issued by an experienced forecaster after considering the statistical model's output. If a forecast model is a complete "black-box", a forecaster will inherently be unable to use knowledge of likely errors in the inputs to improve the estimate of the outcome or relate the current forecast to past scenarios where both the forecast and outcome are known, among other techniques frequently adopted by human forecasters to produce a forecast more skillful than that generated by automated guidance. With a more transparent and comprehensible model forecast process, however, a forecaster may be able to improve upon the guidance in some situations using these sorts of corrections. Of course, if a "black-box" statistical model produces demonstrably and substantially superior forecasts to any competing guidance, it may well still outperform other less skillful models where the forecaster is able to add more value. However, as has been demonstrated in this study, machine learning algorithms including but not limited to RFs can provide forecasting insights that allow improved interpretability both of the output from the statistical model, but also reveal insights about the dynamical model that allow improved interpretation of the dynamical model guidance even absent any machine learning-based model guidance. Although machine learning can identify novel properties and relationships, it should be emphasized that it is not a panacea. The diagnostics presented herein do not directly identify physical reasons for its findings; while some may be readily apparent, others require further investigation to fully understand the identified patterns in these machine learning models. That said, with existing machine learning models demonstrating considerable skill in forecasting locally extreme precipitation as well as a host of other sensible weather phenomena, it is recommended that expected future forays into NWP with machine learning consider not only the properties of the raw forecasts that the developed models produce, but also the visualizability of the model construction and what physical insights and understanding may be gleaned from such visualization.

There are various concrete ways that these diagnostics may assist human forecasters, as well as help guide future research. Even absent using the statistical model output, these diagnostics can help a human forecaster better interpret raw dynamical guidance from the parent model—the GEFS/R in this case. For example, the diagnostics suggest that a forecaster should shift his or her area of highest excessive precipitation risk to the south of where the heaviest precipitation is portrayed over the Great

Plains and eastern regions of CONUS, while shifting to the north along the Pacific Coast. It also suggests that convective systems portrayed in the GEFS/R may be systematically too progressive, particularly in the NGP and MDWST regions—something that likely warrants further investigation. The diagnostics also help point forecasters at which fields to devote the most attention towards; in PCST, the GEFS/R's QPFs should be given considerable credence, while in NGP and MDWST, more attention should be paid to the GEFS/R PWAT field in trying to determine risk of locally extreme precipitation. The diagnostics presented in this paper provide some ability to modifying the statistical model output based on external assessments as well. For example, if a forecaster judges that the GEFS/R is much too dry aloft in a region, he or she may consult the regression coefficients and adjust probabilities accordingly depending on the sign of the PWAT regression coefficients for the region. For RFs, if PWAT FIs are very low, the forecaster can maintain confidence in the forecast, while if they are quite high, the forecaster may choose to discount the output from the machine learning model. Additional corrections may be identifiable by performing a detailed meteorology-dependent verification of the machine learning-based forecasts over an extended historical record. A start to this type of analysis was performed in Chapter 3; future work should further break down model performance by meteorological regime, analysis of which would provide even further aid to the human forecaster.

It is imperative for statistical modelers to investigate the internals of trained models to the extent possible. When performance is not appreciably degraded—and it certainly can be—it may in some instances be preferable to employ algorithms that are more easily interpretable, such as RFs in lieu of algorithms whose output is more difficult to visualize such as support vector machines or neural networks (e.g. Rozas-Larraondo et al. 2014). Additionally, while traditional PCA was applied because the orthogonality and maximum variance constraints were believed to be beneficial for model skill and yield desirable independence properties, their potential for less physically grounded components suggests applying more directly interpretable pre-processing instead, such as sparse PCA (Zou et al. 2006) or rotated PCA (e.g. Richman 1986; Mercer et al. 2012; Peters and Schumacher 2014) could yield more directly and easily interpretable statistical model results. Future work will seek to both further explore the comparison of machine learning algorithms for NWP in additional settings in addition to working to invent or apply improved methods for understanding what the machine learning informs us about the phenomenon of study.

CHAPTER 5

PROBABILISTIC VERIFICATION OF STORM PREDICTION CENTER CONVECTIVE OUTLOOKS

5.1 INTRODUCTION

Severe weather—defined as the presence of one or more tornadoes of any intensity, convectively induced wind gusts of at least 58 mph (93 km h^{-1}), or thunderstorms producing 1 in (2.54 cm) or larger hail—poses a substantial threat to life and property over much of the United States, and is collectively responsible for an annual mean of 137 fatalities and $4.69 billion in damages (NWS 2017c) over the past eight years. Outlooks and other forecasts from the Storm Prediction Center (SPC) are one of the leading sources of severe weather forecast information for National Weather Service (NWS) meteorologists, broadcast meteorologists, emergency managers, and the public. The SPC routinely produces and updates numerous products from nowcasts to forecasts with 8 days of lead time, and these forecasts are publicly archived—some as far back as 2003 (SPC cited 2017a). However, despite the substantial viewership and reliance of end-user communities on SPC products, the specificity and concreteness of their forecast predictands, their standing as a "gold standard" for severe-weather forecasting (e.g. Stough et al. 2010), and SPC's transparency in making available both their contemporary and historical forecasts, much remains unknown about the quality of their outlook products due to gaps in published verification of SPC outlooks. This study seeks to rectify this by performing quantitative verification of their forecast products and in particular their probabilistic convective outlooks.

As alluded above, SPC is responsible for the routine issuance of a wide variety of products, including their convective outlooks which are the focus of this study. Convective outlooks are produced for Day 1–Day 8, where the valid period for a forecast day spans 1200 UTC–1159 UTC, but the outlook specifics vary as a function of forecast lead time. For Day 1, each of the three severe weather elements—tornadoes, hail, and wind—are treated separately, while for Day 2 and beyond, all three are instead treated collectively. Outlooks include both categorical and probabilistic components; the latter use neighborhood probabilities whereby contours are drawn to define lines of constant probability of observing the given severe weather predictand within approximately 25 miles (40 km) of a point. For Days 1–3, categorical outlooks are also provided but are simply a strict function of those neighborhood probabilities. Day 1 tornado forecasts include 2%, 5%, 10%, 15%, 30%, 45%, and 60% neighborhood probability contours; Day 1 hail, Day 1 wind, and Day 2 and Day 3 aggregate severe forecasts employ a

different contour set: 5%, 15%, 30%, 45%, and 60%. Furthermore, for all Day 1–3 forecasts, a higher intensity level, "significant severe"—defined as a tornado of rating at least EF2, thunderstorms producing hail with diameters of ≥ 2 in (5 cm), or convective wind gusts ≥ 75 mph (65 kt, 33 m s^{-1}; Hales 1988)—is considered, and an additional "significant severe" contour is drawn for 25 mile neighborhood probabilities of significant severe weather ≥ 10%. Days 4–8 probability forecasts, also made collectively for any severe weather, use only two contours, 15% and 30%, and do not directly consider the elevated significant severe criteria. Day 1 forecasts are routinely produced at 0600, 1300, 1630, 2000, and 0100 UTC, with the 0600 UTC outlook being the first Day 1 outlook to cover a given valid period. Day 2 and 3 forecasts are disseminated respectively at 0100 and 0230 CT (with the UTC time varying based on daylight saving time); Day 2 receives an additional update at 1730 UTC (Hitchens and Brooks 2014; Edwards et al. 2015).

The majority of previous published severe-weather verification has focused on severe thunderstorm or tornado watch (e.g. Doswell et al. 1990; Anthony and Leftwich 1992; Doswell et al. 1993; Vescio and Thompson 2001; Schneider and Dean 2008) or warning (e.g. Polger et al. 1994; Bieringer and Ray 1996; Simmons and Sutter 2005; Barnes et al. 2007; Simmons and Sutter 2008; Brotzge et al. 2011; Anderson-Frey et al. 2016) verification. However, there has been some limited verification of convective outlooks in the literature. In particular, Hitchens and Brooks (2012) and Hitchens and Brooks (2014) verified Day 1 convective outlooks, the latter also verifying Day 2 and 3 outlooks. The primary purpose of these studies was to evaluate skill of SPC outlooks over a very long period of record—decades—to ascertain temporal trends in performance and the effects of changes in SPC forecasting philosophies on forecast skill over time. However, both of these studies only deterministically considered the verification of the *categorical* versions of the convective outlooks via contingency table statistics, and did not quantitatively consider the *probabilistic* verification of the SPC probability contours. Additionally, the choice to verify categorical outlook contours rather than probabilistic ones made it impossible to perform verification broken out individually by severe-weather predictand for Day 1 outlooks, since the categorical outlooks are objectively determined as a combination of the individual severe weather predictand probabilities—even if probabilities may be subjectively modified to match forecaster conceptions of categorical severity levels—and, historically, were not broken out by phenomenon at all. There has been some very limited published work on verification of probabilistic SPC convective forecasts (e.g. Kay and Brooks 2000), but it is quite dated, preceding the introduction of operational convection-allowing model guidance (e.g. Kain et al. 2006), and substantial changes to both SPC outlook products

and available operational guidance in the forecast process have been introduced in the intervening years (Edwards et al. 2015). Recently, Hitchens and Brooks (2017) have begun to investigate verification of probabilistic SPC convective outlooks. However, while substantially advancing the literature in this area, the verification presented still employs deterministic contingency table-based frameworks, and thus largely neglects the specific quantitative information associated with the probability contours being evaluated.

In this paper, we seek to quantify probabilistic verification properties of SPC severe weather outlooks for Days 1–3, in particular forecast reliability and forecast skill. The following section describes the methods performed to conduct this verification, and outlines two different verification frameworks employed in the study: a so-called "Traditional" framework and an "Interpolation" one. Section 3 presents verification results using the Traditional analysis approach, while section 4 describes the Interpolation framework results and provides a comparison between the two. The paper concludes with a synthesis of the findings and a discussion of broader applications and implications of this work.

5.2 Data and Methods

Forecast data for this study comes from the shapefiles in the public SPC outlook archive (SPC cited 2017a). SPC has changed various aspects of both their product definitions and their archive over the past 10–15 years. Importantly and consequentially, the NWS changed the definition of severe hail from a minimum hail diameter of 0.75 in (1.9 cm) to 1 in (2.54 cm) beginning on 5 January 2010 (Ferree 2009), considerably reducing the number of annual severe hail reports after that date. Verification in this study is performed relative to the effective severe hail criteria at the forecast issuance time. Their categorical convective outlooks were also substantially innovated in October 2014, adding "Marginal" and "Enhanced" categories to the existing classes of "Slight", "Moderate", and "High" (Jacks 2014). However, these changes only affected how severe weather probability contours mapped to categorical risk definitions, and did not directly affect any of the probabilistic forecast contours. The public online forecast archive dates back to 23 January 2003, but forecasts are not available in shapefile format at that time. Shapefiles become available for Day 1 beginning 1 January 2009 and for Days 2 and 3 from around 11 April 2012. File format is consistent for all Day 1 shapefiles from the beginning of their archival, but a significant format change is incurred in the Day 2 and Day 3 shapefiles on approximately 13 September 2012. For these reasons, the period of record for forecast verification spans 1 January 2009–31 December 2016 (2,922 total outlooks) for Day 1 forecasts and 13 September 2012–31 December 2016 for

Day 2 and 3 forecasts (1,569 and 1,568 total outlooks, respectively). In order to maximize the period of record length, limit forecasts already affected by ongoing convection from the day of forecast issuance, and keep the issuance lead time separation—especially for Days 2 and 3—as close as possible, verification in this study is based on the 1300 UTC probabilistic convective outlooks for Day 1, the 0100 CT probabilistic convective outlooks for Day 2, and the 0230 CT outlooks for Day 3.

Archived forecast shapefiles store a list of points defining each polygon issued in the given probabilistic forecast. The verification for this study seeks to compute CONUS-wide verification statistics making use of the quantitative forecast probabilities in a consistent, repeatable manner for each forecast day throughout the period of record. These objectives are by far the most easily equitably achieved using probability forecasts on a uniform grid for each forecast day. This thus requires the conversion of the contour definitions provided in the SPC shapefiles to probability grids.

Verification in this study is performed using two distinct, but complementary approaches. The first approach closely follows the verification performed internally at SPC (R. Edwards 2017, personal communication). Probabilities are gridded onto a CONUS-wide grid with 80-km grid spacing. Probabilities on this grid simply correspond to the value of the innermost probability contour enclosing the grid point, or are zero if no such contour exists. No interpolation is performed between probability contours in this verification scheme, and the forecast grids instead reflect a discrete number of possible forecast probabilities as determined by the allowed contour levels for the given predictand. This verification is performed to correspond with the current state of the science and for a direct comparison with internal statistics historically computed at SPC.

Recognizing however that appropriate verification is determined by the predictand definition(s) in conjunction with the forecast objectives and not on historical practice alone, a second, parallel verification approach is performed with the aim of advancing the state of the science in probabilistic severe-weather forecast verification and obtaining consistency with public interpretation and use of SPC outlooks. Ultimately, verification of neighborhood-based predictands such as those used in SPC's probabilistic convective outlooks occurs in continuous—rather than gridded or discrete—space. A severe weather observation occurs at an arbitrary physical point in space, and of course is not constrained to occur on any grid of finite size; the circle defining 40 km centered about that point is similarly unconstrained to any particular grid. This argues against the use of grids at all, since their use only distorts the "true" relationship between the forecast and the observations; the distortion extent

is proportional to the grid spacing, with no noise added over the "true" relationship in the case of infinitesimal grid spacing. Grids are used to provide a common and convenient quantitative analysis framework for comparing forecasts and observations, but in order to best represent the true relationship between these fields, it is desirable to have as small of a grid spacing as is computationally feasible, and it is certainly desirable to have a grid spacing appreciably smaller than the neighborhood radius of the predictand. Additionally, although SPC forecasters may only issue forecast probabilities at a given point corresponding to discrete levels defined by the allowable contours, some end-users may reach the interpretation that within a region bounded by two contours, the verification probability is higher at a point directly adjacent to the higher probability contour when compared with a point adjacent to the lower probability one. Regardless of forecaster interpretation and intent, this is consistent with how the public and broader meteorological community may interpret plots of other continuous fields with a discrete number of contours (e.g. Lackmann 2011; NWS 2017b), and it is important to evaluate forecasts in a manner consistent with how forecasts are perceived by an educated end-user. To this end, in addition to performing verification on an 80-km grid without probability interpolation in the so-called "Traditional" approach, verification is also performed on a finer resolution grid with interpolated probabilities in the "Interpolation" approach described below.

For the Interpolation approach, ArcGIS was used to process the SPC shapefiles into probability grids using a dynamic workflow divided into three different methods based upon the characteristics of the probabilities issued on a specific day. The first of these methods (hereafter referred to as INTERP) interpolates between the SPC probability contours when more than one contour is present in the daily convective outlook and outputs the results on the specified probability grid. The second method (hereafter referred to as CONSTANT) does no interpolation and is used when one contour level is present in the daily convective outlook since the lack of a defined probability gradient leaves the interpolation problem unconstrained. The third method (hereafter referred to as NODATA) is used when no contours are present in the daily convective outlook and outputs a constant grid of zero values over a CONUS-wide analysis domain that extends so far as the center of the neighborhood is over CONUS land. Open contours that end due to intersection with a CONUS boundary are closed using the CONUS edge as the remaining contour boundary such that all of that area is enclosed. Due to differences between the Day 1, 2, and 3 outlook shapefiles, the workflow was carried about at horizontal resolutions

of $0.03227^\circ \times 0.03227^\circ$, $0.05^\circ \times 0.05^\circ$, and $0.1^\circ \times 0.1^\circ$, respectively. However, for consistency, the interpolated Day 2 and Day 3 outlooks were regridded bilinearly onto the $0.03227^\circ \times 0.03227^\circ$ used for the Day 1 forecasts. All subsequent verification for the Interpolation approach is performed on this grid.

Being a grid with a fixed degree increment in latitude and longitude, the physical area spanned varies with latitude. While the difference in physical distance of a fixed degree increment in the latitudinal dimension varies negligibly as a function of latitude, the physical distance in the *longitudinal* dimension does vary appreciably over the domain. One degree of longitude is approximately 73 km at the northern border of CONUS, while the same increment corresponds to near 100 km at the southern extremities. Consequently, area in northern CONUS is weighted slightly more—a factor of around 1.35 more in the extremes—than in southern CONUS in the calculation of bulk verification statistics. However, with CONUS broadly being confined to the mid-latitudes, this effect has only a small quantitative effect and is not believed to appreciably impact any conclusions drawn from the analysis. Each method within the dynamic workflow is described in detail below.

The INTERP method first converts the native SPC shapefile polygons (e.g. Fig. 5.1a,d) to contours (e.g. Fig. 5.1b,e) that maintain the probability values of the original forecast. These contours are then used as input to the ArcGIS Topo-to-Raster function (Childs 2004) and the output is then extracted only over the spatial extent that was in union with the original SPC probabilities (visually depicted in the fill in Fig. 5.1c,f). Trial and error revealed that the output from the Topo-to-Raster function of the native probabilities tended to be lower than the initial contoured input: along an explicit intermediate contour, interpolated values tended to be approximately half of a probability bin lower than the value of that contour. This is mainly attributed to the interpolation problem becoming more unconstrained as fewer contours are present to be analyzed by the function. In order to correct for this, a second step was added to the INTERP process where another raster was created using the Topo-to-Raster function, but this time, using the SPC probabilities for the same day that were incremented up approximately one probability bin. With a half-bin negative bias in the interpolation procedure, the resulting values from the interpolation on the native probabilities were approximately half of a probability bin too low, and when adjusted upward by one probability bin, the resultant field was approximately half of a probability bin too high. The arithmetic mean of the two separate interpolated rasters (i.e. one from the original forecast probabilities and one from the probabilities incremented up one interval) therefore serves as an unbiased interpolated representation of the probability field. This mean thus serves as the output of the INTERP method (e.g. Fig. 5.1c,d). The only time this did not hold was when there was

an increase in the contour interval with increasing probability; then the bias tended to be smaller at approximately one third of a probability interval, requiring less upward adjustment to produce an unbiased derived interpolated field. For example, in the case of a Day 1 convective outlook that contained tornado probabilities of 2%, 5%, 10%, 15%, and 30% (e.g. Fig. 5.1b), the corresponding probabilities that where incremented up in the second Topo-to-Raster run are 3%, 10%, 15%, 20%, and 45%. The output of the INTERP method creates a raster that maintains a representative depiction of the SPC contours (cf. Fig 5.1b,e to Fig. 5.1c,f), but interpolates in a manner that produces a smooth gradient between the contours and increases the maximum probabilities within the highest contour. There are still instances where the highest probability contour is slightly distorted compared to the original (cf. 30% in Fig. 5.1b,c and 15% in Fig. 5.1e,f); however, the differences in these regions are rarely larger than 1%.

The CONSTANT method, similar to the INTERP method, converts the native SPC shapefile polygons (e.g. Fig. 5.1g) to a constant raster (e.g. Fig. 5.1i) using the ArcGIS Feature-to-Raster tool. Since there is only one contour in the cases that the CONSTANT method is used (see above), no attempt is made to interpolate. As in the INTERP method, the CONSTANT method then outputs the probability raster onto the analysis domain at the resolution corresponding to the convective outlook lead time. The NODATA method simply outputs a constant grid of zero values over the analysis domain, again, at the corresponding resolution for the analyzed convective outlook. The probabilities for each threat (i.e., Tornado, Wind, and Hail) and outlook lead time (i.e. Day 1, 2, and 3) are run through this dynamic workflow to create the analyzed grids that are used for the verification undertaken in this chapter.

The Traditional framework, in contrast, compares the effect the INTERP part of the dynamic workflow plays on the probabilistic verification when compared with the Interpolation method. In the Traditional approach, all of the threat and lead time combinations are run through a simplified workflow that contains only the CONSTANT and NODATA methods and outputs to a lower resolution grid with 80 km grid spacing.

Once all outlooks have been gridded, the same verification methods are employed in each verification framework to assess the quality of these outlooks. These include several commonly used probabilistic verification tools. Specifically, forecast reliability, the extent to which forecasts verify at the

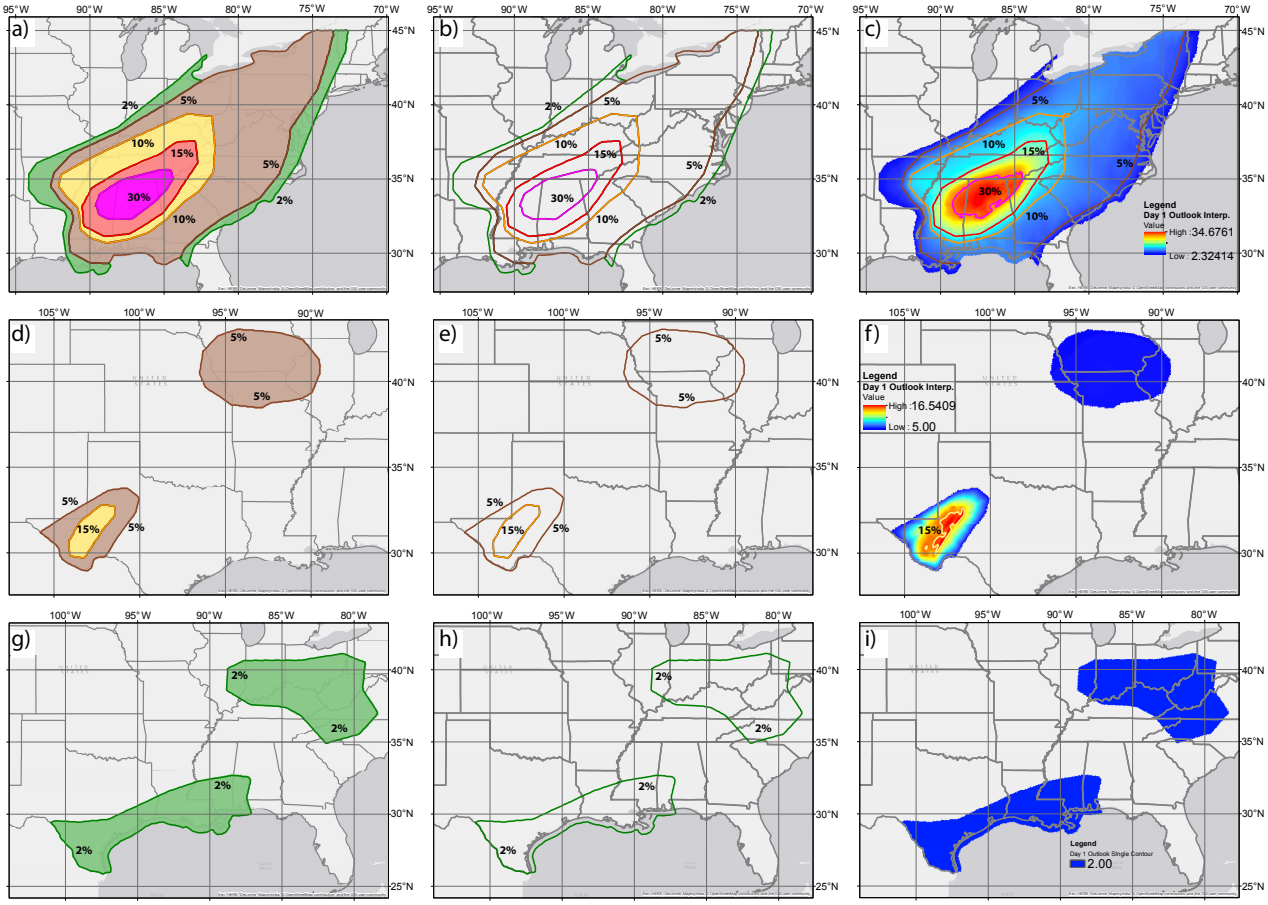


FIG. 5.1. Step-by-step examples of the regridding process performed on the SPC probabilities. (a–c) show the INTERP method example for the Day 1 tornado probabilities valid on 1300 UTC 27 April 2011, (d–f) show the INTERP method example for the Day 1 hail probabilities valid 1300 UTC 21 October 2012, and (g–i) shows the CONSTANT method example for the Day 1 tornado probabilities valid 1300 UTC 1 May 2016. (a,d,g) show the ArcGIS depiction of the native SPC forecast probability polygons with the colors matching the standard SPC graphic. (b,e,h) depict the contours derived from the native SPC forecast probability polygons that are used for input into the Topo-to-Raster function, where the line colors also correspond to the colors used in the standard SPC graphic. (c,f,i) depict the final gridded output from the INTERP (c,f) and CONSTANT (i) methods of the dynamic workflow, where the colored contours represent the location of the constant probability contours in the gridded output to compare to the original input in (b,e,h).

same frequency as indicated by the forecast probability, is assessed by inspection of reliability diagrams (Murphy and Winkler 1977; Bröcker and Smith 2007; Wilks 2011). Forecast skill is quantified using Brier Skill Scores (BSS; Brier 1950), defined as:

$$BSS = 1.0 - \frac{BS}{BS_{clim}} = 1.0 - \frac{\sum_c (p_c - o_c)^2}{\sum_c (p_{clim_c} - o_c)^2}, \tag{5.1}$$

where p_c denotes the forecast probability of a case, p_{clim_c} denotes the climatological forecast probability for the case, and o_c is a binary variable indicating whether the forecast predictand was observed for the given case. Forecasts—defined by forecast day, latitude, longitude triplets on the prescribed analysis grid—are aggregated several different ways to ascertain various properties of SPC outlooks. In particular, they are aggregated spatially in order to determine the regional distribution of forecast skill for the different severe weather predictands, temporally both by month and year to ascertain whether there is any persistent seasonality to forecast skill and whether forecasts are generally improving or degrading year-to-year over the period of record, and meteorologically based on the conditions of the forecast point to deduce whether there is any relationship between the forecast environment and SPC severe-weather forecast skill, which may also speak to the larger predictability of severe weather under different meteorological conditions.

Despite their limitations to be discussed in more detail below (e.g. Trapp et al. 2006), verification uses storm reports as archived in SPC's Severe Weather Database (SPC cited 2017b), and these reports are taken to be "truth" whereby reports are taken to be certain events and non-reports are taken to be certain non-events. All verification is performed on the CONUS-wide 0.03227° × 0.03227° grid and 80 km×80 km grid in the Interpolation and Traditional frameworks, respectively. In particular, for each severe weather predictand, all points on this grid within 40 km of a severe report from the database are encoded as an observed event for the 24-hour forecast day corresponding to the report; all other points in this spacetime matrix are encoded as non-events. Calculating skill scores as described above uses a climatological reference forecast; these are calculated identically to the official SPC severe climatologies (Kay and Brooks 2000; SPC cited 2017c), except on a finer grid in the case of the reference in the Interpolation framework. Raw verification grids are calculated as described above for 1982–2011 to match the period of the severe-weather climatologies published on SPC's website at the time of this study. The 30 annual verification grids are then collectively employed to derive raw frequency grids for each day of year, latitude, longitude triplet for each of the severe weather predictands. These raw frequencies are then smoothed over time using a Gaussian filter with a 15-day standard deviation and using a "wrap" filter mode to handle treatment with respect to year-beginning and year-end. Finally, these are followed by smoothing over the two spatial dimensions with a 120-km standard deviation Gaussian filter with a "reflect" mode treatment for domain edges; the resulting grids are said to be the climatological event probabilities (Kay and Brooks 2000).

In addition to space and time, severe-weather forecasts are often viewed with respect to the prevailing environmental conditions for the forecast, and especially the CAPE-versus-shear parameter space (e.g. Schneider and Dean 2008). Verification with respect to this meteorological regime-based parameter space can provide useful forecast insights into environments of greater and lesser forecast skill. This does however require the use of an external data source to quantify the forecast environment for each forecast. While RAP/RUC analyses are frequently employed in this sort of context (e.g. Bothwell et al. 2002; Dean et al. 2009), the transition from the RAP to RUC in May 2012 (NWS 2012) provides an undesirable potential source of inconsistency in the middle of the study period, and both received smaller changes to their data assimilation systems throughout which may also result in changes to the analysis creation. In order to use a data source that is created in a consistent manner throughout the analysis period, this study uses the North American Regional Reanalysis (NARR; Mesinger et al. 2006) to determine the local meteorology at a point for a given forecast period. The NARR also has been used in analysis of severe weather environments in past studies (e.g. Gensini and Ashley 2011; Nielsen et al. 2015; Vaughan et al. 2017), and while there are some quantitative differences, errors, and regional biases—particularly in the thermodynamics (e.g. Gensini et al. 2014)—compared with other analysis and reanalysis products, the use in this study is largely to classify the general regime of the environment and not to exactly quantify the CAPE or shear at a particular point. To this extent, the NARR has been found to be qualitatively consistent with other analysis products (e.g. Vaughan et al. 2017). The NARR has 3-hour temporal and 0.25° spatial resolution and includes an assortment of fields at various vertical levels from subsurface to 100 hPa. Mean Layer Convective Available Potential Energy (MLCAPE) used for this study comes directly from the NARR and performs averaging over the layer encompassing the lowest 180 hPa of the atmosphere. Deep-layer shear (DSHEAR), another important severe-weather parameter (e.g., Doswell et al. 1993; Gallus et al. 2008; Markowski and Richardson 2010), is often expressed as the bulk wind difference between surface (10 m) wind and 6 km above ground level. This is not available in the NARR; instead, surface to 450-hPa wind difference is used for this study, and this value is rescaled based on the geopotential height at 450 hPa to approximate the value of the 0–6 km shear. All days are classified based on the maximum 3-hourly value over the 24-hour 1200–1200 UTC period for each parameter.

In order to better ascertain the robustness of the various findings, uncertainty analysis is performed for each phase of verification. A bootstrapping procedure is employed to generate confidence intervals for BSS analysis. For BSS over space and time, forecasts are resampled randomly with replacement from

each analyzed grid point and time period studied—both year and month—to ascertain the uncertainty in the skill score for the given subspace. A similar method is employed for parameter space, subsampling instead for each 250 J kg^{-1} × 2.5 m s^{-1} subregion of CAPE vs. deep-layer shear parameter space. For all of this analysis, due to the small spatial scales of storms most commonly associated with severe weather and the spatiotemporally scattered nature of observed reports, points with non-overlapping 40-km radius neighborhoods and all forecasts on separate days are considered to be independent from one another, while forecasts on the same day with overlapping neighborhoods are, necessarily, considered to be non-independent. Reliability uncertainty is also assessed using the methods of Agresti and Coull (1998) as described also in Wilks (2011).

5.3 Results: Traditional Framework

In the Traditional verification framework, highest BSS values for all spatial fields are generally seen in the eastern and especially central United States, with lower skill observed in the West (Fig. 5.2). More generally, spatially, the highest skill is often observed where the climatological event frequency is higher, as evidenced by higher climatological BSs in Figure 2. This holds comparing across the Day 1 outlooks for the individual severe phenomena as well. Severe winds (Fig. 5.2e), for example, have the highest BSS over all of CONUS at 0.093, followed by severe hail at 0.076 and tornadoes—the rarest phenomenon—coming in last of the three with a score of 0.049. The same does not hold true, however, for the "significant" severe phenomena. While the skill for all three phenomena (Fig. 5.2b,d,f) is lower at the significant criteria compared with the outlooks for the same phenomena that include events of lesser severity (Fig. 5.2a,c,e), significant-tornado outlooks (Fig. 2b) are the most skillful in aggregate of the three significant-severe outlooks with an aggregate BSS of 0.028. Significant hail (Fig. 5.2d) lags substantially with an aggregate score of just 0.008, and significant wind events (Fig. 5.2f) have approximately no skill at all over climatology with an aggregate score of -0.001. Positive skill in significant wind events is largely confined to the Ohio River Valley and middle Mississippi River Valley areas, and this skill is not statistically significant. Due to a small sample size, the negative skill in much of the West, and particularly the Arid Southwest, is not found to be statistically significant either. This holds even for significant-hail (Fig. 5.2d) events, where very negative climatology-relative skill is observed in those regions. Given the relative rarity of significant-severe events, just a few events with higher (lower) predictability and large spatial coverage can drive a large degree of positive (negative) skill in a given region. The one parameter with areas of statistically significant negative skill occurs also with significant-wind

events, where weak but statistically significant climatology-relative skill is exhibited along a strip of the Atlantic Coast from the Florida Panhandle through central New York, and secondarily in a region from the Nebraska Panhandle south through the Texas Panhandle. In the West, where larger negative BSS values are obtained, the sample size is insufficient to produce statistical significance. In no other regions is the sign of the obtained skill score, positive or negative, found to be statistically significant for significant-severe forecasts. Statistically significant positive skill is seen for the regular severe convective outlooks, however. While in tornado outlooks (Fig. 5.2a) the statistical significance is confined to the regions such as the Tennessee River Valley region and parts of the Central Plains where highest climatological event frequencies overlap highest skill scores, for severe hail (Fig. 5.2c), most of central and southeastern CONUS excluding the immediate Gulf Coast area exhibits significantly positive skill, in addition to parts of New England. These are also, unsurprisingly, where the skill scores and event frequencies are highest for this phenomenon. With three exceptions, statistically significant positive skill is observed over all of CONUS east of the continental divide for severe-wind forecasts (Fig. 5.2e). The first two exceptions, in south Texas and the Florida Peninsula, skill scores are lower than in neighboring regions, whereas in the far northern Great Plains states, the event frequency is somewhat lower as evidenced by smaller climatological BSs in that region. In these three areas, conditions are insufficient to garner statistical significance in the BSS. Despite different verification frameworks, these findings agree with those of Hitchens and Brooks (2017). For significant-severe events, they similarly found the lowest skill among wind outlooks and the highest skill for tornadoes. Also like in this study, they found substantial skill improvement when considering outlooks pertaining to all weather exceeding the minimum severe criteria rather than the outlooks for only the significant-severe events.

The same general tendencies are observed for the outlooks of all severe weather for Days 2 and 3 (Fig. 5.2g,h) with highest skill over the central US and lower skill over the West. Of note, the Southeast region including the Carolinas, Georgia, and Florida has degraded skill compared to the Day 1 outlooks and is largely slightly negative. Due to considerable variability in success of the longer lead time forecasts, none of the skill—positive or negative—was found locally to be of statistically significant sign. With domain-total skill scores of 0.055 and 0.028 for Days 2 and 3, respectively, there is a clear deterioration in forecast skill with increasing forecast lead time from Day 1 through Day 3.

Comparing the verification results across the period of record (Fig. 5.3), no clear trend in skill scores is seen from the beginning of the period to the end. The verification period of this study is admittedly much shorter than that of Hitchens and Brooks (2012, 2014) when a marked improvement in forecasts

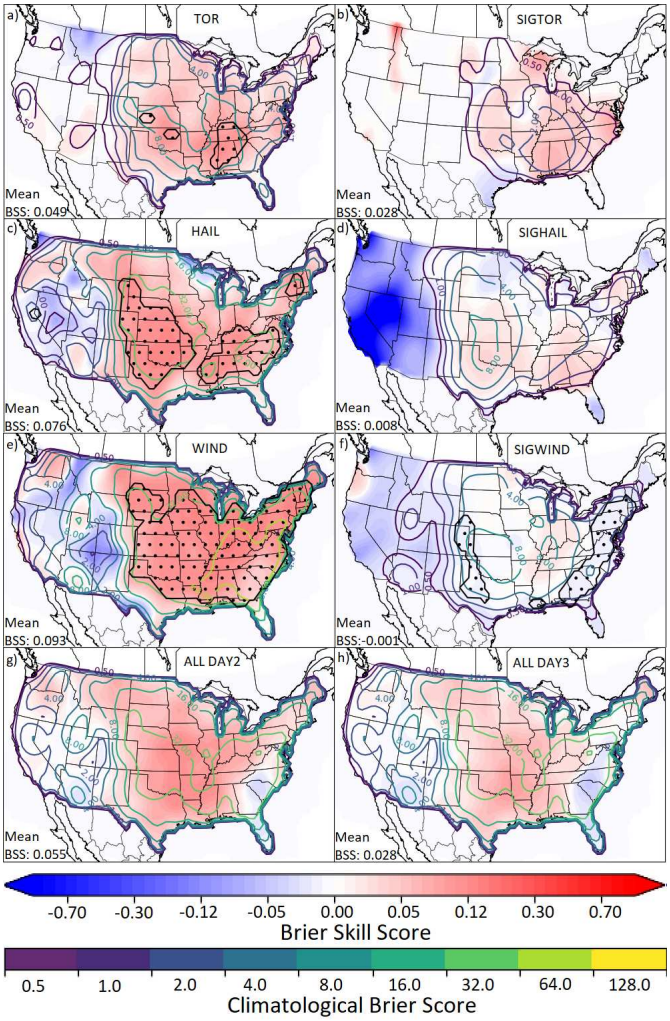


FIG. 5.2. BSS spatial distributions for each of the forecast sets in this study using the Traditional verification framework. Panel (a) plots results of Day 1 tornado probability forecasts, (b) to significant tornado probabilities, (c) and (d) respectively for Day 1 hail and significant hail, (e) and (f) for Day 1 wind and significant severe wind, and (g) and (h) for any severe probabilities for Days 2 and 3, respectively. The "mean" BSS (unity minus ratio of sum of Brier scores at all grid points for the given variable divided by sum of climatological Brier scores at all grid points for the same variable) is depicted in the bottom left of each variable's panel. A 120-km standard deviation Gaussian smoother was applied prior to plotting for panels (a), (c), (e), (g), and (h), while a larger 180-km smoother was applied for the significant severe variables in panels (b), (d), and (f) owing to the smaller sample sizes. Unfilled color contours depict the climatological Brier scores for the verification location; larger numbers indicate locations of more frequent events and more impactful areas towards the mean score. Note that the contour interval and color scale, shown at figure bottom, is nonlinear. Stippling depicts areas where the sign of the indicated skill score is statistically significant with 95% confidence using a bootstrapping procedure as described in the chapter text. Light smoothing of the significance contours has been performed to enhance readability.

was observed over the decades of SPC outlooks analyzed. In general, like with the spatial results where areas of higher event occurrence exhibited more skill, skill tends to be somewhat higher on more-active

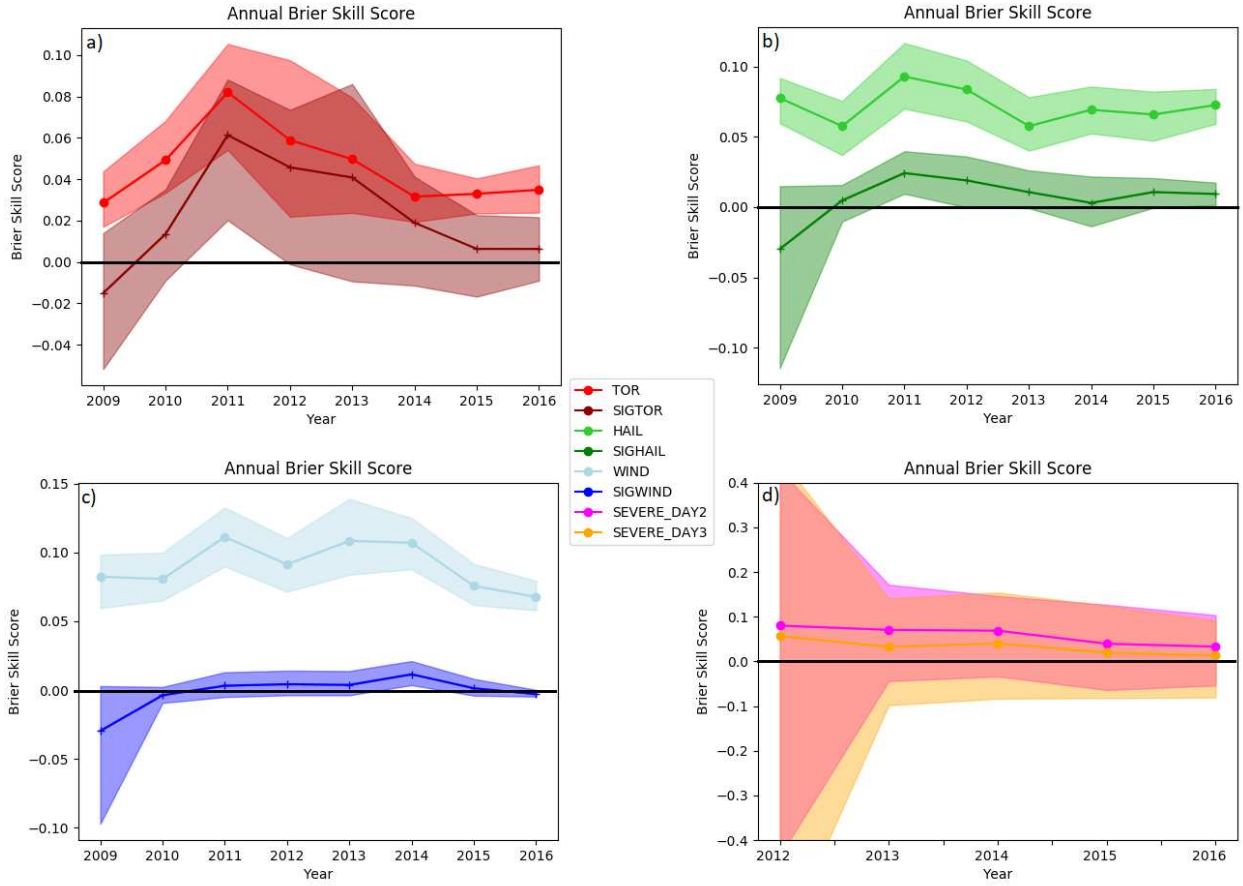


FIG. 5.3. Brier Skill Scores for each forecast set as a function of year of forecast issuance for (a) tornadoes and significant tornadoes, (b) severe hail and significant-severe hail, (c) severe wind and significant-severe wind, and (d) Day 2 and 3 forecasts of any severe weather using the Traditional verification framework. Brier scores and climatological BSs have been summed over space to produce the skill scores shown in (a)–(d). Transparent shading around lines indicate 95% confidence intervals on the BSS obtained via bootstrapping as described in the text. Note that the y-axes vary between panels.

years compared with less-active ones. This is especially true for tornadoes (Fig. 5.3a), where the highest skill—both for all tornadoes and for just significant ones—is seen in the historically active 2011 season; this holds to a lesser extent with hail and wind as well. As a result, and given a particularly low skill 2009, an increase in skill is seen in the 2009–2011 period, with a gradual decline in skill thereafter likely attributable to annual fluctuations in severe weather frequency, especially those associated with relatively high-predictability synoptic-scale regimes. Statistically significant positive skill is seen for all of tornadoes, hail, and wind for all years. For each phenomenon, skill is consistently better year-to-year for the regular severe criteria compared with forecasts of significant-severe events. Compared with their less stringent counterparts, confidence intervals are larger for significant tornadoes and smaller for

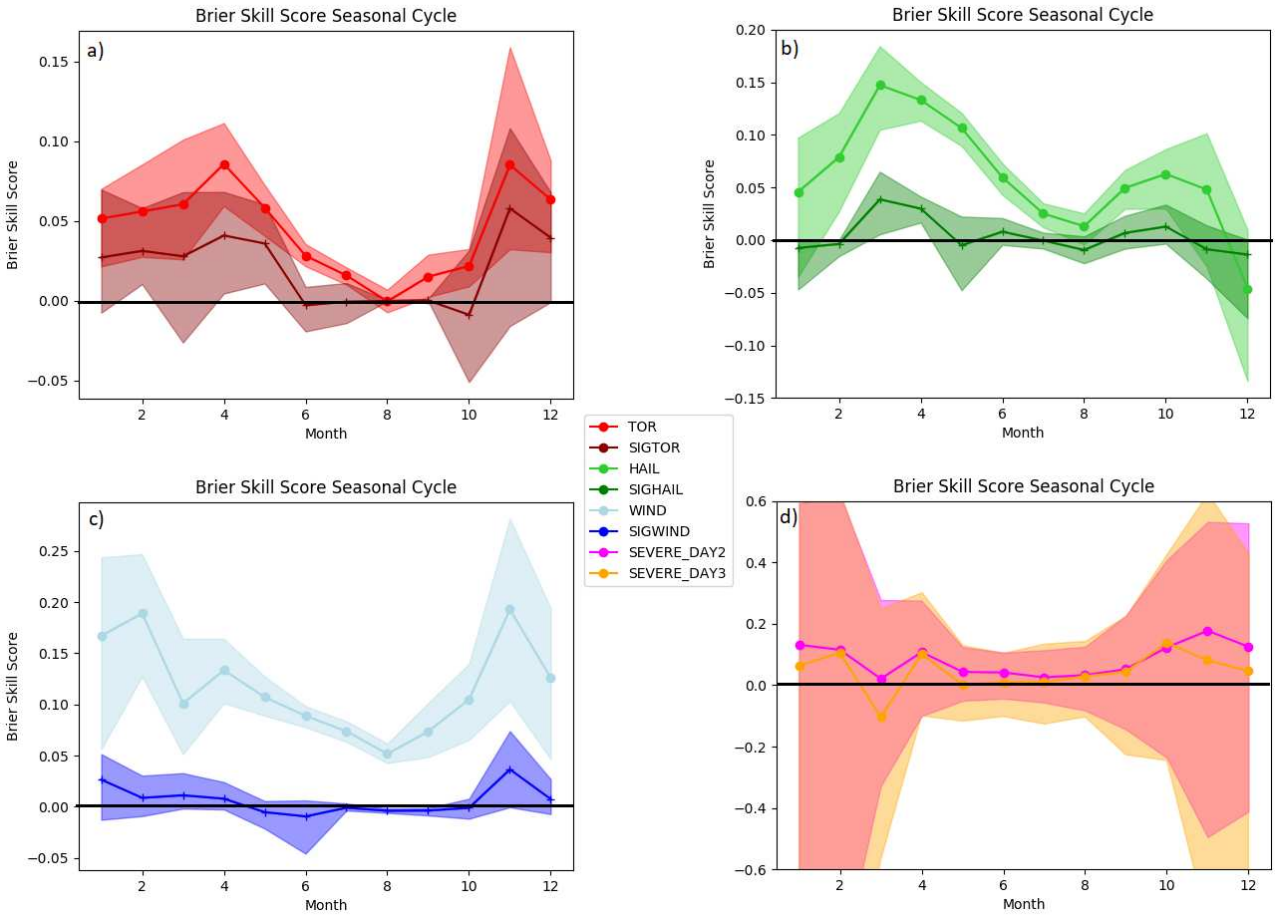


FIG. 5.4. Same as Figure 5.3, except for by month of forecast issuance.

significant-wind events, but with a couple of exceptions, no statistically significant skill of either sign is observed for significant-severe events throughout the period of record. For Day 2 and Day 3 outlooks (Fig. 5.3d), Day 2 forecasts consistently verify slightly better than Day 3 forecasts year-to-year despite slight fluctuations in skill overall. Confidence intervals are large compared to the Day 1 forecasts for specific phenomena, largely due to large forecast-to-forecast variability in the successfulness of individual outlooks, in addition to an overall shorter period of record. The intervals are particularly large in 2012 because only approximately one third of the year falls into the period of record, resulting in a significantly reduced sample size.

Fairly substantial amplitude seasonal cycles of forecast skill in Day 1 outlooks (Fig. 5.4a,b,c) were discovered from this verification. Tornadoes, both for outlooks of any tornado and for only significant ones, exhibit two peaks, one in the spring and a particularly sharp one in the late autumn, maximizing in November. Between the two, there is a broad skill minimum throughout the summer and early

autumn, consistent with prior studies (e.g. Hart and Cohen 2016). In fact, tornado outlooks suffer a skill degradation sufficiently large such that in August, tornado outlooks of both severity levels verify about equally. Confidence bounds are also much tighter in this minimum, and the seasonal differences for the EF0+ outlooks are found to be statistically significant. A somewhat similar trend is seen for hail forecasts (Fig. 5.4b), however the springtime maximum peaks slightly earlier, in March rather than April, and is significantly larger than either of the skill spikes in the tornado outlooks. In addition, the skill maximum in the autumn still exists, but is of a much smaller magnitude and is no higher than the winter months. While a summer local minimum of skill is observed, the primary minimum is seen in December, where appreciably negative aggregate skill is in fact observed during outlooks from that month. The monthly differences between these minima and maxima in forecast skill are found to be statistically significant. Significant hail events, in contrast, are found to have a substantially muted seasonal cycle, with a slight increase in skill during the spring found to be the primary feature. Wind events (Fig. 5.4c) follow a somewhat similar pattern to tornadoes, with a clear minimum in skill reaching its lowest point in August, and a maximum in skill in November, but there is no real secondary springtime peak and instead a gradual degradation in skill throughout the winter and spring months. This cold-season skill maximum coincides with a period whereby a higher proportion of severe weather events are from synoptically-forced systems than in the warm-season, and thus likely have higher predictability, particularly at the extended range, than those with weaker or smaller-scale processes primarily responsible (e.g. Surcel et al. 2016; Nielsen and Schumacher 2016; Herman and Schumacher 2016a). Significant-wind events, like significant hail, feature a muted seasonal cycle with a slight maximum observed during the late autumn and early winter. Significant-severe wind events also verify significantly worse than other severe-wind events throughout the entire year. Lastly, Day 2 and 3 convective outlooks (Fig. 5.4d) show relatively little seasonal cycle in forecast skill, except with a slight enhancement of skill in November as seen in many of the Day 1 outlooks. Confidence intervals are generally very large, but shrink considerably in size in the warm-season where sample size is much larger. With one slight exception in October, Day 2 forecasts continue to perform slightly better than Day 3 forecasts from month-to-month; the magnitude of this difference tends to be smallest in the late summer to early autumn and largest in late winter to early spring.

Skill verification in the CAPE-versus-shear parameter space (Fig. 5.5) depicts positive skill throughout much of the parameter space for each of the different severe phenomena forecasted in Day 1 convective outlooks, as one would expect given the positive aggregate skill (Fig. 5.3a,c,e), with two primary

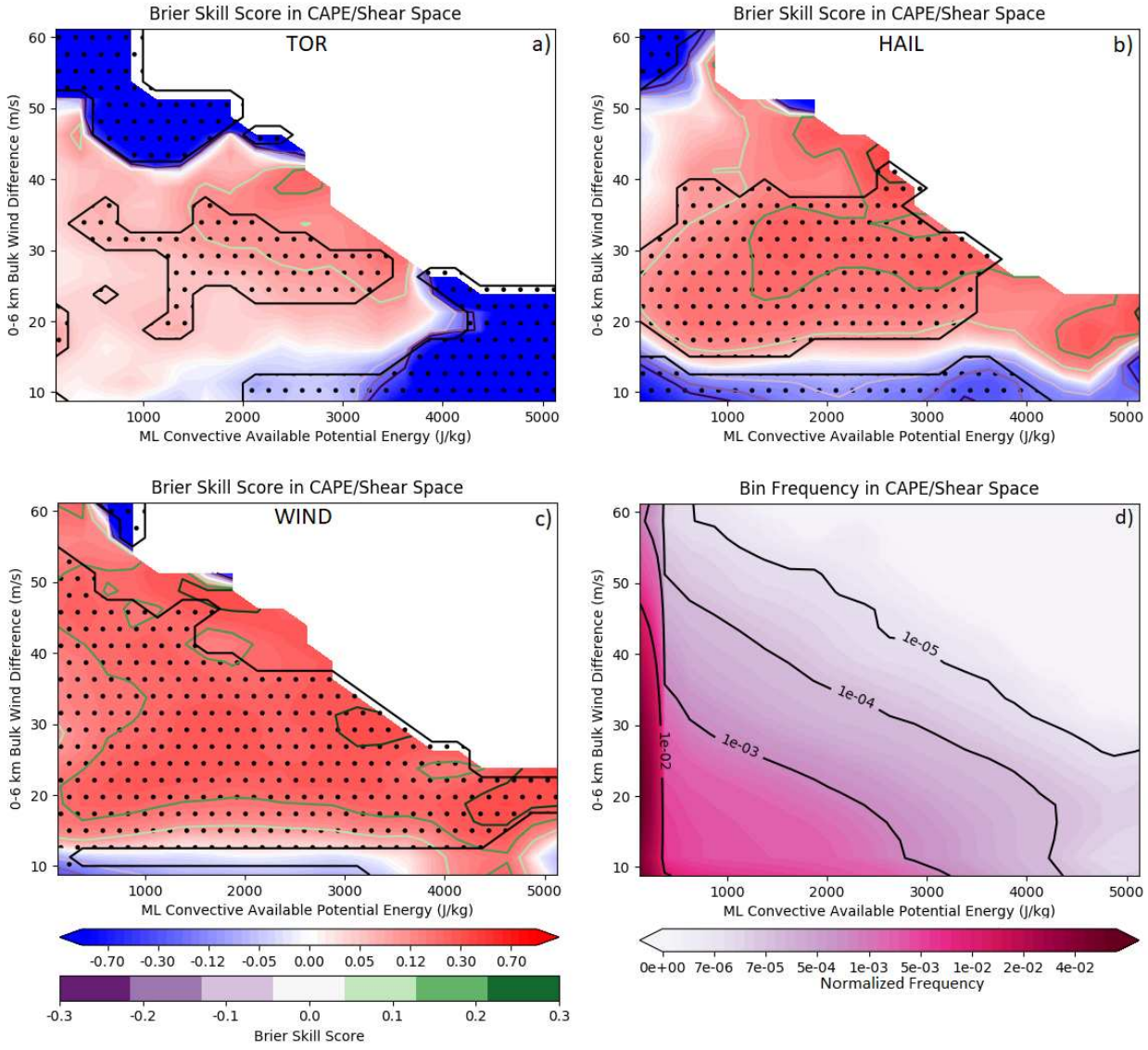


FIG. 5.5. BSS as a function of the prevailing MLCAPE and DSHEAR at the forecast point for (a) Day 1 Tornado, (b) Day 1 Hail, and (c) Day 1 Wind forecasts verified from 1 January 2009–21 August 2014 using the Traditional verification framework. Panel (d) indicates the raw frequencies of points falling into each bin, separated by 250 J kg^{-1} in MLCAPE space and 2.5 m s^{-1} in DSHEAR space, over the verification period. Values have been lightly smoothed with a 187.5 J kg^{-1}, 1.875 m s^{-1} Gaussian smoother for increased clarity. Stippling denotes regions of the parameter space where the sign of the indicated skill score is known with 95% confidence. Note that both the red/blue and magenta scales are nonlinear, particularly the latter one. Both the red/blue and green/purple scales depict the same BSS field, but the explicit contours in green/purple are included for quantitative clarity.

regions of exception. First, low climatology-relative skill is seen in much of the parameter space with very weak deep-layer shear. For hail (Fig. 5.5b), this is true throughout the weak shear region of the parameter space. In contrast, for tornadoes (Fig. 5.5a), this is especially emphasized in the low-shear,

high-CAPE region of the parameter space while for wind (Fig. 5.5c), the opposite is true, with the most negative scores found in the low-shear, low-CAPE environments. The second region of deflated skill is in the low-CAPE, high-shear region of the parameter space, an area frequently noted as a particularly challenging forecast region in the phase space (e.g. Evans and Doswell 2001; Davis and Parker 2014; Sherburn and Parker 2014; Sherburn et al. 2016); this degradation in skill is evident for all variables, but particularly pronounced for tornadoes (Fig. 5.5a) and least apparent for wind (Fig. 5.5c). Despite both of these environments being quite rare (Fig. 5.5d), the signs of the skill score in these subregions of parameter space are found to be statistically significant. Much of the positive skill regions are also found to be statistically significant. For hail and wind (Fig. 5.5b,c), the main exception is portions of the high-shear, high-CAPE regions of parameter space, where the sample size is too small (Fig. 5.5d) to obtain significance. For tornadoes, some of the more common lower-shear, lower-CAPE environments also fail to acquire statistical significance, in these subregions owing to deflated skill scores (Fig. 5.2a) compared with hail and wind. These results on outlook verification largely agree with findings of Anderson-Frey et al. (2016) and others on tornado warning verification, which similarly found best performance when both ingredients were highest, and the worst outcomes when one ingredient or the other was lacking while the other remained relatively large. These findings are with respect to a climatological baseline, and other qualitative findings may emerge with comparison with respect to a different reference forecast.

Lastly, with regards to the attributes diagrams characterizing these forecast sets (Fig. 5.6), one can note that the vast majority of forecasts of all variables have probability zero, with forecasts becoming increasingly rare with increasing probability. This is especially true of tornadoes, which have at at least an order of magnitude fewer forecasts than other variables at probability thresholds at and above 5%. At the extreme, 60% probabilities have been issued for wind during each variable's period of record, and have only been issued for approximately 1 in 100,000 forecast points. Tornado forecasts in the Traditional framework appear to be quite negatively biased—observed relative frequencies are substantially higher than their corresponding forecast probabilities for probabilities at and above 5%. This is also true for Day 2 and Day 3 convective outlooks. At forecast probabilities of 5%, the observed relative frequency is over 10% for each variable (Fig. 5.6b). This improves slightly at higher probabilities, but the extension of negative bias extends there as well. At the highest observed forecast probabilities during the Day 2/Day 3 period of record (Fig. 5.6a), 45%, Day 3 forecasts verify with approximately that frequency, but Day 2 forecasts remain statistically significantly negatively biased. Wind and hail forecasts

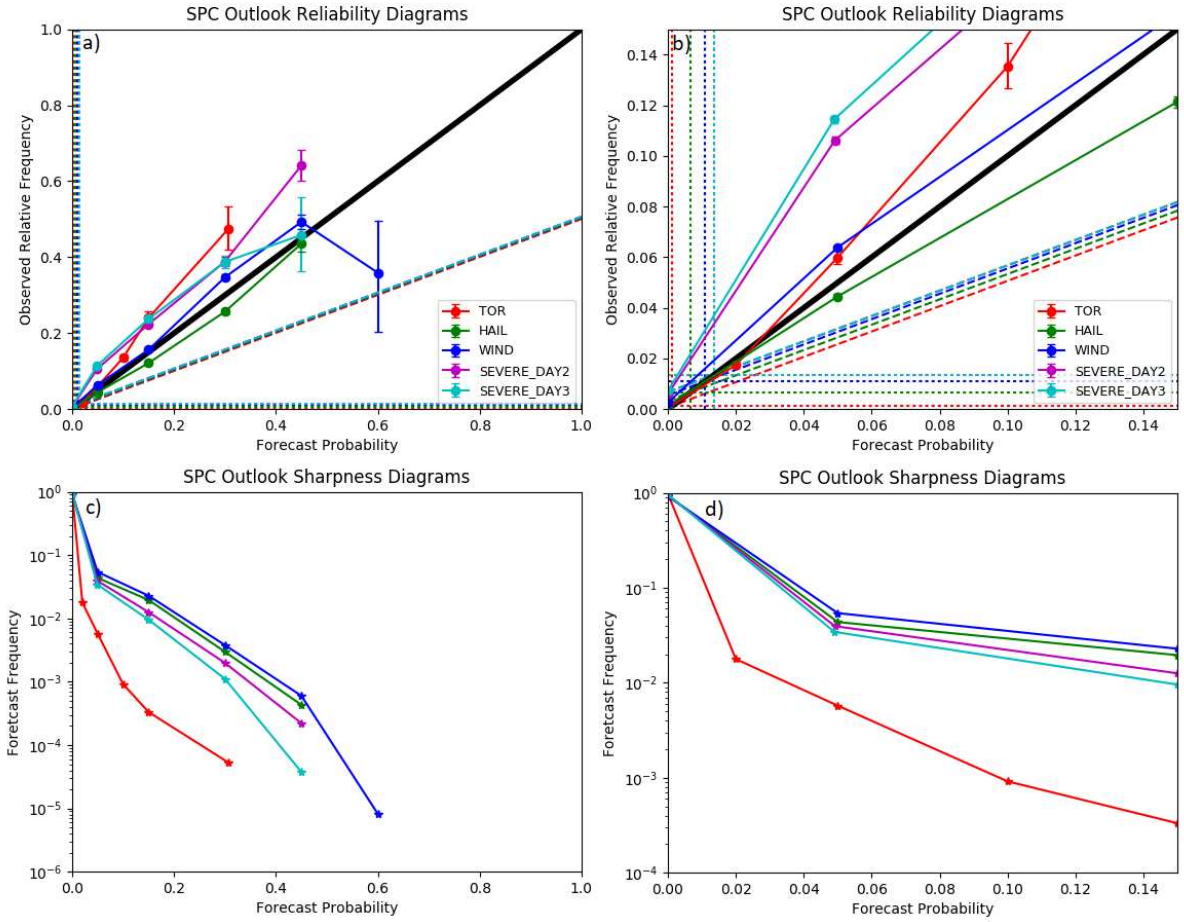


FIG. 5.6. Reliability and sharpness diagrams using the Traditional verification framework. Panels (a), (b): colored lines with circular points indicate observed relative frequency as a function of forecast probability; the solid black line is the one-to-one line, indicating perfect reliability. Colors correspond to forecast sets of different parameters and lead times as indicated in the panel legend. Panel (a) portrays the entire reliability diagram, while panel (b) is a zoom of panel (a), restricted to only probabilities of 0.15 or lower. Probability bins correspond to the full range of discrete probabilities that SPC can issue for the given forecast variable. Horizontal and vertical dotted lines denote the "no resolution" lines and correspond to the bulk climatological frequency of the given predictand. The tilted dashed lines depict the "no skill" line following the decomposition of the Brier score. Error bars correspond to 95% reliability confidence intervals using the method of Agresti and Coull (1998), where non-overlapping neighborhoods are assumed to be independent. Panels (c), (d): sharpness curves, whereby lines indicate the total proportion of forecasts falling in each forecast probability bin, using the logarithmic scale shown on the y-axis and using the same color encoding used in panels (a) and (b). X-axes of (c) and (d) correspond with those of (a) and (b), respectively.

at Day 1 are reasonably well calibrated, except for wind forecasts at the 60% probability thresholds, which are actually statistically significantly positively biased despite the small sample size and large associated uncertainty.

5.4 Results: Interpolation Framework

The overall findings of the spatial distributions of the BSSs in the Interpolation verification framework, shown in Figure 5.7, are similar to those in the Traditional framework (Fig. 5.2), but there are some important and interesting differences in the details. Overall, skill scores in aggregate are slightly higher in the Interpolation framework compared to the Traditional framework for multi-contour forecast variables. Aggregate BSS values for tornado forecasts (Fig. 5.7a) are 0.059 in the Interpolation context compared with 0.049 in the Traditional one (Fig. 5.2a), 0.096 vs. 0.076 for hail (cf. Figs. 5.7c, 5.2c), and 0.130 vs. 0.093 for wind (cf. Figs. 5.7e, 5.2e). Similarly for the Day 2 and 3 outlooks, Interpolation scores are 0.066 and 0.040 respectively compared with 0.055 and 0.028 in the Traditional verification framework. All of these differences are found to be statistically significant at a 95% significance level. In contrast, for the significant-severe forecasts (Fig. 5.7b,d,f) which use only a single 10% probability contour and the only difference stems from the choice of analysis grid (i.e. 80 km in Traditional and 0.03227° in Interpolation), the aggregate BSS differences are all 0.001 or smaller. These results strongly suggest that the act of smoothly interpolating between drawn probability isopleths has merit and results in superiorly verifying forecasts, with expected skill improvement on the order of 10% to 40% in the case of severe winds.

With regards to the effects on particular regions, the overall results are fairly similar, with higher scores over the central and eastern states and lower scores in the West, but there are some notable differences. In particular, there is a tendency for forecasts to degrade across the South and improve across the North, and this effect is especially pronounced in the severe hail and wind outlooks (Fig. 5.7c,e). South Texas is especially negatively impacted in the severe hail and wind outlooks, while the Atlantic Southeast including the Carolinas is especially negatively impacted for longer lead time outlooks (Fig. 5.7g,h). Interpolation can make an especially large difference for hail and wind, as 5% and 15% contours are comparatively frequent, and between these contours, probability ratios between frameworks of two to three are common. This effect appears to a lesser extent between higher probability isopleths. If outlooks exhibit an underforecast bias in the North and an overforecast bias in the South within the Traditional framework, perhaps due to terrain and coastal effects inhibiting predictability over southern CONUS, anticipated framework differences in forecast skill would be consistent with what is observed here. This effect affects statistical significance as well, with worse forecasts and less coverage of statistical significance of positive skill over the central and southern Great Plains, with more coverage of significantly skillful forecasts in the northern Great Plains and northern Rockies. For other variables,

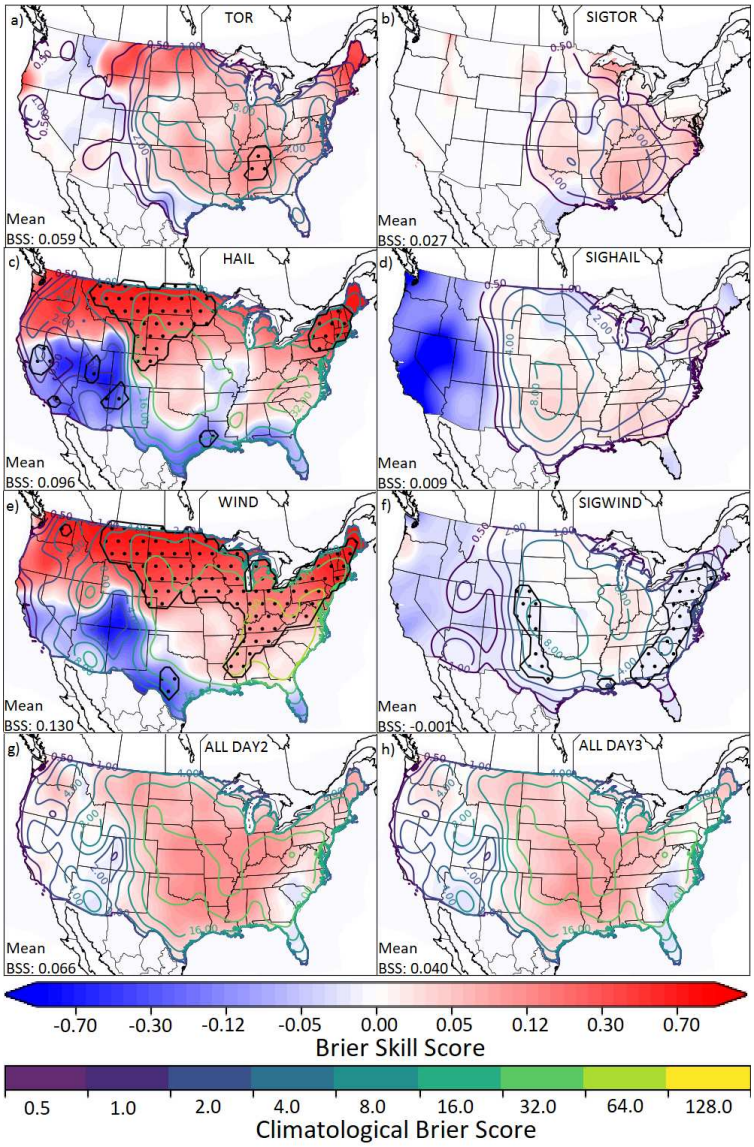


FIG. 5.7. Same as Figure 5.2, except for the Interpolation verification framework.

the spatial patterns of skill and statistical significance are more or less the same as the Traditional approach.

The annual time series of forecast skill in the Interpolation framework (Fig. 5.8) exhibit generally similar trends to the Traditional framework verification results. The climatological BSs by year (Fig. 5.8e) further validate that the higher skill scores tend to occur in more-active years, which have correspondingly higher climatological BSs. By far the most active year of the verification period, 2011 (Fig. 5.8e), also featured the most skillful forecasts, consistent with the spatial findings of Figure 5.7 and in accordance with one would typically expect forecasting very rare events (e.g. Baldwin and Kain 2006;

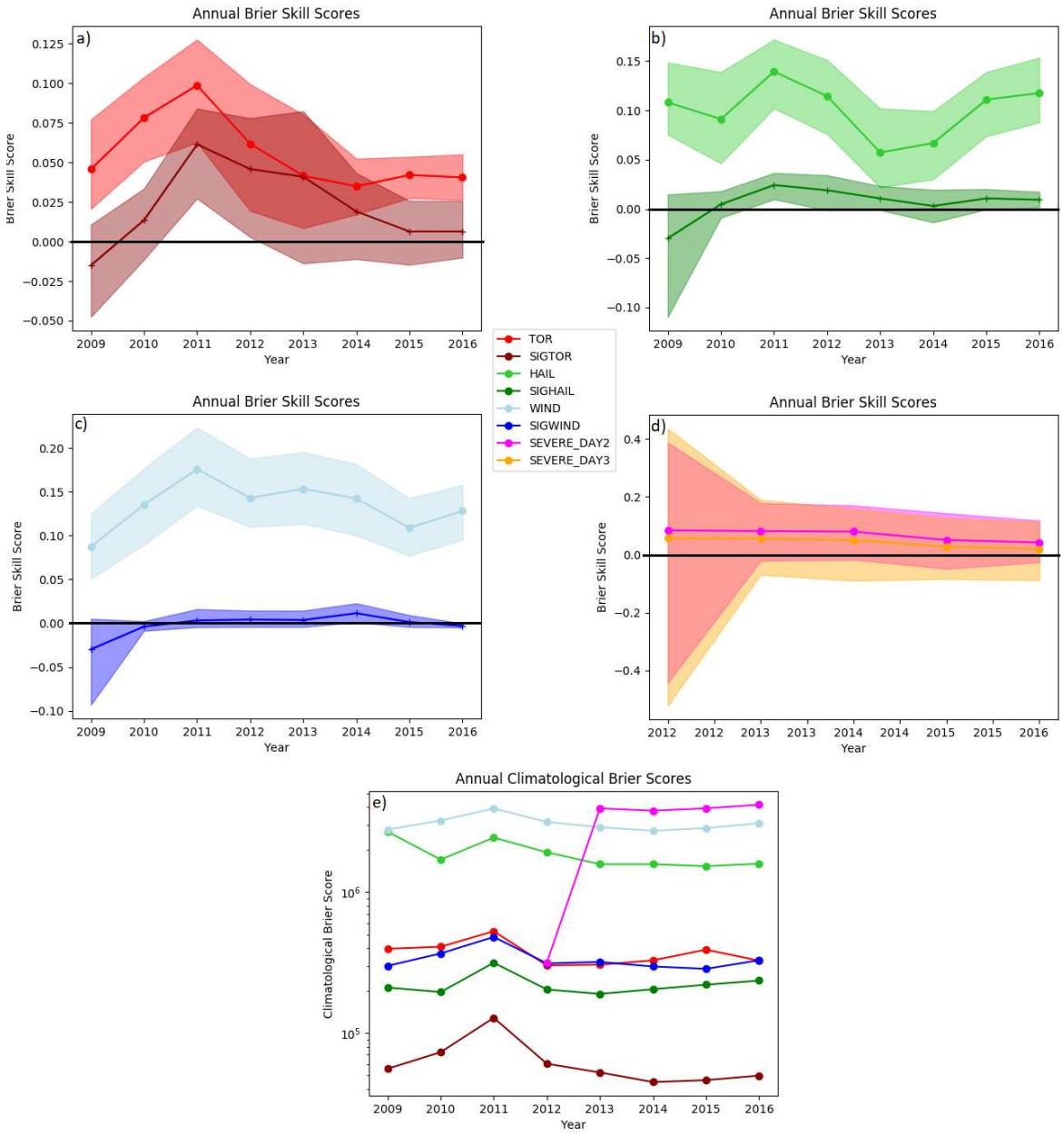


FIG. 5.8. As in Figure 5.3, except using the Interpolation verification framework. Additionally, the corresponding climatological Brier scores to panels (a)–(d) appear in panel (e) on a logarithmic axis using the same color coding as indicated in the figure legend.

Stephenson et al. 2008; Wilks 2011). Severe hail events were more common in 2009 owing to the lower threshold definition valid at that time, and skill was also correspondingly somewhat higher during that year compared to most other years in the period. The seasonal cycles of skill (Fig. 5.9), in contrast, portray both some similarities and some notable differences compared to the Traditional approach results. Tornado forecasts exhibit almost the exact same seasonal cycle of skill in both analysis frameworks (cf.

Fig. 5.4a, 5.9a), with a broad spring peak, a sharp peak in late autumn, and a skill minimum in the late summer and early autumn. However, notable differences appear in the severe hail forecasts (Fig. 5.9b). While the large pattern is generally the same as seen in the Traditional approach, the late autumn peak is more muted and, more importantly, there is a substantial performance spike in July and to a lesser extent in surrounding months that is entirely absent from the Traditional results. This same spike also appears in the severe wind forecasts (Fig. 5.9c) and is absent from those Traditional verification results as well. Like with tornadoes, Day 2 and 3 forecasts (Fig. 5.9d) exhibit very similar skill seasonal cycles in both frameworks. Unlike with years and space, there is not in general a correspondence by month between event frequency, depicted most explicitly with the climatological BSs in Fig. 5.9e, and forecast skill. Tornadoes have a primary peak in the spring and a coincident maximum in skill during that time, but the variables maximize in frequency in the late spring and early summer, and skill is largely quite low there except for the July skill spike in the interpolation framework. Severe weather environments often feature fewer higher predictability, synoptic-scale forcing scenarios such as strong fronts during this period, and this may be at least partly responsible for this apparent discrepancy (e.g., Hart and Cohen 2016).

The verification results in the CAPE-versus-shear parameter space are also largely similar in the Interpolation framework (Fig. 5.10) compared with the Traditional framework (Fig. 5.5). The primary difference, in addition to more regions of statistical significance, is the improvement in skill in the low-CAPE, high-shear region of the parameter space. The negative skill region is much smaller for tornado outlooks (Fig. 5.10a) and completely vanishes for the wind outlooks (Fig. 5.10c). Improvement is also seen, although the sign of skill remains the same, across much of the moderate-to-high shear and moderate-to-high CAPE regions of the parameter space. The region of negative skill in the very low shear (< 10 m s^{-1}) region of the parameter space, remains, however, and perhaps even amplifies to an extent.

Perhaps the biggest difference between the verification regimes emerges in the analysis of forecast reliability. First and most obviously, interpolation acts to distribute the probability between different explicit probability contours and create a continuous probability field rather than a stepwise discrete one. This results in more probability bins for the attributes diagrams in the Interpolation framework (Fig. 5.11). But perhaps more significantly, the "redistribution" of probability is not symmetric per se in that the total probability of the forecast is not conserved. The drawn contours define isopleths of constant forecast probability of value equal to the contour label. In the Traditional framework, all

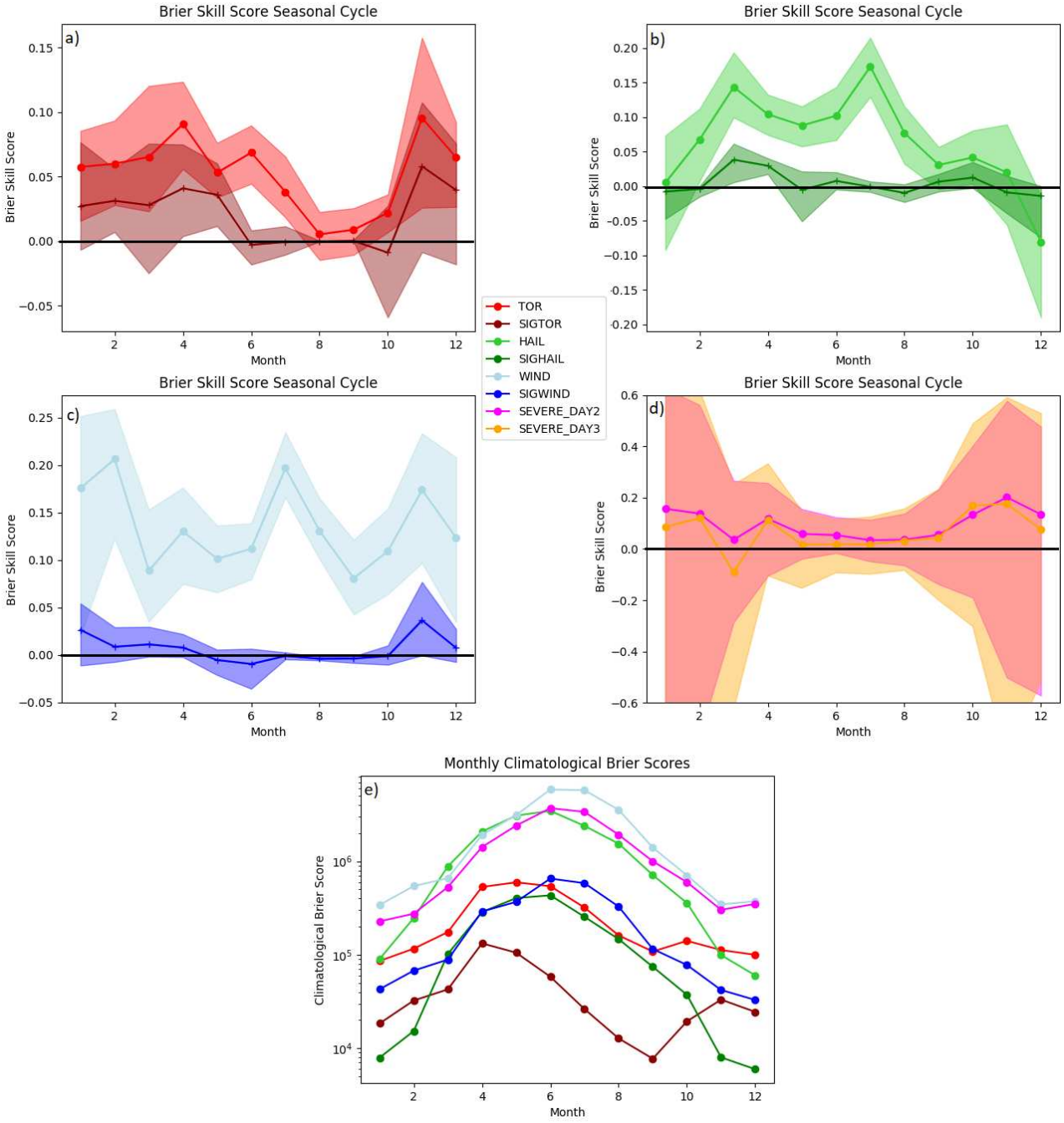


FIG. 5.9. Same as Figure 5.8, but by month of forecast issuance.

points within a given contour have an associated forecast probability in accordance with that contour label until a new interior contour is drawn. In the Interpolation framework, however, the probability values are at least that large and may be larger—up to the value of the next explicit contour level,

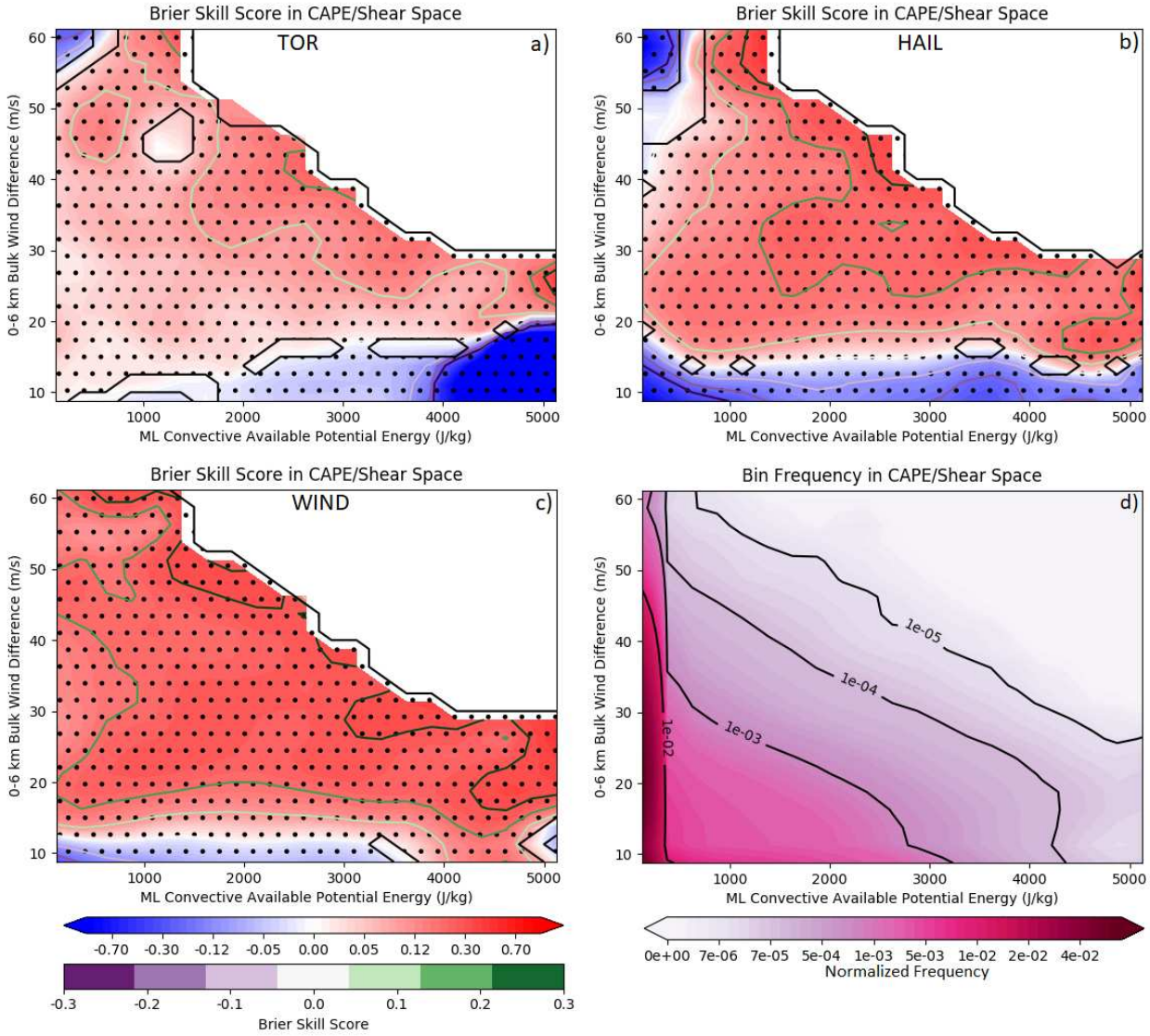


FIG. 5.10. Same as Figure 5.5, but for the Interpolation verification framework.

whether that contour was drawn or not. This essentially acts to strictly increase (or maintain) probabilities compared with the Traditional framework and never to decrease them. This has substantial implications on the reliability of the forecast sets, and reflects in the attributes diagrams by rotating all reliability lines clockwise. For negatively biased forecasts in the Traditional framework (Fig. 5.6) such as tornado Day 1 outlooks and the Day 2 and 3 outlooks, this acts to better calibrate the probabilities by bringing them closer to the one-to-one line, even though a slight negative bias is still evident at the higher probabilities. Tornadoes, however, become slightly positively biased at lower probabilities, and as a result tornado outlooks may be characterized as underconfident. The Day 1 hail and wind

outlooks, in contrast, which were better calibrated in the Traditional framework, are now positively biased except at the highest probabilities where a drastic increase in observed relative frequency with increasing forecast probability occurs above probabilities of approximately 0.5. For probabilities between 0.05 and 0.3, both hail and wind forecasts fall along or near the no skill line. One does see (Fig. 5.11a) that within explicit probability contours, observed relative frequency does tend to increase for points closer to higher numbered contours and in the center of closed contours where interpolated probabilities are higher than near contour edges, though some exceptions can be seen, particularly at lower forecast probabilities (Fig. 5.11b). Overall, these reliability findings in both the Traditional and Interpolation frameworks do contrast some with those of Hitchens and Brooks (2017), which noted positive frequency biases for essentially all forecast sets, but this discrepancy is likely attributable to differences in how that study treated the probability contours as categorical predictions as opposed to the more probabilistic treatment employed here.

Explicitly comparing skill in the two frameworks as a function of time (Fig. 5.12), one sees that Interpolations scores are consistently higher than Traditional scores from year to year (Fig. 5.12a) with the one minor exception of tornado forecasts in 2013. Both the magnitude of the differences within years and the uncertainty in the difference is largest for wind, with the smallest uncertainty in the difference for Day 2 and 3 outlooks. Statistically significantly positive differences for all variables occur for all forecast sets in at least one year, and no significant negative difference occurs for any variable in any year. In the seasonal cycle comparison (Fig. 5.12b), the interpolation adds very substantial and significant skill during the summer months for the hail and wind outlooks, maximizing with BSS enhancements of over 0.1 in July. The rest of the year, however, there is little difference, and in the case of hail, even a slight decrease in performance using interpolated forecasts. The other variables have much less dependence on forecast month in the skill difference, but generally have the largest improvement in forecasts in the late autumn and winter months, particularly November, with tornado forecasts also having a significant peak in skill difference in June.

5.5 Discussion and Conclusions

Up to 8 years of probabilistic SPC convective outlooks for Days 1–3 were gridded onto CONUS-wide grids and evaluated using two different analysis frameworks. The first, the so-called Traditional framework, uses a grid with 80-km grid spacing and does not interpolate between drawn probability contours, representing the forecast probability fields as stepwise discrete with discontinuities along

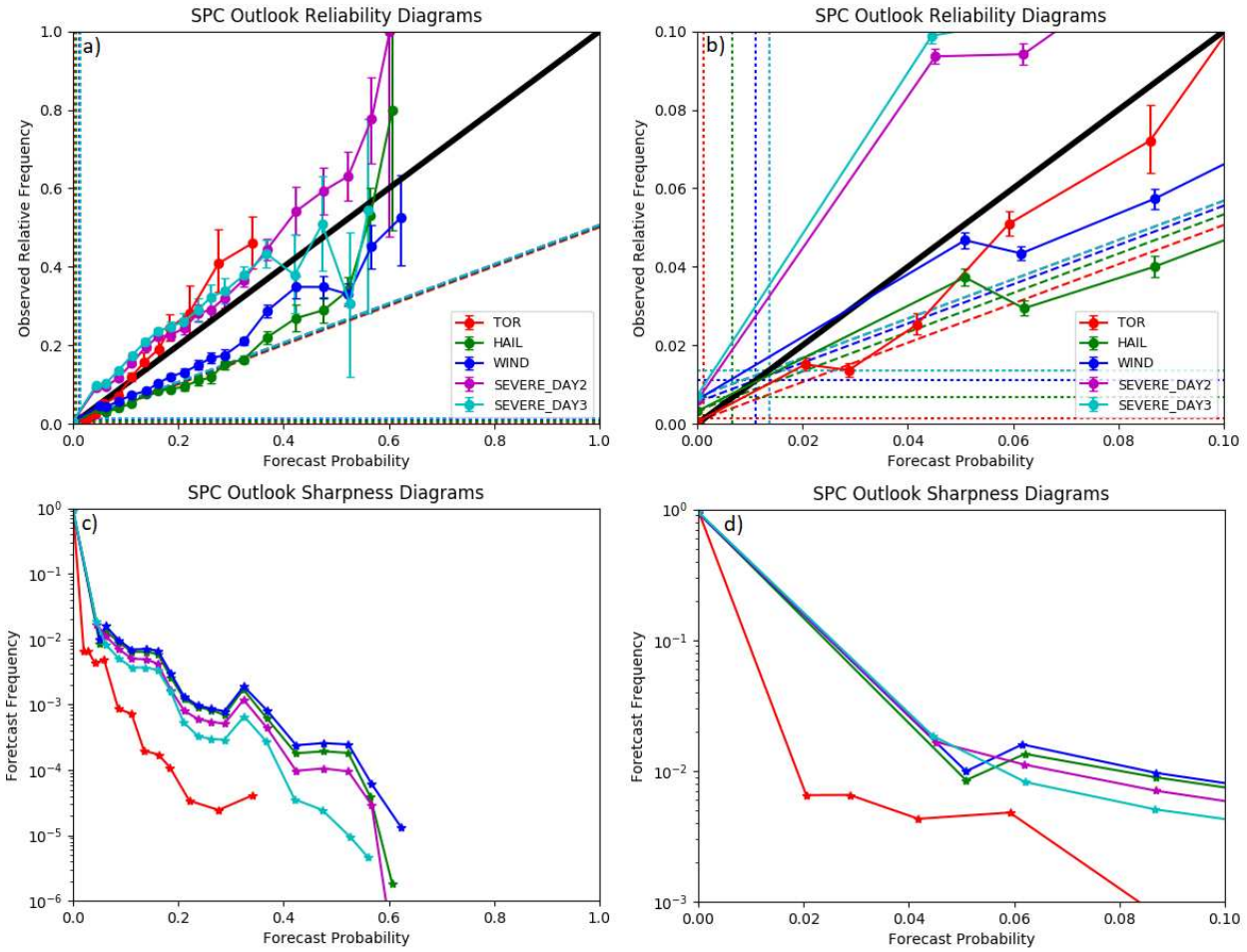


FIG. 5.11. As in Figure 5.6, except for the Interpolation framework and zoom in panels (b) and (d) is to 0.1 rather than 0.15. Probability bins are delineated by 2%, 3.5%, 5%, 7.5%, 10%, 12.5%, 15%, 17.5%, 20%, 25%, and 30% thresholds for Day 1 tornado forecasts, and by 5%, 7.5%, 10%, 12.5%, 15%, 17.5%, 20%, 22.5%, 25%, 27.5%, 30%, 35%, 40%, 45%, 50%, 55%, and 60% for all other forecast sets.

contour edges. This is performed to match historical internal verification practice at SPC and allow direct comparison with past findings. A second approach, the so-called Interpolation framework, uses a higher-resolution grid with 0.03227° spacing and instead interpolates between probability contours when two or more contour levels are depicted. Below the lowest allowable contour level for the forecast variable or when only one contour level is drawn, no interpolation is performed. The analysis period spans January 2009 through December 2016 for Day 1 forecasts and begins in September 2012 with the same end date for Day 2 and 3 outlooks. The gridded outlooks were then verified using BSSs and reliability diagrams.

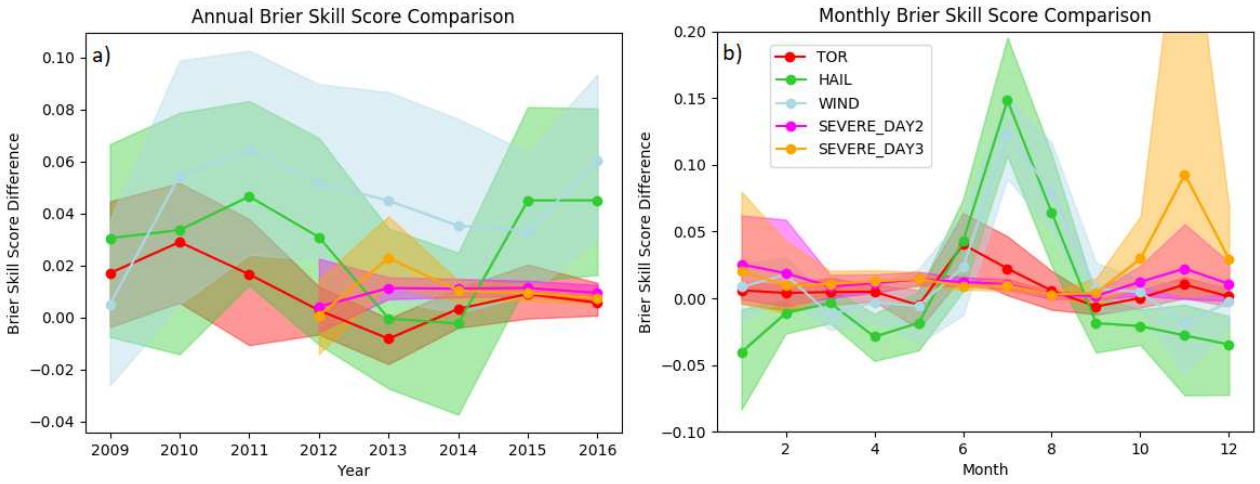


FIG. 5.12. Difference of verification results from the Interpolation framework minus results from the Traditional framework as a function of (a) forecast year and (b) forecast month for each forecast variable as indicated in the figure legend. Transparent shading corresponds to 95% confidence bounds on the difference obtained through bootstrapping and explained in greater depth in the chapter text.

In general, skill verification was best when and where events were most common, and when forecast lead time was shortest. Among Day 1 forecasts, severe winds are the most skillfully forecast by SPC in both the Traditional (BSS = 0.093) and Interpolation (BSS = 0.130) frameworks, followed by severe hail (BSS = 0.076 Traditional; 0.096 Interpolation) and then tornadoes (BSS = 0.049 Traditional; 0.059 Interpolation). The opposite trend, however, was observed at the significant-severe threshold, with significant tornadoes (BSS = 0.027) being the best forecast, and significant-severe winds (BSS = 0.00) being the worst. Forecasts were generally best in the north and eastern parts of the country and worst in the southern and western parts of the country. Little trend was seen in the skill of SPC outlooks over the analysis period, with the most skillful forecast years coinciding with the years of highest event totals. Considerable month-to-month variability was found both between adjacent months and between variables was found in SPC outlook forecast skill; highest skill was typically found in the spring and late autumn. For Days 2 and 3, forecast skill was also high during winter. Forecasts were also evaluated in the CAPE-versus-shear parameter space using classifications derived from the NARR; skillful forecasts were found over the vast majority of the parameter space. Exceptionally, forecasts in the entire very low-shear end of parameter space were found generally not to be skillful relative to climatology for all severe weather elements. Additionally, the very high-shear, very low-CAPE region of parameter space was found to be a secondary environment of forecast struggles, particularly for tornadoes and hail. In

aggregate, Interpolation framework forecasts consistently yielded higher forecast skill than analogous sets in the Traditional framework, suggesting that the intuitive practice of interpolating between the finite number of allowable contours yields superior forecast probabilities compared with simply using the probability associated with the nearest contour enclosing the point.

Forecasts were further analyzed to ascertain reliability; results contrasted based on the verification framework. For Traditional forecasts, Day 1 tornado outlooks in addition to Day 2 and 3 forecasts exhibited an underforecast bias while Day 1 hail in wind outlooks were relatively well calibrated along the spectrum. In the Interpolation framework, in contrast, hail and wind forecasts have a moderate to strong overforecast bias, while Day 2 and Day 3 forecasts have a mild underforecast bias that is alleviated compared with their Traditional counterparts; tornadoes were found to have what could be considered a mild underconfidence bias, but again to a lesser extent than in the Traditional framework.

There are several limitations or shortcomings of this work that should be noted. SPC outlooks, while being treated as probability grids for the purpose of this study and have grids generated internal to SPC in a similar manner, are not publicly disseminated or archived in grid format; instead, what is publicly available are equivalent to finite sets of probability contour outlines. This makes the quantitative SPC outlooks somewhat unconstrained everywhere not immediately on or directly adjacent to a drawn probability contour. Two sets of assumptions were made to convert these fixed contours to gridded probabilities, but one could certainly argue that—at least under some circumstances—the methodologies employed in this study would produce a grid from the contours that is appreciably different from what the human forecaster would have made had they produced a grid directly. Particularly consequential is that the fact that no interpolation could be performed outside the lowest probability contour due to the unconstrained nature of the problem, resulting in event probabilities being uniformly zero outside of SPC contours. In tandem with the fact that the lowest probability contours are generally many times larger than the climatological event frequency, this inherently inhibits SPC from gaining resolution in the lower end of the probability spectrum near the climatological frequencies. The climatological reference, in contrast, has considerable resolution in this subdomain of the probability space, and at times this can result in an uneven comparison between the SPC outlooks and the climatological reference. This effect is particularly pronounced for the significant-severe forecasts, since the climatological frequencies are so low and the forecast process effectively constrains forecasters to issue forecast probabilities of either 0 or 0.1 for any given point. This is also seen, albeit to a

lesser degree, in the verification of other fields. In particular, while forecasters draw 2% and 10% probability contours for tornado forecasts, these contours are not drawn for the remaining multi-contour probabilistic forecasts, which begin at 5% and skip to 15%. Tornado forecasts therefore have some enhanced native precision; this is apparent in Figure 5.12, where the probability interpolation is able to add substantially more to forecast skill in variables other than tornado forecasts owing to their increased comparative contour granularity. These effects all work in the mean to harm the verification of SPC outlooks relative to what they would likely be if a forecaster were to adopt the operationally impractical approach of issuing continuous, subjective forecast probabilities on a point-by-point basis. More contours, particularly at the lower ends of the probability spectrum, would allow a more quantitative interpretation of the forecast probabilities by end-users and would also result in more representative probability grids for verification. One way this could plausibly be addressed is by using the general thunderstorm contour from the categorical version of the convective outlooks as a 0% probability contour and interpolate between that and the lowest probability contour. Given that the general thunderstorm contour encompasses regions where non-thunderstorm-induced severe weather is considered possible, areas outside the general thunderstorm contour can be reasonably considered to have forecast a 0% severe probability, and so this is a reasonable attempt to gain resolution on the low probability end of the forecast spectrum. However, since this study is focused on the verification of the probabilistic convective outlooks and this contour is not included in those outlooks, applying this reinterpretation of the categorical thunderstorm line and merging it with the probabilistic convective outlooks is beyond the scope of the present study. It is, however, a worthwhile endeavor to explore in future work to attempt to address this important limitation within the confines of existing practices.

While several years of convective outlooks have been used in this study in an attempt to obtain robust verification results, one must still recognize that severe weather is a rare phenomenon, particularly during certain times of year and for particular regions of CONUS. Consequently, despite this large temporal sample, the event sample, especially in certain subclasses, is still rather small and some caution should be exercised in generating conclusions from the findings. Formal uncertainty analysis and significance testing has been performed to attempt to ascertain a realistic range of true possibilities given the data sample analyzed and ascertain which conclusions may be robustly made. This revealed, for example, that low skill scores in the West may be just an artifact of the sample owing to the small size and large variability, while comparably smaller magnitude scores to the east are significant owing to the higher climatological event frequency. An important and related, but distinct, point concerns the

pitfalls of skill calculation for a phenomenon with varying climatological frequency (Hamill and Juras 2006). As a result, added care must be exercised when comparing skill scores across variables, regions, or times where the frequency of occurrence may vary substantially. Concerns are lessened comparing across verification frameworks or between Day 2 and Day 3 outlooks when the references are identical.

Furthermore, all of this analysis uses SPC storm reports as "truth". This is a sensible choice given its continuous coverage—it is generally recognized as the best dataset for severe weather reports except for on a case-by-case basis when more thorough analysis has been conducted (e.g. Hitchens and Brooks 2012, 2014, 2017). However, this dataset has numerous limitations. Human reports of course require physical observation of either the phenomenon or lasting damage it produces. Events can occur and go unreported in rural areas where few or no people are impacted (e.g. Anderson et al. 2007). Nocturnal events, and events in heavily forested areas or areas of complex terrain also pose reporting challenges, particularly for tornadoes, due to the difficulty in visual observation of the event (e.g. Anderson et al. 2007). For a multitude of reasons, including but not limited to the increasing population density (e.g. Verbout et al. 2006), increases in radar coverage (e.g. Agee and Childs 2014), and improved spotter networks and reporting practices (e.g. Trapp et al. 2006; Doswell 2007), there have also been numerous changes over time in report frequency and density. Fortunately, from a climatological perspective, the period of record employed by this paper is rather short, and most of these report trend considerations are not significant concerns. The unreliability and inconsistency in EF0 tornado reports (e.g. Anderson et al. 2007), the change in severe hail criteria (Ferree 2009), and particularly reporting issues associated with severe convective winds (e.g. Trapp et al. 2006; Edwards and Carbin 2016), all present additional concerns that can harm the reliability of the database and adversely impact the validity of the verification results such as those presented herein. An additional, but related, limitation concerns the actual treatment of the reports in this study. Here, reports have been used to form binary grids of event observance, which doesn't account for the density of reports within a verification grid box like a practically perfect approach (e.g. Hitchens and Brooks 2014) would. However, given the high resolution nature of the verification grid, this is not considered to be a substantial concern in the end results.

Despite these limitations, this analysis can provide utility in a variety of ways. It can help end-users determine under which situations SPC outlooks exhibit more or less skill to the extent this may assist with uncertainty assessment and decision making. As a "gold standard" of severe-weather forecasting, these results can also help direct both operational forecasters and researchers into which particular

areas could use further attention, both in the forecast process and in modeling and physical understanding. The results, and particularly the reliability findings, may invoke changes in forecasting philosophy that are of benefit to end users (Hitchens and Brooks 2012). For example, the reliability results suggest that, at least under some circumstances, SPC forecasters may benefit from increased conservatism with their Day 1 hail and wind contours, and more liberal usage of probability contours for their Day 2 and Day 3 forecasts. This analysis also provides robust, quantitative benchmarks for comparison of newly developed severe weather forecast guidance. In isolation, a skill score—other than 0 or 1—doesn't have much quantitative physical meaning. A positive value less than one indicates non-perfect forecasts that nevertheless have skill over the reference, of course, but the interpretation of a specific number—0.2, for example—depends both on the quality of the reference forecast and the feasibility of perfect forecasts. Having benchmarks against a robust, respected standard such as the SPC outlooks is particularly important in the severe-weather forecast problem since, unlike other forecast predictands, operational models are not able to simulate or forecast most severe weather phenomena directly, further reducing the possibility to compare new methods with existing guidance and contextualize the results. There have been numerous forays into improving aspects of severe-weather forecast guidance in recent years, some already in operational use (e.g. Brimelow et al. 2006; Sobash et al. 2011) and others more recent work under development (e.g. McGovern et al. 2014; Sobash et al. 2016a,b; McGovern et al. 2017; Gagne et al. 2017); having these results will help better place their results and future like studies in the context of existing forecasts.

Future verification work will seek to perform similar analysis for flash flooding using Excessive Rainfall Outlooks issued by the Weather Prediction Center and explore other issues in flood and flash flood verification (e.g. Drobot and Parker 2007; Gourley et al. 2013; Schroeder et al. 2016; Herman and Schumacher 2016a, among others). From there, we will also explore the overlaps and intersections of probabilistic forecasts for different weather hazards to glean additional operational insight on forecasting performance and challenges in predicting these elevated threat scenarios, such as concurrent and collocated tornado and flash flood hazards (Nielsen et al. 2015). Other work will seek to provide improved gridded probabilistic forecast guidance for these high-impact weather hazards to help yield improvement in future verification of these operational forecasts.

CHAPTER 6

FORECASTING SEVERE WEATHER WITH RANDOM FORESTS

6.1 INTRODUCTION

Severe weather is comprised of three distinct phenomena: 1) the presence of one or more tornadoes of any intensity, 2) the presence of 1 in. (2.54 cm) or larger hail, or 3) convectively-induced wind gusts of at least 58 mph (93 km h^{-1}). Beyond this, tornadoes of F2 or EF2 strength or greater, hail 2 in (5.08 cm) or larger in diameter, or wind gusts of at least 74 mph (119 km h^{-1}), pose particularly elevated threats to life and property and are considered supplementarily in a "significant severe" weather class (Hales 1988; Edwards et al. 2015). Collectively, these hazards have inflicted more than 1100 fatalities and $36.4B in damages across the contiguous United States (CONUS) in 2010–2018 (NWS 2018). While inherently dangerous and damaging phenomena, accurate severe weather forecasts can increase preparedness and help mitigate inclement weather losses.

The hazards associated with severe weather are further encumbered by the challenge in accurately forecasting the phenomena. Due to the very small spatial scales associated with severe weather, it is often exceedingly difficult to model dynamically with operational weather models. Production of large hail involves a plethora of very small-scale microphysical processes which are necessarily parameterized in numerical models. The microphysical simplifications involved to hasten production of operational model output, including bulk rather than bin schemes (e.g. Khain et al. 2015), single moment microphysics (e.g. Igel et al. 2015), and in some cases, not having an explicit category for hail at all (e.g. Hong and Lim 2006), all make direct prediction of severe hail from operational dynamical model output a perilous task. Tornadoes are in some respect even more difficult to simulate; while numerical tornado simulations have been conducted in a research setting (e.g. Orf et al. 2017), they occur on much too small of spatial scales to be resolved by any operational model. In forecasting severe weather, it is therefore necessary to relate simulated environmental factors across various scales, from storm-scale up to the synoptic scale, to severe weather risk. This is routinely performed in the human severe weather forecast process (e.g. Johns and Doswell 1992; Doswell III 2004; Doswell III and Schultz 2006), but in terms of producing automated guidance, statistical in addition to dynamical approaches are necessary for this important forecast problem.

CONUS-wide operational severe weather forecasts are issued routinely by the Storm Prediction Center (SPC) for Days 1–8 via their convective outlooks (Edwards et al. 2015). In these products, forecasts are issued for 24-hour 1200–1200 UTC periods, and are given as probabilities of observing the corresponding severe weather phenomenon within 40 km of the forecast point during the period. An additional categorical risk outlook is provided for Days 1–3, defined based on the probabilistic outlook values. For Day 1, SPC issues separate probabilistic outlooks for each of the three severe weather predictands; for Day 2 and beyond, they are treated collectively in a single outlook. In the forecast process, the forecaster draws from a discrete set of allowable probability isopleths, where applicable. For Day 1 hail and wind outlooks, and Day 2 and 3 outlooks, permitted isopleths are 5%, 15%, 30%, 45%, and 60%; Day 1 tornado outlooks include 2% and 10% probability contours as well. For Day 4 and beyond, only 15% and 30% contours are issued, and for signifcant severe risk, only a single 10% contour is drawn. Chapter 5 gives more information on SPC's forecasting process, including historical changes to severe weather and product definitions; more information can also be found in Hitchens and Brooks (2014), Edwards et al. (2015).

A limited number of published studies have quantified the skill of these convective outlooks and examined their strengths and weaknesses. Hitchens and Brooks (2012) investigated the skill of Day 1 categorical outlooks, and this effort was expanded to include evaluation of Days 2 and 3—among other additions—in Hitchens and Brooks (2014). Early published efforts to verify SPC's convective outlooks probabilistically (e.g. Kay and Brooks 2000) have received renewed attention in Hitchens and Brooks (2017) and more formally in Chapter 5. Collectively, these studies have demonstrated improving skill in short-to-medium range severe weather forecasts in association with improved numerical weather prediction (NWP; e.g. Hitchens and Brooks 2012, 2014), though advances have been stagnating somewhat in recent years. As demonstrated also in Chapter 5, forecast skill is highest at the shortest lead times and gets progressively lower with increasing lead time (e.g. Hitchens and Brooks 2014). In general, wind is the most skillfully predicted severe weather phenomenon with tornado outlooks exhibiting the lowest skill, but this is reversed for significant severe events (Hitchens and Brooks 2017). Additionally, skill was found to be maximum over the Midwest and Great Plains, and lowest over the South and West. Outlooks are generally most skillful in the winter and spring, and least skillful in the late summer into early autumn. Furthermore, skill is high when at least moderate amounts of both CAPE and wind shear are present, but struggle when CAPE is limited and shear is large, or vice versa (e.g. Sherburn and Parker

2014, , as shown also in Chapter 5). As noted above, SPC's convective outlooks are based on only a finite set of probability contours, producing discontinuous jumps in gridded probability fields. Chapter 5 demonstrated that forecast skill is improved, albeit not uniformly, when probabilities are interpreted as interpolated between confining human-drawn probability contours. In these interpolated outlooks, hail and wind forecasts exhibit an overforecast bias, while tornado and Day 2 and 3 outlooks exhibit a slight underforecast bias. Moreover, the evaluation of Chapter 5 provides quantitative benchmarks for placing newly developed statistical guidance in the place of existing operational performance.

There have been numerous forays into statistical prediction of severe weather in existing literature. These include applications for statistical prediction of tornadoes (e.g. Marzban and Stumpf 1996; Alvarez 2014; Sobash et al. 2016a; Gallo et al. 2018), hail (e.g. Marzban and Witt 2001; Brimelow et al. 2006; Adams-Selin and Ziegler 2016; Gagne et al. 2017), wind (e.g. Marzban and Stumpf 1998; Lagerquist et al. 2017), and severe weather more broadly (e.g. Gagne et al. 2009; Sobash et al. 2011; Gagne et al. 2012; Sobash et al. 2016b). Many of these studies have applied machine learning (ML) to the prediction task; in general, ML techniques have demonstrated great promise in applications to high-impact weather prediction (e.g. McGovern et al. 2017). In addition to severe weather, ML has demonstrated success in forecasting heavy precipitation (e.g. Gagne et al. 2014; Whan and Schmeits 2018, , in addition to Chapters 3 and 4), cloud ceiling and visibility (e.g. Herman and Schumacher 2016b; Verlinden and Bright 2017), and tropical cyclones (Loridan et al. 2017; Alessandrini et al. 2018). Furthermore, automated probabilistic guidance, including ML algorithms, have been identified as a priority area for integrating with the operational forecast pipeline (e.g. Rothfusz et al. 2014; Karstens et al. 2018). However, many past applications have focused on either much shorter timescales, such as nowcast settings (e.g. Marzban and Stumpf 1996; Lagerquist et al. 2017), or on much longer timescales (e.g. Tippett et al. 2012; Elsner and Widen 2014; Baggett et al. 2018), with lesser emphasis on the day-ahead time frame and very little model development in the medium-range (e.g. Alvarez 2014). Furthermore, many studies have operated over only a regional domain (e.g. Elsner and Widen 2014) and no study to date has exactly replicated the operational predictands of SPC's convective outlooks, making it difficult to make one-to-one comparisons between ML study outcomes and operational performance.

One such ML algorithm that has demonstrated success in numerous previous high-impact weather forecasting applications (e.g. McGovern et al. 2011; Ahijevych et al. 2016; Herman and Schumacher 2016b; Gagne et al. 2017; Whan and Schmeits 2018) is the Random Forest (RF; Breiman 2001). This

study seeks to apply RF methodology to the generation of calibrated probabilistic CONUS-wide forecasts of severe weather with predictands analogous to those of SPC convective outlooks in the hope that the guidance produced can be used to improve operational severe weather forecasting. Section 2 provides further background and describes the data sources used and methodologies employed to create and evaluate these forecasts. Section 3 investigates the RF-derived severe weather forecasting insights gleaned from the trained models. Section 4 evaluates the RF forecasts produced and places the results in the context of existing operational forecasts. Section 5 concludes the paper with a synthesis of the findings and a discussion of their implications.

6.2 Data and Methods

6.2.1 Overview

Chapters 3 and 4 extensively explored the utility of applying RFs and other machine learning algorithms towards post-processing global ensemble output to forecast locally extreme precipitation events across CONUS at Days 2–3. This study follows analogous methodology. For the sake of brevity, several of the RF model configuration choices selected in this study are motivated by the findings of Chapter 3 rather than reperforming all the same experiments for this forecast problem. Informal replications of those experiments with the severe weather predictands used in this study produced similar findings (not shown).

An RF (Breiman 2001) is an ensemble of unique, weakly-correlated decision trees. A decision tree makes successive splits into branches, with each split based on the value of a single input predictor. The splitting predictor and the value associated with each branch is determined by the combination that best separates severe weather events from non-events in the supplied model training data. This process then continues for progressively smaller branched subsets based on only the training data that satifies the previous branching conditions. This process continues until a termination criterion is satisfied, either because all of the remaining training examples are either all events or all non-events, or because there are too few remaining training examples to continue splitting. At this point, a "leaf" is produced which makes a forecast according to the proportion of remaining training examples associated with each event class. In real-time forecasting, the new inputs are supplied and the tree is traversed from its root according to the input values until a leaf is reached, which becomes the real-time prediction of the tree. An RF produces numerous unique decision trees by considering different subsets of training data

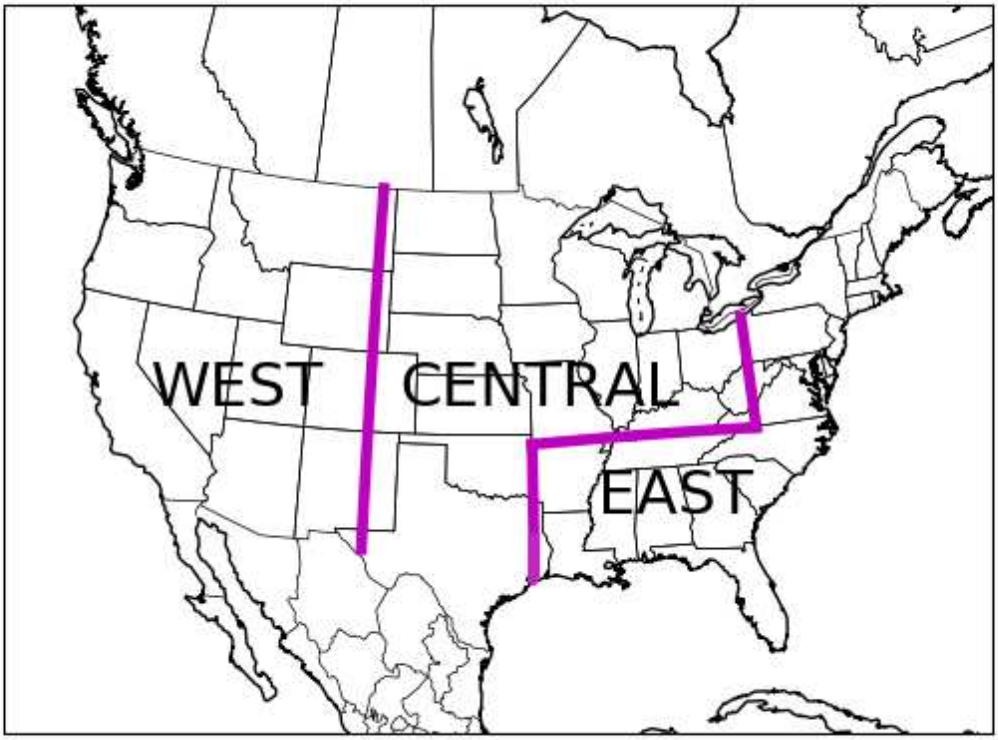


FIG. 6.1. Map depicting the training regions of CONUS for the statistical models used in this study.

and input features, or predictors, for each tree generation process. An RF's forecast is simply calculated as the mean probabilistic forecast issued by the trees within the forest (e.g. Breiman 2001).

Based on different diurnal and seasonal climatologies (e.g. Brooks et al. 2003; Nielsen et al. 2015; Krocak and Brooks 2018), and due to differing regimes and storm systems primarily responsible for severe weather across CONUS (e.g. Smith et al. 2012), the country is partitioned into three regions as shown in Figure 6.1. This study develops separate RFs for each of the three regions of CONUS, with unique forests trained also for each of the five predictand, lead time combinations: 1) Tornado Day 1, 2) Hail Day 1, 3) Wind Day 1, 4) Severe Day 2, and 5) Severe Day 3. For the Day 1 models, the severity levels of the category are retained using a 3-category predictand (none, non-"significant" severe, "significant" severe), while the severity levels are aggregated for longer lead times. Each of the 15 forests is trained using a nine year historical record spanning 12 April 2003–11 April 2012. As noted above, the focus of this study is on the model evaluation rather than on involved sensitivity experiments and parameter

tuning. Models were evaluated using Python's Scikit-Learn library (Pedregosa et al. 2011) as in Chapters 3 and 4; deviations from defaults for this study were made based on a combination of performance considerations and computational constraints. The only parameters varied were the forest size B and minimum number of training examples required to split an impure node in a decision tree, Z. For the interested reader, the final values used are furnished in Table 6.2.

RF predictor information comes from the GEFS/R dataset (Hamill et al. 2013). The GEFS/R is a global, convection-parameterized 11-member ensemble with T254L42 resolution—which corresponds to an effective horizontal grid spacing of ∼55 km at 40° latitude—initialized once daily at 0000 UTC beginning in December 1984. Perturbations are applied only to the initial conditions, and are made using the ensemble transform with rescaling technique (Wei et al. 2008). The ensemble system used to generate these reforecasts is nearly static throughout its 30+ year period of coverage, though updates to the operational data assimilation system over time have resulted in some changes in the bias characteristics of its forecasts over the period of record (Hamill 2017). Most surface (or column-integrated) fields are preserved on the native Gaussian grid ($\sim 0.5°$ spacing), while upper-level and some other fields are available only on a $1° \times 1°$ grid. Based on findings from Chapter 3, this study derives predictors from the GEFS/R ensemble median. Model training employs a 9-year training period, using daily initializations from 12 April 2003–11 April 2012. Temporally, forecast fields are archived every three hours out to 72 hours past initialization, and are available every six hours beyond that. Accordingly, the RFs trained in this study use 3-hourly predictors for Day 1 and 2 forecasts, and 6-hourly temporal resolution for Day 3.

Several different GEFS/R simulated atmospheric fields with known or postulated physical relationships with severe weather are used as RF predictors (Table 6.1), referred to interchangeably as "features". These include surface-based CAPE and CIN, 10-meter winds (U10, V10, UV10); surface temperature and specific humidity (T2M, Q2M), precipitable water (PWAT), accumulated precipitation (APCP), wind shear from the surface to 850 and 500 hPa (MSHR, DSHR), and mean sea level pressure (MSLP). For Day 1, three additional predictors are supplied: surface relative humidity (RH2M), lifting condensation level height above ground (ZLCL), and surface–850 hPa storm relative helicity (SRH), approximated following Ramsay and Doswell (2005) as described in the Appendix. Some of these variables are archived natively by the GEFS/R, while others are derived based on stored fields that are available. The full list of fields, their class, whether they are natively archived or derived, and the grid from which they are sampled is included in Table 6.1. Descriptions of how derived variables are calculated

is provided in Chapter 6.2.2. For each field, in addition to sampling the temporal variation of the fields throughout the forecast period as noted above, spatial variations in the simulated fields are included as inputs to the RF. Specifically, predictors are constructed in a forecast point-relative sense, with predictors up to three grid boxes (1.5° or 3°, depending on the predictor) displaced in any horizontal direction relative to the forecast point. Forecasts are made on the Gaussian grid; for predictors on the 1° grid, the nearest point to the Gaussian point is used as the central point on that grid. In addition to this suite of meteorological predictors, forecast point latitude, longitude, and the Julian day associated with the forecast are included as predictors as well.

6.2.2 Derived Variables

6.2.2.1 Relative Humidity

Relative humidity is calculated as a function of specific humidity q, temperature T, and pressure P, all of which are natively archived. The surface pressure is assumed to be negligibly different from the air pressure two meters above ground. The variables are related through Clausius-Clapeyron, as employed in Bolton (1980) and elsewhere:

$$RH = \frac{0.263 * P * q}{e^{\frac{17.67(T-T_0)}{T-29.65}}} \tag{6.1}$$

for temperature in K and pressure in Pa, where a reference temperature T_0 of 273.15 K is used. RH is calculated on the 1° grid, since surface pressure is only archived on this grid.

6.2.2.2 Lifting Condensation Level Height

An exact formula for the LCL height as a function of temperature, pressure, and relative humidity was described in Romps (2017), and that formulation is employed here. Relative humidity is not natively archived and is supplied to this formulation as calculated in the previous subsection.

6.2.2.3 Wind Shear

SHEAR850 and SHEAR500—bulk wind differences between two vertical levels—are calculated straightforwardly:

$$SHEAR850 = \sqrt{(U_{850} - U_{10m})^2 + (V_{850} - V_{10m})^2} \tag{6.2}$$

$$SHEAR500 = \sqrt{(U_{500} - U_{10m})^2 + (V_{500} - V_{10m})^2} \tag{6.3}$$

Winds were used on the 1° grid for both levels.

6.2.2.4 Storm Relative Helicity

Limited information is available from which to calculate SRH, but given its demonstrated importance in severe environments (e.g. Kuchera and Parker 2006; Parker 2014), the forecast information is used to generate as accurate of SRH estimates as possible. Low-level vertical winds on pressure levels are provided at only 1000, 925, 850, and 700 hPa—quite insufficient for use in an SRH calculation. In height, winds are provided at only 10 and 80 meters above ground level—again, insufficient. Hybrid levels provide some resolution in the low-levels, with winds archived on the 0.996, 0.987, 0.977, and 0.965 sigma levels; geopotential heights are provided for these levels as well. Thus, for calculating SRH from the surface to 850 hPa, five layers are used: 1) 10m–0.996σ, 2) 0.996σ–0.987σ, 3) 0.987σ–0.977σ, 4) 0.977σ–0.965σ, and 5) 0.965σ-850 hPa. Storm motion is estimated as 75% and 30° to the right of the mean wind, a common heuristic employed in Ramsay and Doswell (2005) and others. The mean wind is estimated as the average of the wind at 850, 500, and 200 hPa:

$$\overline{U} = \frac{U_{850} + U_{500} + U_{200}}{3}; \overline{V} = \frac{V_{850} + V_{500} + V_{200}}{3} \tag{6.4}$$

Accordingly:

$$SRH850 = \sum_{l=1}^{5} \max(0, SRH_l) \tag{6.5}$$

where

$$SRH_l = (Z_l - Z_{l-1})\left((\overline{V_l} - V_{st})\frac{U_l - U_{l-1}}{Z_l - Z_{l-1}} - (\overline{U_l} - U_{st})\frac{V_l - V_{l-1}}{Z_l - Z_{l-1}}\right) \tag{6.6}$$

with

$$\overline{U_l} = \frac{U_l + U_{l-1}}{2}; \overline{V_l} = \frac{V_l + V_{l-1}}{2} \tag{6.7}$$

and

$$U_{st} = \sqrt{0.75} * (\overline{U}\cos(-30^\circ) - \overline{V}\sin(-30^\circ)) \tag{6.8}$$

$$V_{st} = \sqrt{0.75} * (\overline{U}\sin(-30^\circ) + \overline{V}\cos(-30^\circ)) \tag{6.9}$$

6.2.3 Evaluation Framework

Trained RFs are evaluated in two distinct ways. First, in Section 3, the statistical relationships diagnosed by the RFs are investigated to determine the insights gleaned about the forecast problem and assess whether the models are making predictions in ways consistent with our external understanding of the forecast problem. Due to the number and size of trees in a forest, it is not practical to investigate

the complete structure of each tree in the forest; instead, summary statistics are used to capture the extent of use of different aspects of supplied forecast information in generating a final prediction. In particular, this is done by means of feature importances (FIs). Though there are several ways that FIs can be quantified (e.g. Strobl et al. 2007, 2008), this study uses the so-called "Gini importance" metric for consistency with prior ML research in the community (e.g. Pedregosa et al. 2011; Whan and Schmeits 2018). A single FI is attributed to each input feature, and may be conceptualized as the number of splits based on the given feature, weighted in proportion to the number of training examples encountering the split (Friedman 2001). This is summed over each split in the tree for each tree in the forest, and then normalized so that the sum of all FIs is unity. FIs thus range between zero and one, with larger values indicating that the associated predictor has more influence on the prediction values. In the extremes, an FI of zero means that the predictor has no influence on the prediction made by the RF, while a value of one indicates that the value of the associated predictor uniquely specifies the predictand. As noted above, input predictors to the RF vary in associated simulated forecast field, forecast time, and in space relative to the forecast point. In many cases, it is convenient to present FIs summed over one or more of these dimensions to provide a summary aspect of which fields, times, and locations are being most and least used in generating predictions for different severe weather phenomena.

Second, in Chapter 6.4, the probabilistic performance of the models is evaluated. The trained RFs are used to generate probabilistic convective outlooks over 4.5 years of withheld model data spanning 12 April 2012–31 December 2016. Model skill is evaluated through the Brier Skill Score (BSS; Brier 1950), using an informed climatological reference as described in Chapter 5, while forecast calibration is assessed via reliability diagrams (Murphy and Winkler 1977; Bröcker and Smith 2007; Wilks 2011). While forecasts are evaluated in aggregate, they are also assessed both spatially and seasonally in order to assess the times and locations where the RFs perform most and least skillfully. Additionally, following Chapter 5, outlook skill is evaluated based on the large-scale environmental conditions associated with the forecast, as quantified based on CAPE and deep-layer bulk wind difference (hereafter referred to as shear) in the North American Regional Reanalysis (NARR; Mesinger et al. 2006). Findings are contextualized by comparing the RF performance against SPC convective outlooks for the same predictands issued with comparable lead times. Consistent with Chapter 5, Day 1 outlooks evaluated in this study come from the 1300 UTC forecast issuance, while Day 2 and 3 outlooks come from the 0100 CT (0600 or 0700 UTC) and 0230 CT (0730 or 0830 UTC) forecast issuances, respectively. Because the interpolated probability grids verified more skillfully than the uninterpolated outlooks (demonstrated in Chapter 5),

the interpolated grids are used as the benchmark for comparison in this study. In most cases, the entire evaluation period is used for the comparison; due to data availability constraints, a slightly shorter 13 September 2012–31 December 2016 period is used for Day 2 and 3 verification, while 12 April 2012–31 December 2014 is used for the evaluation in the CAPE-versus-shear parameter space. As a final evaluation of the operational utility of the ML-based forecast guidance provided by the trained RFs, a weighted blend of the SPC and RF-based convective outlooks is evaluated over the same period; the level of skill improvement, if any, quantifies the value added by the addition of the ML guidance to the operational forecast pipeline. Weights are supplied based on the BSS of the two component outlooks using three temporally-contiguous quarters of the evaluation period that excludes the forecast being weighted, based on the following formula:

$$W_{SPC} = \frac{\frac{1}{1-BSS_{SPC}}}{\frac{1}{1-BSS_{SPC}} + \frac{1}{1-BSS_{RF}}}; W_{RF} = 1 - W_{SPC} \tag{6.10}$$

In the event that one BSS is negative, the weight associated with that forecast is set to zero with the other set to one. In this way, if either forecast set has no climatology-relative skill on the portion of the evaluation period used to generate the weights, it does not contribute to the blended forecasts, while if either forecast set is perfect, it completely determines the blended forecast. Statistical significance of both the absolute climatology-relative skill and comparisons between forecast sets are assessed using bootstrapping whereby random samples of forecast days are sampled with replacement among the evaluation period to produce a realistic range of Brier and climatological Brier Scores for each evaluated forecast set or forecast set comparison. Other uncertainty analysis follows the methods of Chapters 3 and 5; more details may be found there.

6.3 Results: Model Internals

Predictive utility of different simulated atmospheric fields (Fig. 6.2) is found to vary somewhat by forecast region and severe predictand. Under almost all circumstances, CAPE is found to be the most predictive severe weather predictor by a fair margin, particularly for predicting hail and wind. CIN is generally identified as far less predictive, but still more so than other fields. The West is an exception, with CIN identified as quite predictive of hail and especially severe wind, with CIN actually having higher FIs than CAPE for wind (Fig. 6.2a). All fields contribute some to the output of each model, with a relatively balanced distribution outside of the more predictive fields. In addition to CAPE and CIN, DSHR is found to be fairly predictive as well, and this is most evident for hail (Fig. 6.2). For tornadoes,

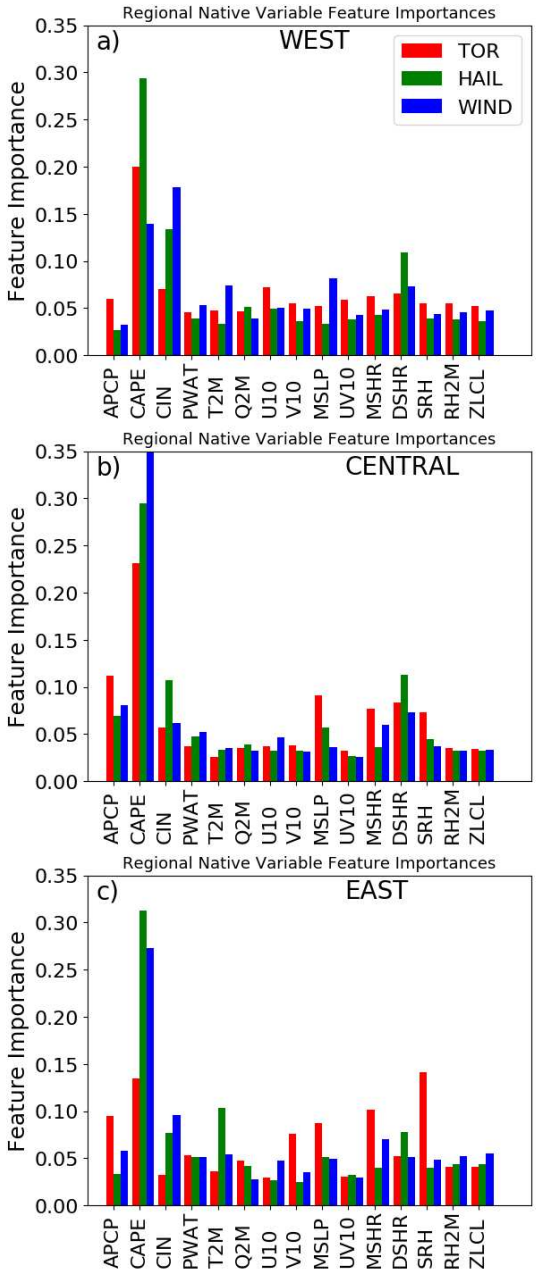


FIG. 6.2. FIs aggregated by atmospheric field for the Day 1 models in the WEST, CENTRAL, and EAST regions in panels (a)–(c), respectively. Red bars correspond to FIs for the tornado predictive model, green bars to the hail predictive model, and blue bars to the wind predictive model for each region.

shear over a shallower layer in MSHR is found to be equally (e.g. Fig. 6.2b) or more (Fig. 6.2c) predictive than DSHR, and one of the more predictive variables overall. Other variables with high RF FIs for tornadoes include APCP, MSLP, and SRH. The high FI attributed to model APCP in predicting tornadoes may be surprising, but heavy precipitation is often found to be associated with low-level rotation (e.g. Smith

et al. 2001; Hitchens and Brooks 2013; Nielsen and Schumacher 2018). MSLP serves to characterize the synoptic environment and help distinguish favorable from unfavorable environmental conditions for tornadoes. SRH has often been noted as a predictive variable for determining tornado potential (e.g. Davies and Johns 1993; Thompson et al. 2007), and is found to be the most predictive field in the East (Fig. 6.2c). Overall, the RFs are largely following conventional wisdom about human forecasting of severe weather: CAPE and shear are some of the most important fields to consider, shear should be considered over a deeper layer for hail and wind to ascertain supercell potential and over a shallow layer and in conjunction with helicity for tornado prediction in order to ascertain potential for low-level rotation, and the kinematics play a more significant role overall for tornadoes than for severe hail and wind. The RFs have simply learned these facts objectively and empirically based on analysis of many historical cases, and have provided a quantitative assessment of their findings.

In predicting any severe weather beyond Day 1 (Fig. 6.3), the trends largely follow the findings for hail and wind in their respective regions. Considering that the vast majority of severe observations are either hail or wind, that the FIs track those of hail and wind more closely than tornadoes is not surprising. CAPE and CIN are about equally predictive of severe weather at Days 2 and 3 in the West (Fig. 3a), with DSHR the next most predictive. The relative ranking mostly holds for the Central and East regions (Fig. 6.3b,c), although CAPE is much more predictive than CIN, especially in the Central region. MSHR becomes increasingly important with longitude, and is interestingly identified as more indicative of severe weather in the East region at these longer lead times. Importances are mostly similar between days, though CAPE importance tends to decline slightly from Day 2 to 3 (Fig. 6.3) and is distributed among the other fields. This is perhaps attributable to the noisy and highly sensitive nature of the CAPE field yielding less predictive utility with increasing forecast lead time and associated increasing uncertainty.

FI time series (Fig. 6.4) reveal a clear diurnal peak in importance of model information throughout the forecast period, although in all cases the peak is much more uniformly distributed relative to the diurnal event climatology in the region. In the extreme, tornadoes in the West (Fig. 6.4a), there is little peak at all. In some cases, notably in the East (Fig. 6.4c,f,i), the importance peak is aligned with the climatological event maximum, while in other situations, it leads (e.g. Fig. 6.4h) or lags (e.g. Fig 6.4d,e,g) it. In some cases, this could be an initiation bias—particularly in the lagging cases—while it could also be attributable to the forecasted pre- (or post-) event environment being more predictive than the simulated evolution at event time. Breakdowns into thermodynamic and kinematic variables

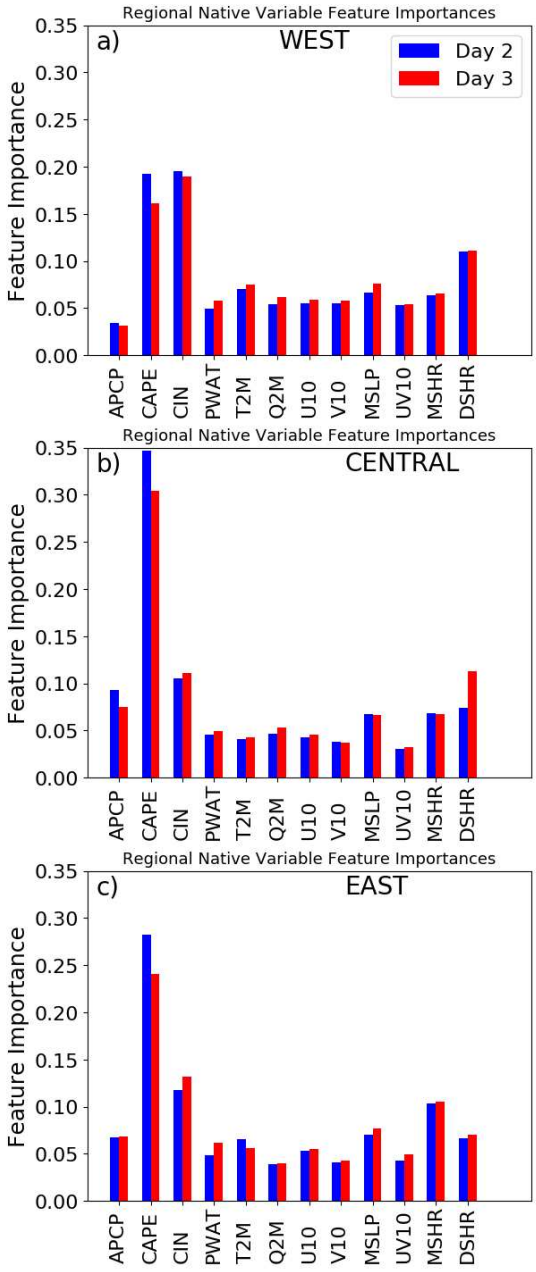


FIG. 6.3. Same as Figure 6.2, but for the Day 2 and 3 models. Day 2 and 3 FIs are indicated in red and blue bars, respectively.

(Table 6.1) reveals that the thermodynamic variables are much more predictive of hail and wind than the kinematics, while the two classes are about equally predictive for tornadoes. Furthermore, while the thermodynamics have a sharp diurnal peak, the importance of the kinematic variables has little temporal dependence throughout the forecast period (Fig. 6.4).

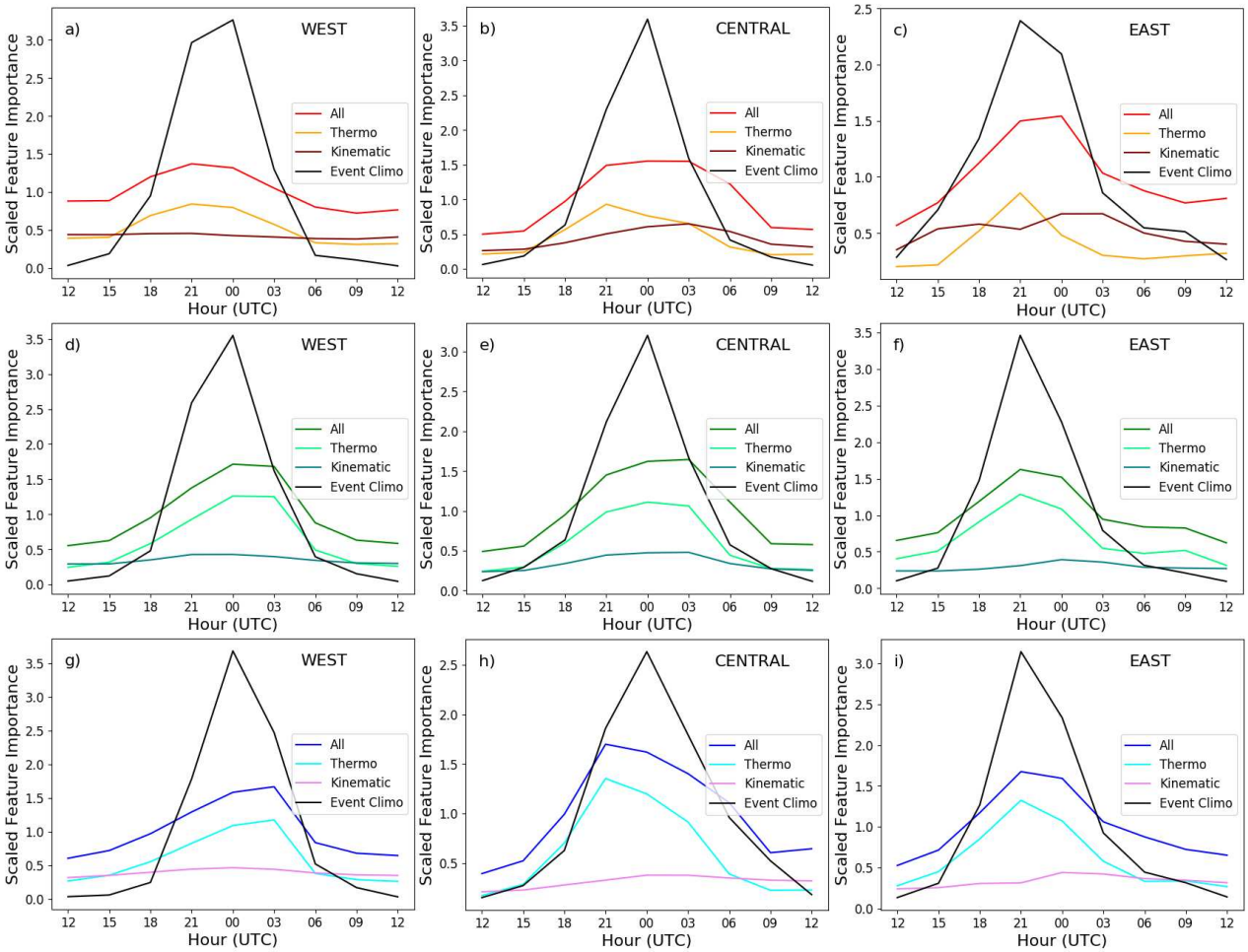


FIG. 6.4. Normalized FIs aggregated as a function of forecast hour for the Day 1 models. The top, middle, and bottom rows depict FIs for the tornado, hail, and wind models, respectively, while the left, center, and right columns respectively depict FIs for the WEST, CENTRAL, and EAST regions. Severe phenomenon diurnal climatologies are depicted for each region in black. These and the total FIs, colored as indicated in the panel legend, are normalized so that the curve integrates to unity. FI time series broken down by thermodynamic and kinematic variables are also included, with lines as colored in the panel legend and using the variable partitioning depicted in Table 6.1.

RF FI time series for Day 2 and 3 models (Fig. 6.5) again share similarities with their Day 1 counterparts. Like with the Day 1 models, importance peaks come earliest in the East (Fig. 6.5c) and latest in the West (Fig. 6.5a), ranging from 2100–0300 UTC. Interestingly, there is a shift in peak importance between Day 2 and 3 models towards earlier times, especially pronounced in the West and Central regions (Fig. 6.5a,b). This may simply be attributable to the degradation in temporal resolution between the two models, but it is possible that there is some lead time dependence on the diurnal climatology and biases in the GEFS/R. As was seen for kinematic variables overall in Day 1 (Fig. 6.4), the predictive

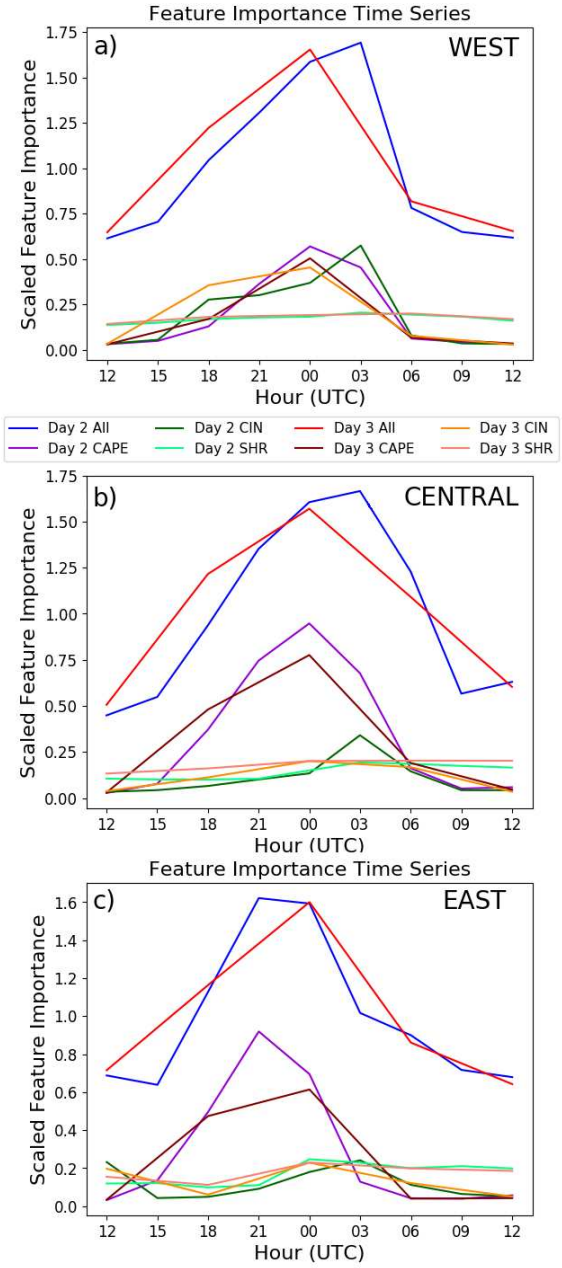


FIG. 6.5. Similar to Figure 6.4, except for the Day 2 and 3 models, which are combined onto single panels for the (a) WEST, (b) CENTRAL, and (c) EAST regions. FI time series of CAPE, CIN, shear, and all variables combined are shown for each forecast region, colored as indicated in the panel legend.

utility of simulated shear is nearly constant at all times throughout the forecast period for both forecast lead times (Fig. 6.5). CAPE and CIN both have more pronounced diurnal signatures, but they are different from one another (e.g. Fig. 6.5a). CAPE FIs peak in association with the maximum in climatological event time frequency, while CIN has a primary peak after this and, in many circumstances

(e.g. Fig 6.5a,c), a secondary peak before it. The secondary peak is perhaps the more intuitive of the two; the environmental CIN in the pre-event environment determines how much of a cap storms must overcome, and the potential for instability to build or storms to be prevented from initiating entirely. The primary peak may speak to the degree of stabilization associated with cold pool strength, instability release, anvil shading, and other factors as portrayed in the convection-parameterized GEFS/R, and the severe weather potential associated with these factors. However, more investigation into the causes of this peak may prove fruitful.

In space (Fig. 6.6), RF FIs are typically highest near the forecast point and decrease with increasing distance from the point, but there are some notably anomalies. FIs are generally most spatially uniform for tornado prediction and have the sharpest peak in predicting severe hail; this is especially true in the West (cf. Fig. 6.4a,d). In the West, while FI importance maxima are collocated with the forecast point for tornadoes and wind, information to the east of the forecast point is more predictive of conditions at that point than the collocated simulated forecast values for hail and the medium-range forecasts. A variety of factors could be attributable to this observation, including a displacement or initiation bias in the model's placement of storms in the region, or the lopsided event climatology in the region, with most events occurring on the eastern fringes with the primary storm ingredients just to the east over the Great Plains. Especially because this appears prominently in the hail signature but not in other fields, the interface between the simulated fields over the Great Plains and events over the far eastern Intermountain West appears a likely source, with more usefully predicted values over the Great Plains, but more investigation is required to validate that hypothesis. In the Central region, FIs are highest from the forecast point south, with downstream maxima for every predictand except severe winds, which has an identified maximum in predictive utility just upstream of the forecast point (Fig. 6.6h). The southern displacement in importance appears to become more pronounced with increasing forecast lead time, and is especially evident at Day 3 (Fig. 6.6n). FI maxima also become less pronounced with increasing forecast lead time (Fig. 6.6j–o), consistent with past studies such as Chapter 4. In the East, importances for all severe weather models maximize near the forecast point and extend to the south and west.

The so-called ring plots of Figures 6.7–6.9 provide a more complete representation of the models' diagnoses and how the summary statistics of Figures 6.2, 6.4, and especially 6.6 were obtained. In the West (Fig. 6.7), the most predictive fields, CAPE and CIN (Fig. 6.2) are seen clearly for all three predictands. In general the importance maxima for these fields occur near the forecast point, though CAPE

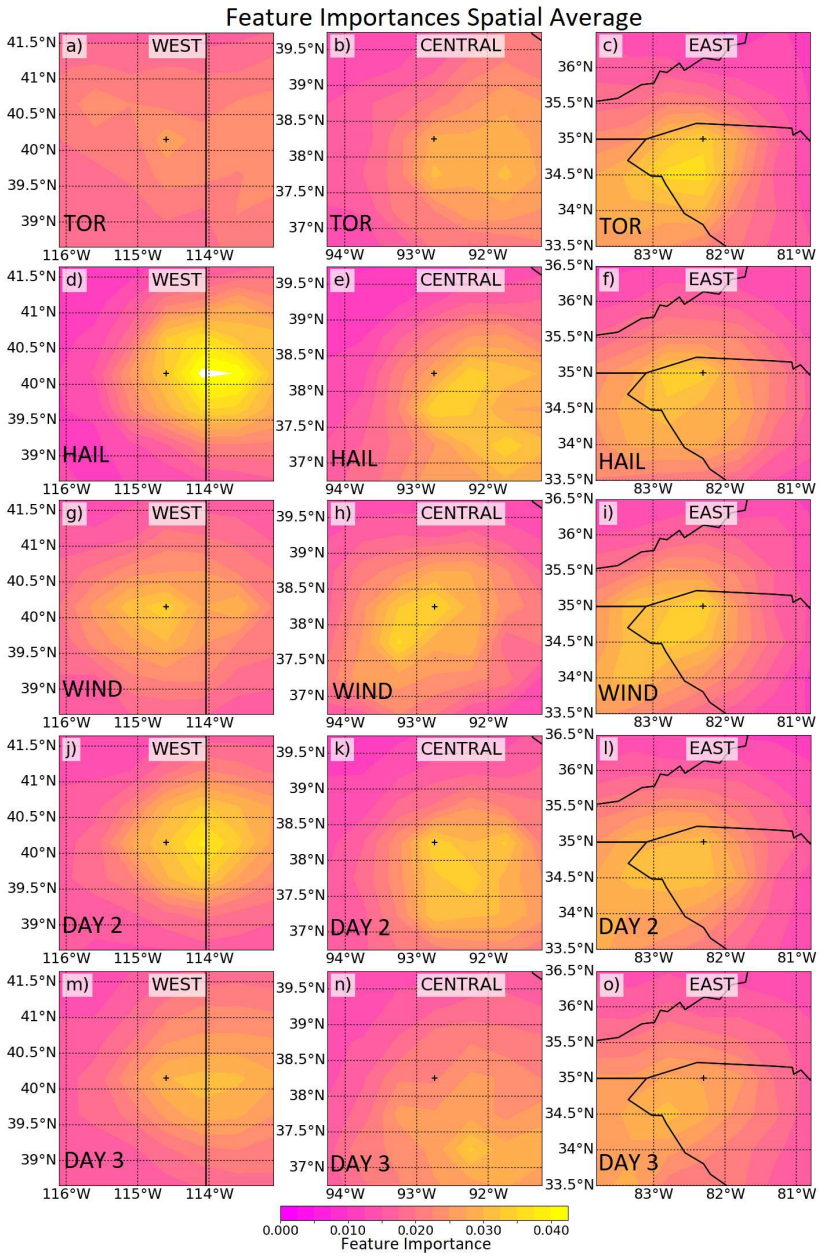


FIG. 6.6. FIs summed according to the corresponding predictor's position in point-relative space for the WEST, CENTRAL, and EAST regions respectively in the left, center, and right columns. Tornado model FIs are depicted in the top row, followed by hail, wind, Day 2, and finally the Day 3 model on the bottom row. Yellows indicate high importance of information at the point, while magentas indicate lesser importance. The forecast point is shown with a black cross; latitude and longitude are presented using the region centroid, and are shown merely to provide improved sense of spatial scale.

FI maxima are displaced farther north relative to the forecast point in predicting tornadoes compared with hail and wind (Fig. 6.7a). For CIN (Fig. 6.7d), importances maximize downstream of the forecast

point, particularly for wind. DSHR is predictive for both hail and wind (Fig. 6.2a, 6.7m), but is maximized on the upstream side of the forecast point for wind and downstream side for hail. A different moisture variable is found to be most predictive of for each severe weather predictand: APCP, Q2M, and PWAT for tornadoes, hail, and wind, respectively (Fig. 6.7h,e,g). In all cases, the spatial maximum in importance is found displaced to the north of the forecast point, likely associated with biases in the GEFS/R's positioning of precipitation systems (e.g. as noted in Chapter 6.4), also seen in other models with parameterized convection (e.g. Clark et al. 2010).

The RF FI maxima and spatial placement thereof displays some similarities and some differences between the West (Fig. 6.7) and Central (Fig. 6.8) regions. In the Central region, CAPE FI (Fig. 6.8a) are still of course paramount for all predictands (per Fig. 6.2), but unlike in the West region, the maxima are found to the south of the forecast point. This southern displacement is even more pronounced in CIN (Fig. 6.8d), particularly for forecasting severe hail. In the moisture variables, there is a shift from the West to Central region, with APCP becoming the preferred moisture variable for each predictand. Interestingly, APCP FI importance is consistently maximized late in the period to the northeast of the forecast point, perhaps noting with its late and eastward-displaced elevated FIs that many tornadoes occur during the afternoon hours with discrete supercell activity and during the upscale growth phase leading up to vigorous evening mesoscale convective systems which are common during the warm-season in this region (e.g. Nielsen et al. 2015). The northern displacement is again consistent with the documented displacement bias in the positioning of convective systems in convection-parameterized models such as the GEFS/R (e.g. Wang et al. 2009). DSHR's FI maxima (Fig. 6.8m) are again centered near the forecast point, although MSHR (Fig. 6.8j) and SRH (Fig. 6.8l), which are particularly predictive for tornadoes, have maximum predictive utility to the southeast of the forecast point. Finally, MSLP is also found to a useful severe weather predictor (Fig. 6.8i), and one observes its importances track from west to east across the forecast point domain throughout the forecast period.

The East region FIs (Fig. 6.9) display very similar spatial patterns in CAPE (Fig. 6.9a) and CIN (Fig. 6.9d) as seen in the Central region. In both cases, it appears that these thermodynamic indicators forecasted in the source region of moisture and instability are more predictive than at the point itself, particularly for CIN. A similar pattern is also seen in APCP (Fig. 6.9h), including the northward displacement. However, the late maximum in the northeast corner is entirely removed, as nocturnal mesoscale convective systems are not climatologically frequent over much of this region, and the synoptic conditions associated with tornadoes are often different between the regions (e.g. Smith et al.

2012). Shear is again most important nearly collocated with the forecast point (Fig. 6.9j,m) with MSHR (Fig. 6.9j)—especially late in the period—being more predictive for tornadoes and wind, while DSHR (Fig. 6.9m) is the dominant shear variable for predicting hail. In predicting tornadoes, meridional

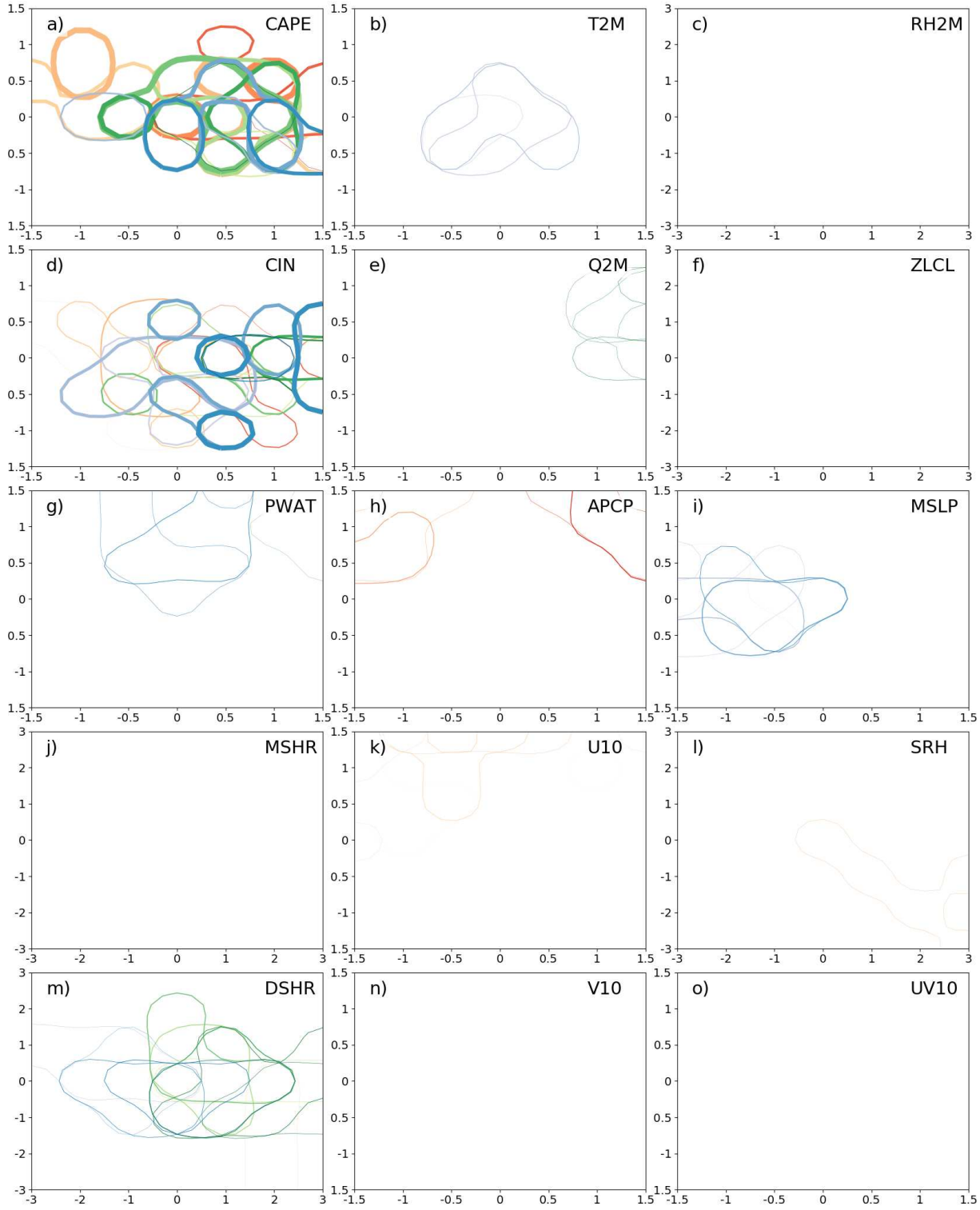


FIG. 6.7. Feature importances by space and atmospheric field for the Day 1 tornado, hail, and wind models in the WEST region. Rings enclose regions where the FI for the variable and time exceeds 1.5 standard deviations above the spatial mean FI for that variable and time. Ring colors vary according to the predictand of the model, with oranges and reds corresponding to FIs associated with predicting tornadoes, greens to predicting hail, and blues to predicting wind. Within these, colors darken and transition from orange (tornado), green-yellow (hail), and purple-blue (wind) to solid red, green, and blue with time throughout the forecast period, from the front-end 1200 UTC (forecast hour 12) to the back-end 1200 UTC (forecast hour 36). Line thickness is determined by the FI threshold associated with the ring, with thicker lines indicating higher FI and rings associated with below average thresholds (based on the +1.5 standard deviation exceedance given the predictand, predictor field, and time) are excluded entirely. Panels (a)–(o) correspond respectively to FIs for the CAPE, T2M, RH2M, CIN, Q2M, ZLCL, PWAT, APCP, MSLP, MSHR, U10, SRH, DSHR, V10, and UV10 fields.

winds (Fig. 6.9n) to the south of the forecast point and MSLP (Fig. 6.9i) upstream of the forecast point are found to be good discriminators of tornado events and non-tornado events, speaking to both the degree of advection of convective ingredients from the south and the level of synoptic-scale forcing for ascent advecting into the region. SRH (Fig. 6.9l) is found to be predictive of tornadoes throughout the period, with FI maxima generally tracking west to east to the immediate south of the forecast point during the period. One other major difference between the East region and other regions is the importance of nighttime T2M (Fig. 6.9b) in predicting hail in the East; the exact reasoning for this identification is not obvious.

In summary, the RFs trained in this study appear to be making statistical deductions that are in strong agreement with our physical understanding of severe weather processes, and identify values to inspect—such as CAPE and shear near the forecast point and APCP to its north, and inspecting DSHR for hail but MSHR for tornadoes—that agree with conventional operational severe weather forecast practices (e.g. Johns and Doswell 1992). However, the RF provides an automated, objective, and quantitative synthesis of these many important factors that contribute to a skillful severe weather forecast, in addition to identifying some factors, such as the southward CIN FI maxima displacement, that may be less well-documented but still contribute to a skillful forecast. The following section investigates the predictive performance of these models.

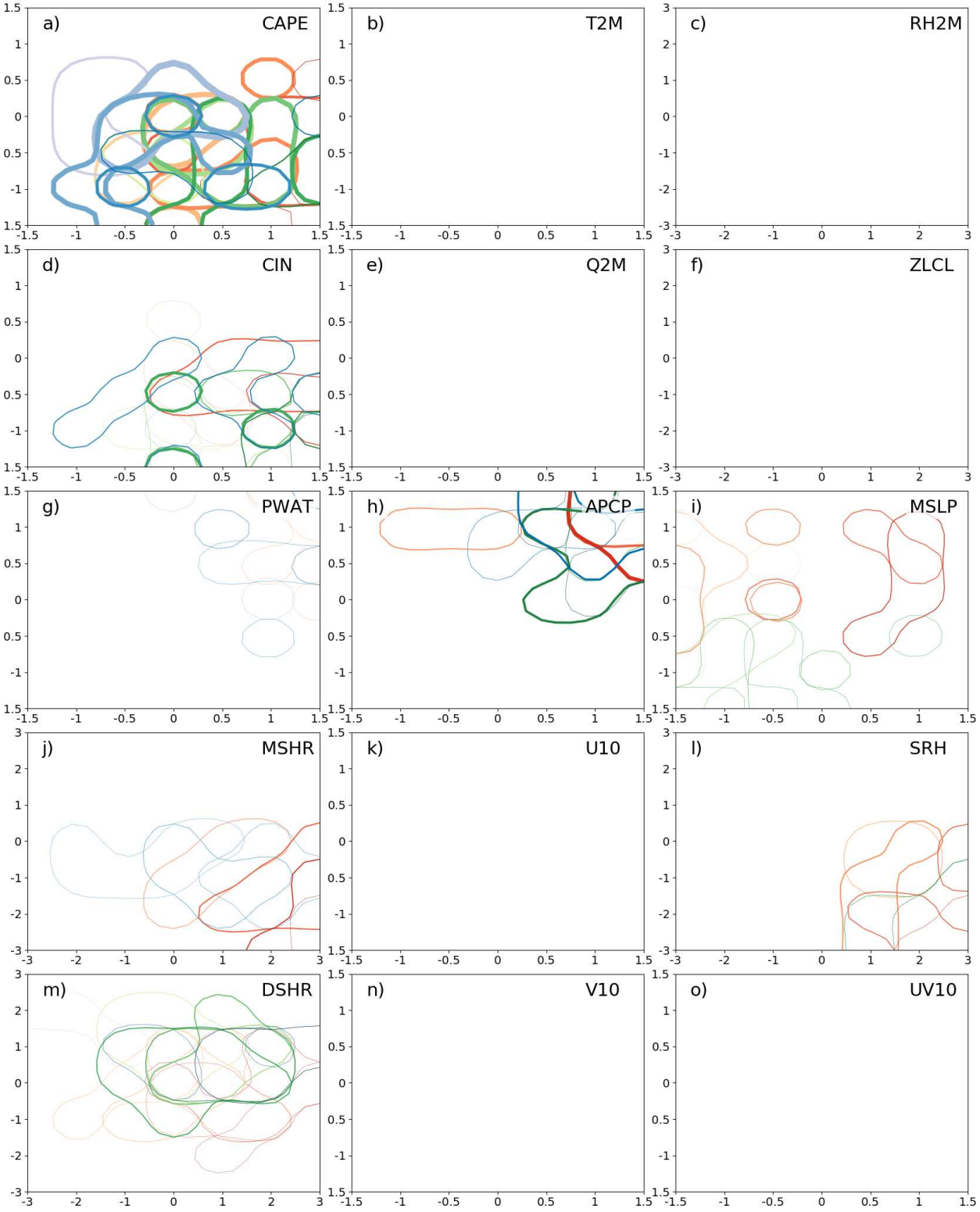


FIG. 6.8. Same as Figure 6.7, but for the CENTRAL region.

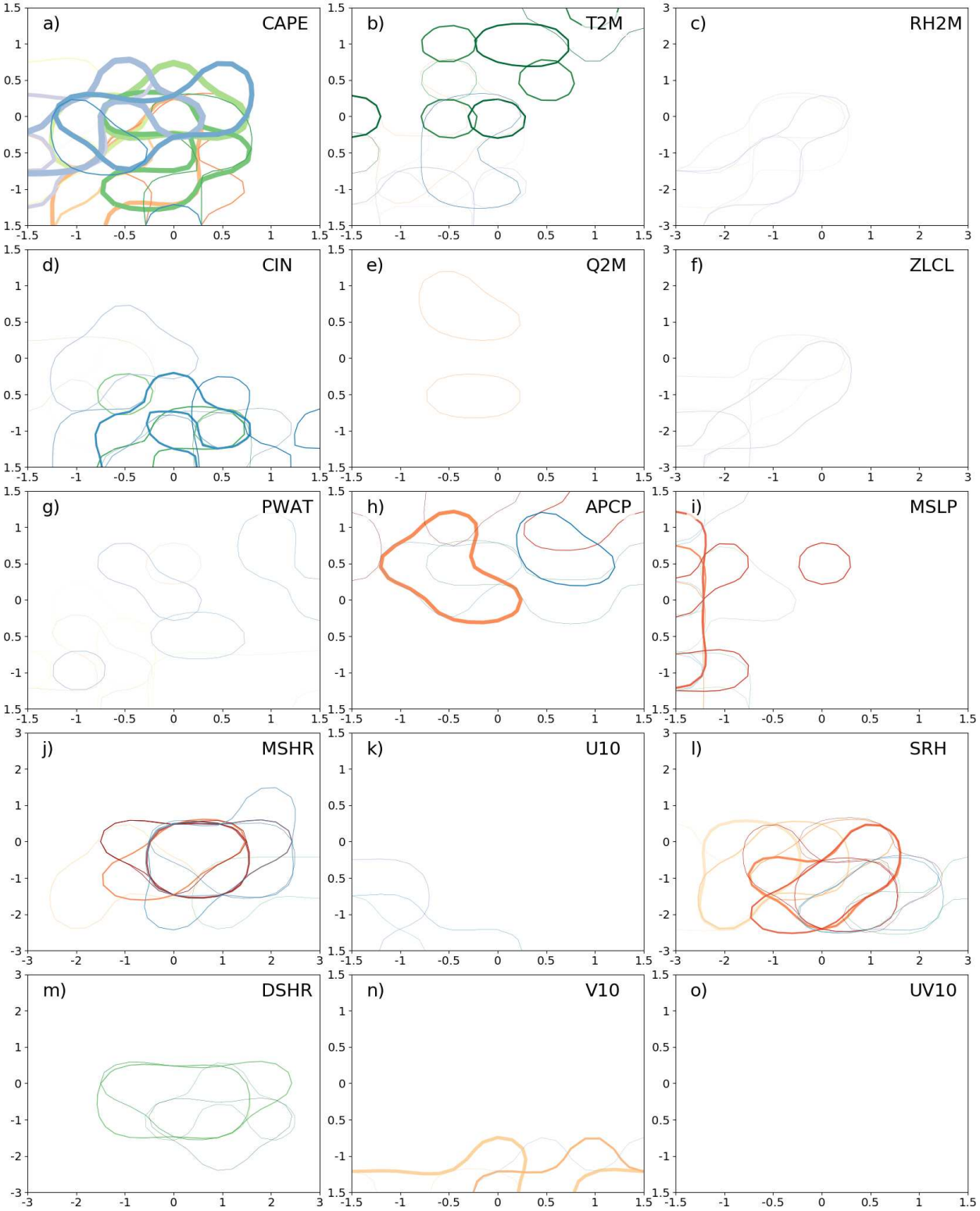


FIG. 6.9. Same as Figure 6.7, but for the EAST region.

6.4 Results: Model Performance

The RFs show ability to skillfully predict all severe weather predictands (Fig. 6.10), though there are some differences in the details. Prediction of tornadoes (Fig. 6.10a) produced the most mixed verification results, with statistically significant positive skill over the Central Great Plains, Mississippi Valley, Ohio River Valley, and parts of the Mid-Atlantic region and Floridian Peninsula. However, BSSs are lower and in many cases less skillful than climatology—albeit not statistically significantly so—over the West, Northeast, Upper Midwest, far northern and southern Plains, and the Carolinas. These same general findings extend for significant tornadoes (Fig. 6.10b) but with lower skill overall, with CONUS-wide skill decreasing from 0.029 for tornadoes to 0.013 for significant torrnadoes. The large area of extremely negative skill over the West is simply reflective of the fact that no significant tornadoes were observed over this region during the verification period, and the model had above climatological probabilities for some events. Due to the small or even non-existent sample, the negative skill observed here is not statistically significant. Hail (Fig. 6.10c), wind (Fig. 6.10e), and the Day 2 and 3 (Fig. 6.10g,h) models all exhibit very similar spatial patterns of forecast skill, with near uniform and statistically significant positive skill over much of CONUS east of the Rocky Mountains. Somewhat degraded skill is seen over Southern Texas, Florida, and pockets of the Upper Midwest; these spatial variations are particularly pronounced in the hail verification (Fig. 6.10c). In the West, fewer of the results are found to be statistically significant due to the reduced event frequency. Nevertheless, positive skill is still noted for these predictands over much of the West, with the exceptions of a pocket of southwestern Colorado and surroundings and the Pacific Coast. As with SPC convective outlooks (see Chapter 5), Day 1 forecast skill is highest for severe winds at 0.105, with hail in the middle at 0.079. Skill unsurprisingly decreases with increasing forecast lead time, and CONUS-wide BSSs of 0.108 and 0.089 are observed for Day 2 and Day 3 RF outlooks, respectively (Fig. 6.10g,h). Like with tornadoes, the spatial patterns are similar between hail and wind and their significant severe counterparts (Fig. 6.10d,f), except with lower skill magnitudes with CONUS-wide numbers of 0.023 and 0.022 for significant hail and wind. The highest (and statistically significant) skill is seen over the Central Plains for these variables; positive but insignificant skill is observed in the East, and skill near climatology observed over much of the West.

Relative to SPC (Fig. 6.11), the RF outlooks verify quite competitively. On Day 1, where human forecasters have access to more skillful convection-allowing guidance and more updated observations and simulations, SPC outlooks are generally more skillful than the RF, with aggregate skill score differences of 0.007 for hail (Fig. 11c) increasing to 0.013 for tornadoes (Fig. 6.11a) and 0.024 for severe

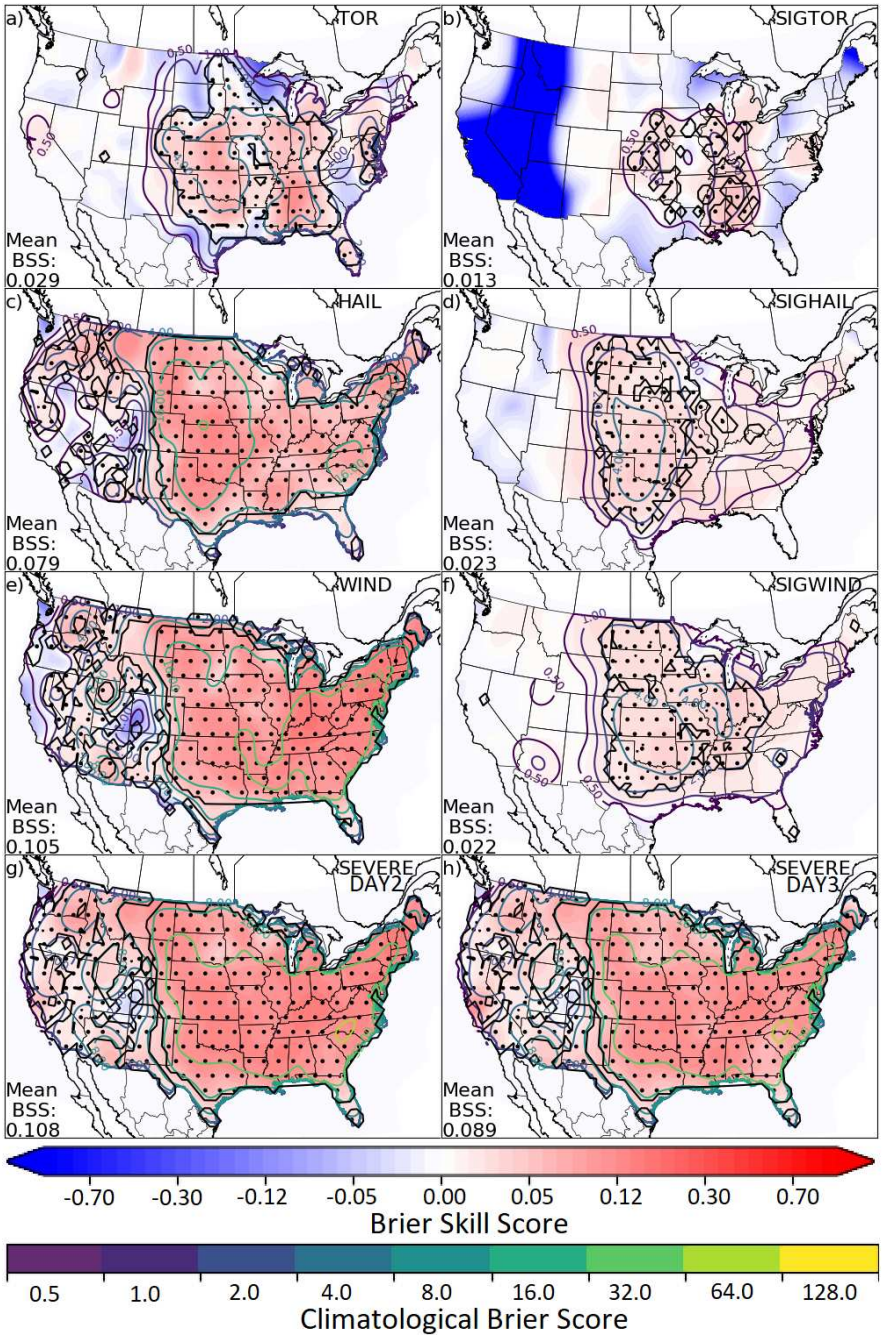


FIG. 6.10. Brier skill scores (filled contours) in space evaluated over the 12 April 2012–31 December 2016 verification period for each of the ML models trained in this study. Panels (a)–(h) correspond respectively to the performance of the tornado, significant tornado, hail, significant hail, severe wind, significant severe wind, Day 2, and Day 3 outlooks. Unfilled contours depict the Brier score of climatology at the point over the verification period; higher values indicate more common events. Stippling indicates areas where the sign of the skill score is statistically significant at 95% obtained from bootstrapping as described in the text.

wind forecasts (Fig. 6.11e). However, the CONUS-wide summary gives an incomplete picture, as there are significant regional variations in skill differences. Unlike the RF outlooks, which exhibited fairly uniform skill in hail and wind across the eastern two-thirds of CONUS (Fig. 6.10c,e), SPC interpolated convective outlooks exhibited a strong latitudinal gradient in BSS, with higher skill to the north (Chapter 5). This is reflected in the skill comparison, with SPC outlooks substantially outperforming the RF outlooks over far northern CONUS in predicting severe hail and wind (Fig. 6.11c,e). However, over the southern two-thirds of CONUS, the RF outlooks outperform the SPC outlooks in these fields. There is much more spatial inhomogeneity in the tornado outlooks (Fig. 6.11a). The magnitudes of the skill differences at a point are usually much smaller than in the hail and wind outlooks, but SPC outlooks still outperform the RF forecasts the most in the northern tier of states. The mixed spatial skill comparisons for tornadoes extend to verification of significant tornadoes (Fig. 6.11b) as well, but the comparison is much different for significant hail (Fig. 6.11d) and wind (Fig. 6.11f) events. Here, RF outlooks are actually found to exhibit higher probabilistic skill overall than the SPC outlooks, with skill differences of 0.012 and 0.020 respectively for the significant severe hail and wind outlooks. The gains are largest over the Central region.

For Day 2 and 3 outlooks (Fig. 6.11g,h), the RF outlooks exhibit substantially higher probabilistic skill than the analogous SPC convective outlooks, with aggregate CONUS-wide skill differences of 0.043 and 0.045 respectively for the Day 2 and 3 outlooks. RF outlooks demonstrate higher skill over almost all parts of CONUS, the primary exceptions being the Pacific Coast and western Colorado where the RFs had lower absolute skill (e.g. Fig. 6.10g), and over Louisiana and Arkansas. The biggest skill differences over SPC are in the East region domain, particularly the Mid-Atlantic and southern New England. The general finding that the RF outlook skill becomes increasingly skillful relative to SPC outlooks with increasing forecast lead time is consistent with there being less information beyond global, convection-parameterized ensemble guidance on which to base a skillful forecast with increasing lead time, with the biggest jump between Days 1 and 2.

Except for hail (Fig. 6.12b), which exhibits a springtime maximum in skill, all RF outlooks exhibit a climatology-relative peak in skill during the cold-season (Fig. 6.12a,c,d). In fact, hail exhibits essentially an inverted seasonal cycle in forecast skill compared with the other variables, since hail outlooks verify worst in the winter and other variables verify worst in March. Tornadoes and wind also exhibit a skill minimum in late summer–early autumn, consistent with SPC outlooks verified in Chapter 5.

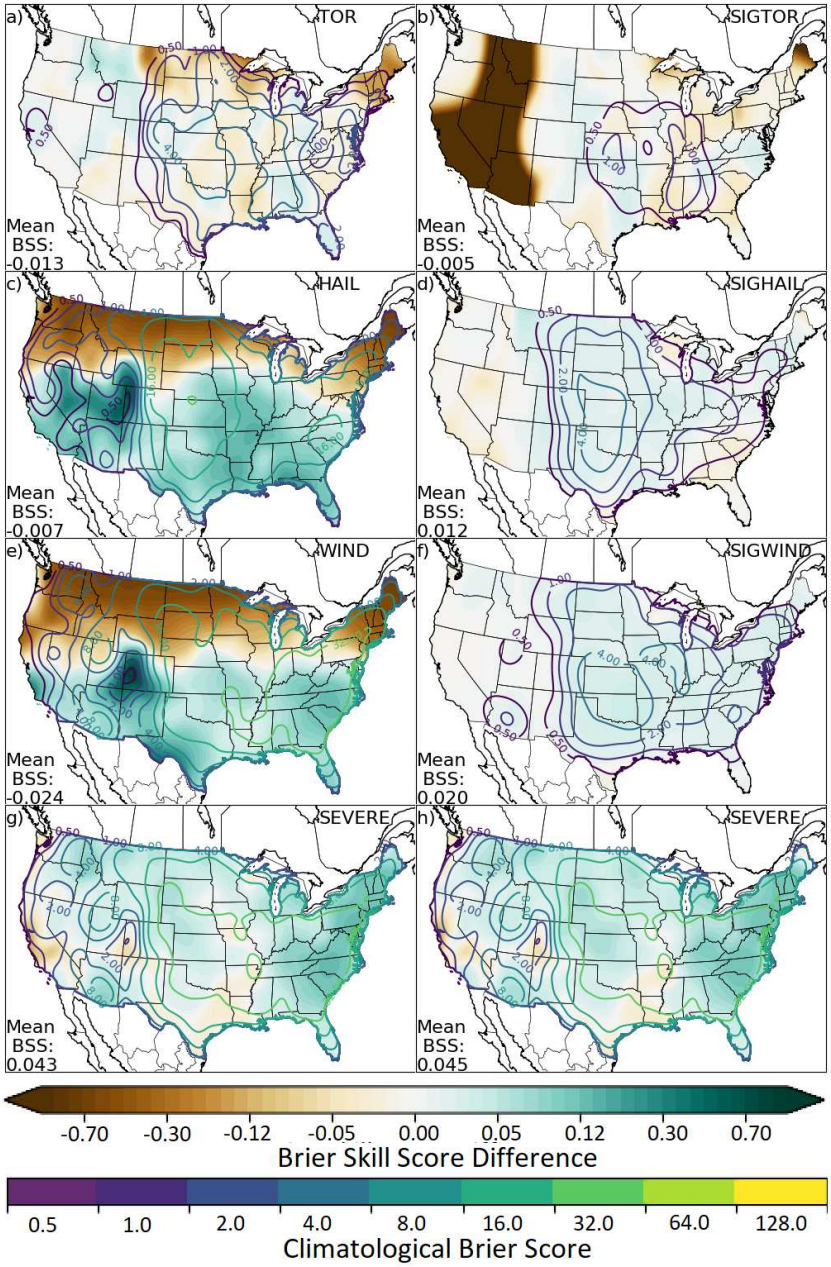


FIG. 6.11. Same as Figure 6.10, except depicts the difference in BSS between ML outlooks and the analogous outlooks issued by SPC. Greens indicate ML forecasts outperform SPC; browns suggest the opposite. Due to data availability, a slightly shorter 13 September 2012–31 December 2016 period is used for the Day 2 and 3 outlook verification comparison.

For all severe weather predictands, the severe and significant severe events have nearly identical seasonal cycles in forecast skill (Fig. 6.12). Comparing against SPC, while there does not appear to be a clear seasonal or monthly signal in the skill difference for tornado outlooks (Fig. 6.12a), the primary

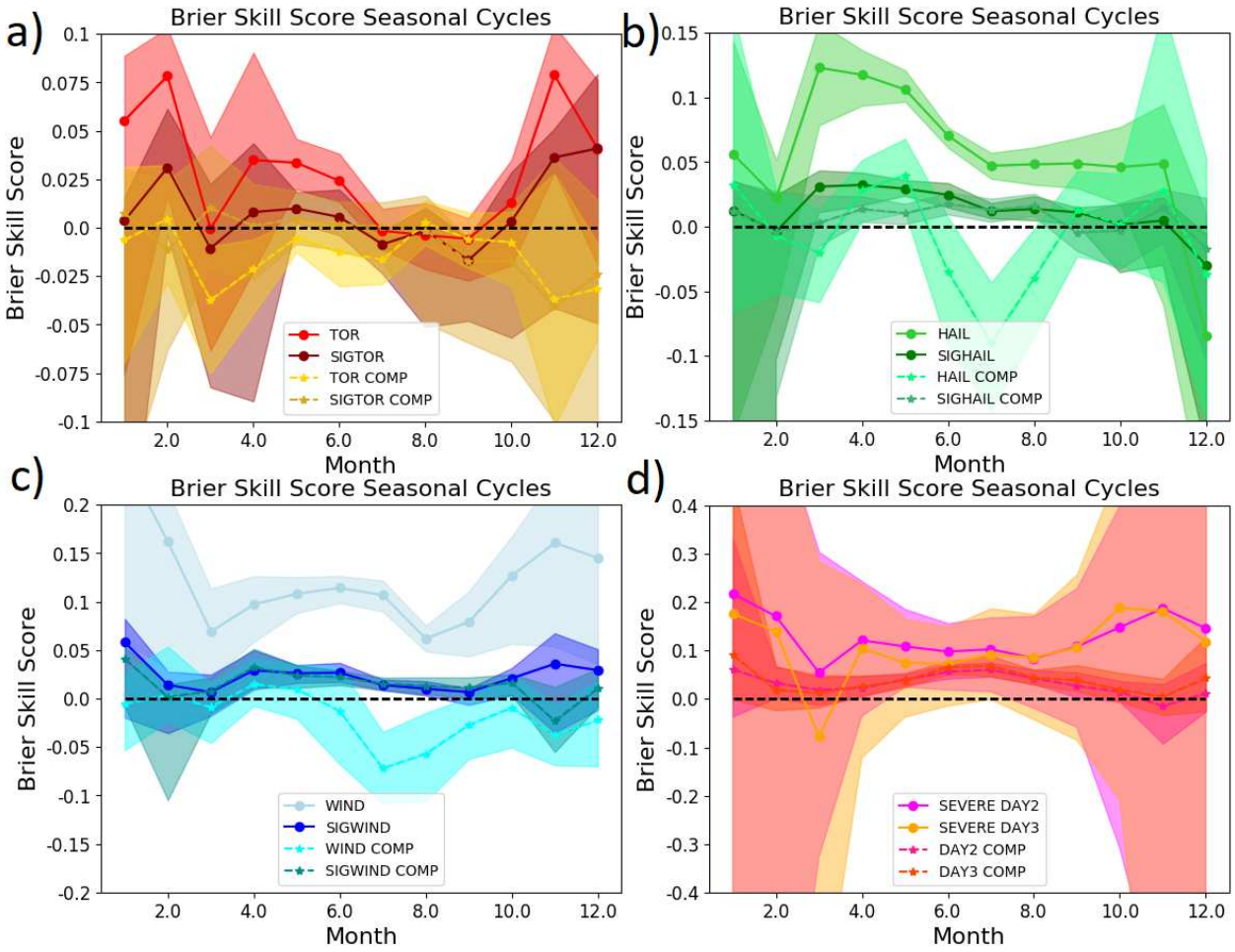


FIG. 6.12. BSSs by month and comparison between ML and SPC outlooks for (a) tornado and significant tornado, (b) hail and significant hail, (c) wind and significant wind, and (d) Day 2 and 3 outlooks. Lines are colored as indicated in the panel legend; shading about the line indicates 95% confidence bounds obtained by bootstrapping. Differences are ML-SPC, positive numbers indicating ML outperforms SPC. Note that the y-axis varies between panels.

advantage for SPC outlooks over the RF counterparts in hail and wind appears to come in the month of July, where SPC outlooks performed very well (Chapter 5) and substantially outperform the RF outlooks. In contrast, in the Day 2 and Day 3 comparison, RF outlooks outperform SPC by the most during the summer, maximizing in July. These differences are all consistent with the SPC being able to effectively harness the advantages of convection-allowing guidance for their Day 1 convective outlooks over the warm-season, where the responsible physical processes are predominantly smaller-scale and more weakly forced than cold-season events. At Day 2 and 3, where convection-allowing guidance is

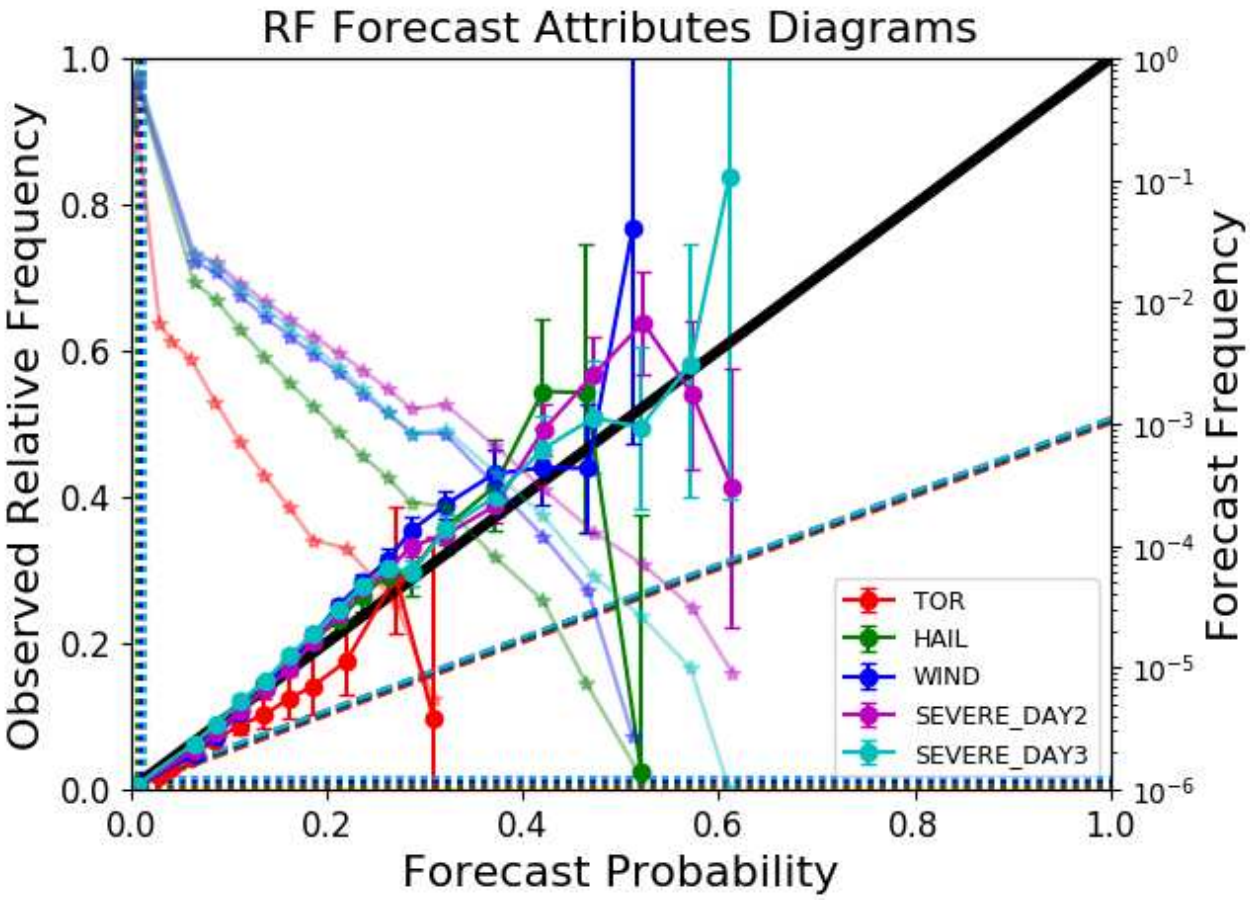


FIG. 6.13. Attributes diagrams for ML-based outlooks. Colored opaque lines with circular points indicate observed relative frequency as a function of forecast probability; the solid black line is the one-to-one line, indicating perfect reliability. Colors correspond to different severe predictands and lead times as indicated in the panel legend. Semi-transparent lines indicate the total proportion of forecasts falling in each forecast probability bin, using the logarithmic scale on the right hand side of the figure. Probability bins are delineated by 2.5%, 3.5%, 5%, 7.5%, 10%, 12.5%, 15%, 17.5%, 20%, 25%, and 30% thresholds for Day 1 tornado forecasts, and by 5.5%, 7.5%, 10%, 12.5%, 15%, 17.5%, 20%, 22.5%, 25%, 27.5%, 30%, 35%, 40%, 45%, 50%, 55%, and 60% for all other forecast sets. Horizontal and vertical dotted lines denote the "no resolution" lines and correspond to the bulk climatological frequency of the given predictand. The tilted dashed lines depict the "no skill" line following the decomposition of the Brier score. Error bars correspond to 95% reliability confidence intervals using the method of Agresti and Coull (1998), where non-overlapping neighborhoods are assumed to be independent.

largely unavailable, SPC outlooks suffer from biased guidance that cannot come close to resolving the responsible physical processes. These biases are largest in the convectively-active warm-season; the RF outlooks, using years of historical data, are able to robustly identify and correct for many of these biases, leading to the largest improvements in skill when the model biases are largest and the least skillful external guidance is available to the human forecaster.

Reliability diagrams for the RF outlooks (Fig. 6.13) demonstrate quite calibrated forecasts along the spectrum of the probability distribution. A slight underconfidence bias is observed for most predictands, but otherwise calibration remains quite good until the highest probability bins, where sample size is very small. Maximum forecast probabilities get as high as approximately 30% for tornadoes, into the lower 50% range for hail and wind, and into the lower 60s for any severe at Days 2 and 3. The main exception to calibration is the tornado forecasts, which are characterized by a slight overforecast bias. This may be attributable to large differences in the event frequency between the training sample, which featured many highly active tornadic years, and the test period, which as discussed in Chapter 5, was relatively quiescent.

The weighted blend of SPC and RF outlooks described in Chapter 6.2 (Fig. 6.14) unsurprisingly demonstrates forecast skill spatial characteristics of both the interpolated SPC Chapter 5 and RF (Fig. 6.10) outlooks. Most prominently, the high skill in the northern states in the SPC outlooks is reintroduced to the blend in the hail and wind outlooks (cf. Fig. 6.10c,6.14c; 6.10e,6.14e). For predictands in which the skill difference is large between the two outlook sources, such as for significant wind (Fig. 6.14f) and the medium-range outlooks (Fig. 6.14g,h), the blended outlooks verify very similarly to the more skillful component, in part simply because the weights direct the blend heavily towards that component. Across the board, the SPC RF blend verifies as or more skillfully than the SPC outlooks alone—both in space (Fig. 6.15) and when aggregated across CONUS (Fig. 6.16)—a testament to the utility of the RF guidance in improving operational severe weather forecasts. Even at Day 1, where SPC outlooks outperform the raw RF guidance (Fig. 6.16), the blended forecasts outperform both the raw SPC and raw RF outlooks. In the case of hail and wind, the margin of improvement is considerable, with BSS improvements of 0.061 and 0.053 respectively (Fig. 6.15c,e). At Day 2 and 3, while the blend is not able to improve skill over the RF outlooks (Fig. 6.16), that difference is already considerable when compared with the SPC outlooks at 0.044 and 0.048 (Fig. 6.15g,h). Consequently, the blended forecast exhibits much improved skill compared with the raw SPC outlooks for all eight forecast predictands evaluated (Fig. 6.16). Even more encouragingly, the skill improvements are seen across all regions of CONUS (Fig. 6.15) with fairly uniform distribution. For hail, wind, and the medium-range outlooks, the skill differences are statistically significant over all except for pockets of western CONUS where the climatological event frequencies are insufficient to produce a robust sample. Hail outlooks ae most improved over the Mississippi Valley region into the Midwest, while wind outlooks are most improved

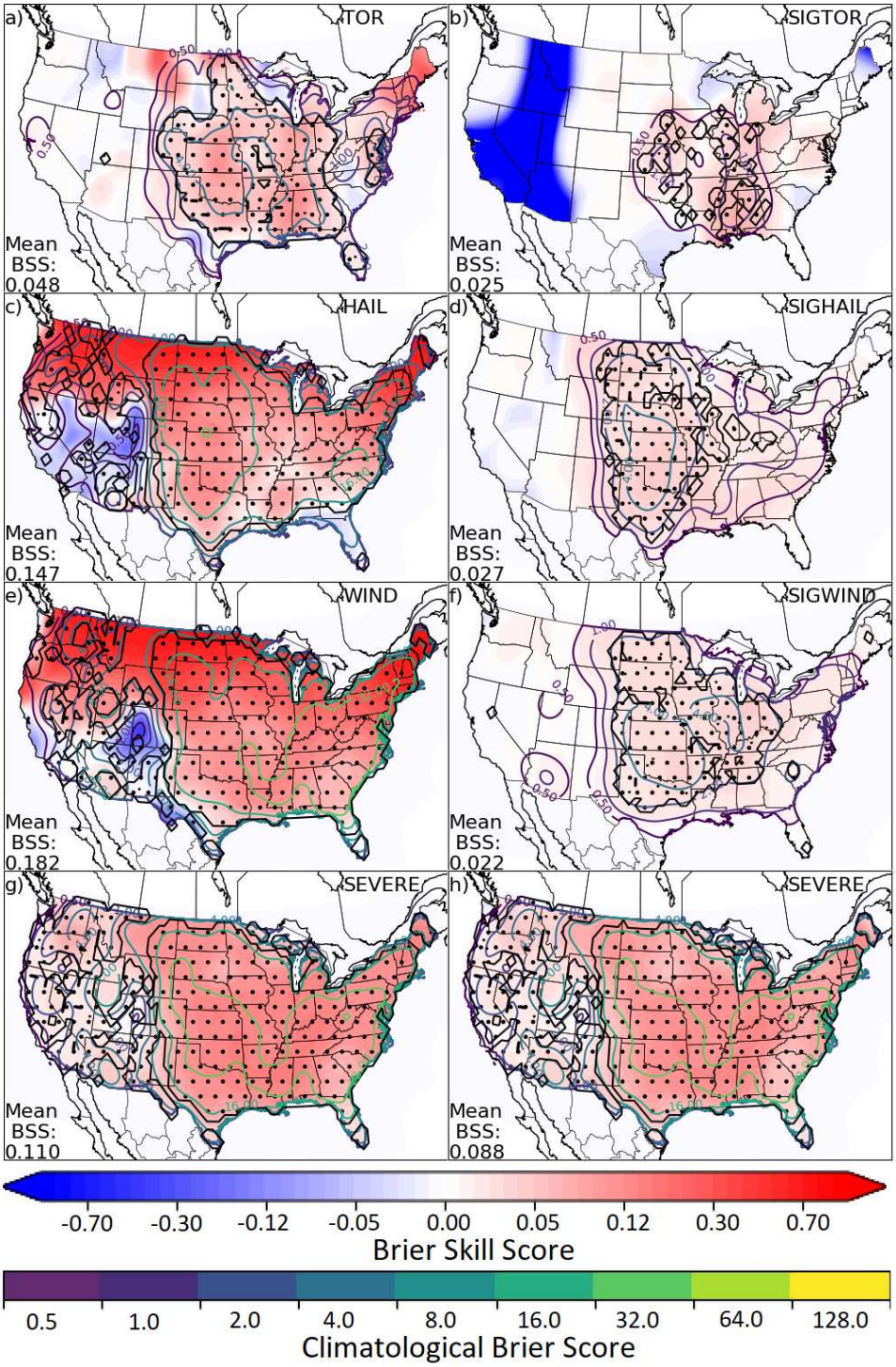


FIG. 6.14. Same as Figure 6.10, except for the weighted blend of SPC and ML outlooks.

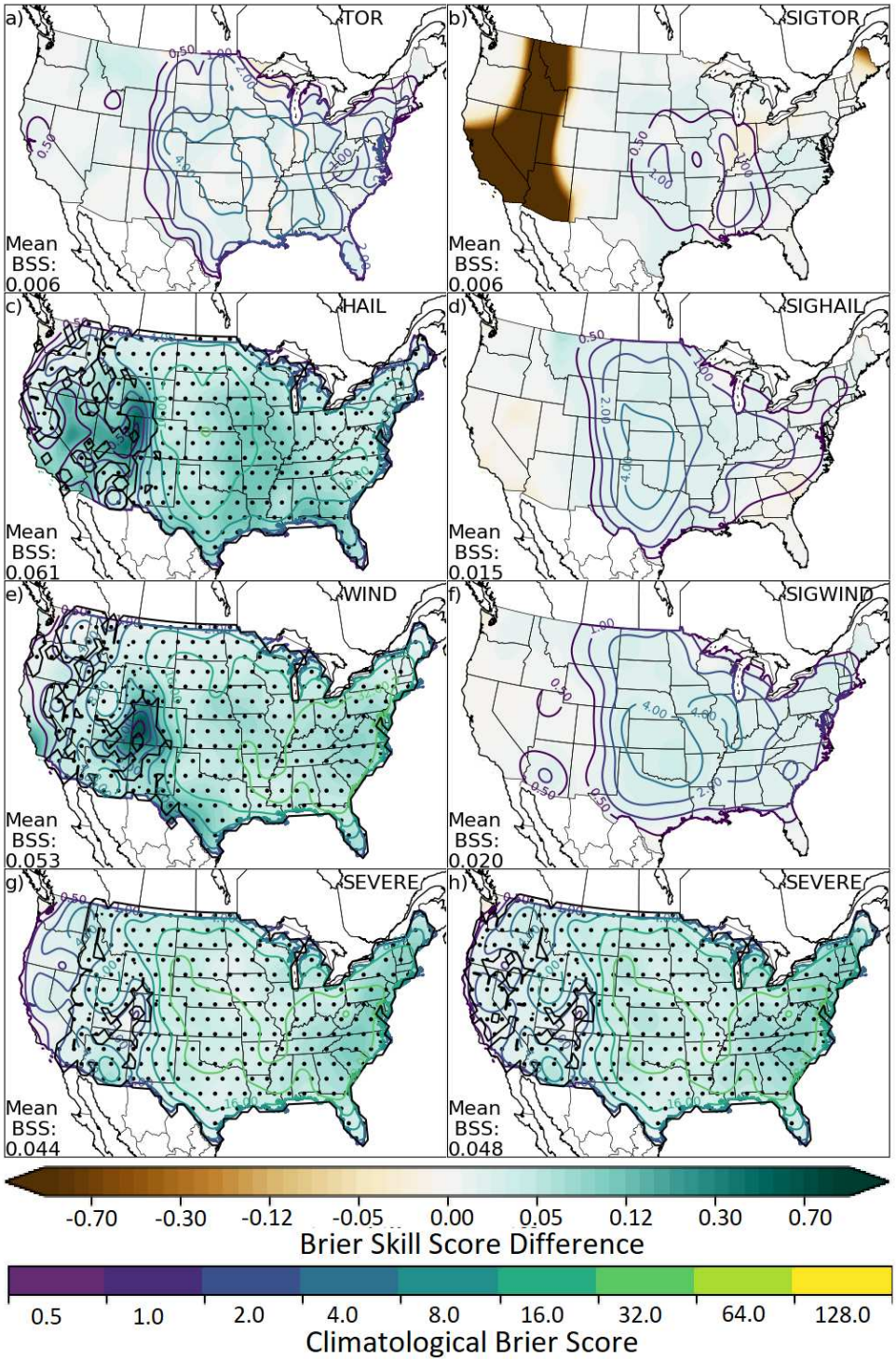


FIG. 6.15. Same as Figure 6.11, except for the weighted blend of SPC and ML outlooks.

over the southern Plains, and the medium-range outlooks most improved over the East Coast urban corridor.

One additional instructive skill decomposition inspects forecast verification in the CAPE vs. shear parameter space. The raw RF hail (Fig. 6.17d) and wind (Fig. 6.17g) forecasts exhibit high skill throughout much of the parameter space. Wind forecasts are skillful throughout essentially the entire space, with a skill minimum in the low CAPE, low shear corner of the parameter space. Hail (Fig. 6.17g) exhibits a local BSS minimum in this region as well, but has primary skill minima in the high CAPE, low shear and especially the low CAPE, high shear corners of the parameter space. Tornado forecast (Fig. 6.17a) verification results are more mixed. Like hail, forecast skill suffers in scenarios with ample supply of CAPE or shear, but little of the other. Skill is significantly positive when sufficient amounts of both ingredients are in place, but outlooks are not always skillful relative to climatology with less pronounced convective ingredients, as evidenced by the interior pockets of blue in Figure 6.17a. The addition of the weighted average with SPC outlooks (Fig. 6.17b,e,h) improve outlook skill across the parameter space while leaving the character of the skill distribution much the same. Skill improvement is especially evident in low CAPE scenarios with low to moderate wind shear (e.g. Fig. 6.17e); skill improvement is minimal in the high CAPE, low shear and low CAPE, high shear corners of the parameter space, where as shown in Chapter 5, SPC outlooks also struggle. In comparison to the raw SPC outlooks, the blend of the RF-based ML forecasts with the SPC outlooks yields skill improvements across the parameter space for hail (Fig. 6.17f) and wind (Fig. 6.17i) forecasts, and across much of the domain for tornadoes (Fig. 6.17c). The skill improvements are largest in the low shear end of the parameter space, especially with high CAPE. Moderate to high wind shear is a necessary ingredient for supercell activity, processes which can be much better resolved by convection-allowing models than parameterized guidance like the GEFS/R. Benefit of employing these RF outlooks can likely be maximized on the low shear end of the parameter space because the benefits from the statistical learning are more offset by an inferior representation of the underlying dynamics in the GEFS/R in high wind shear scenarios.

Finally, a brief case study example is provided in order to illustrate the real-time character of the ML model forecasts. Across many cases, the spatial character of the ML-based outlooks are often very similar to those produced by SPC. This is seen for the outlooks valid 1200 UTC May 9 2016–1200 UTC May 10 2016 (Fig. 6.18), a period in the middle of a moderate-severity multi-day outbreak which spread from the Colorado Plains out to Appalachia. This 24-hour period, while not the most intense outbreak of the evaluation period, garnered a considerable number of reports for each severe weather phenomenon in different areas, including significant severe observations for each. Tornadoes (Fig. 6.18a,b) occurred primarily in two groups. One cluster centered about southern and southeastern Oklahoma,

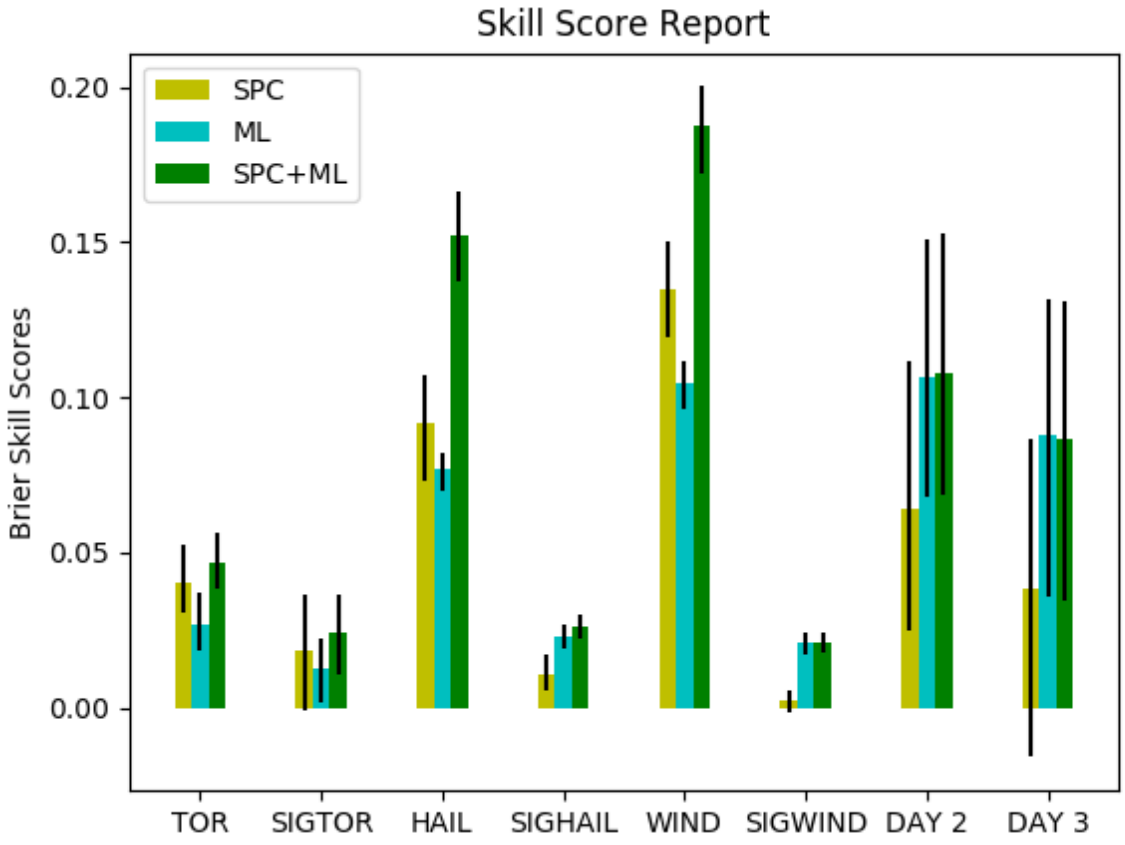


FIG. 6.16. CONUS-total BSS for each of the eight verified predictands for the SPC outlooks (yellow bars), ML forecasts (blue bars), and weighted average of the two (green bars). Error bars indicate 95% BSS confidence bounds obtained via bootstrapping.

with scattered reports up into central Oklahoma and south and east into Arkansas and far northeastern Texas. The second cluster was more broadly spread out from southern Nebraska and northern Kansas east across Iowa and Missouri into western Illinois. Both had at least one significant tornado embedded. Hail observations (Fig. 6.18c,d) were more focused in a north-south oriented region extending from the Oklahoma-Texas border into far northern and northeastern Nebraska, with significant observations seen throughout this region. Wind observations (Fig. 6.18e,f), in contrast, were observed only in two regions: a tightly clustered region in south central Kansas, and a broader region from the Texas/Oklahoma/Arkansas triple point extending northeast across Arkansas into southeastern Missouri. SPC's Day 1 tornado outlook (Fig. 6.18b) highlighted the southern domain reasonably well, with a 10% risk contour, but was generally too far southeast with many tornadoes occurring on the edge of the 2% probability contour, and most of the northern cluster was missed entirely. They identified

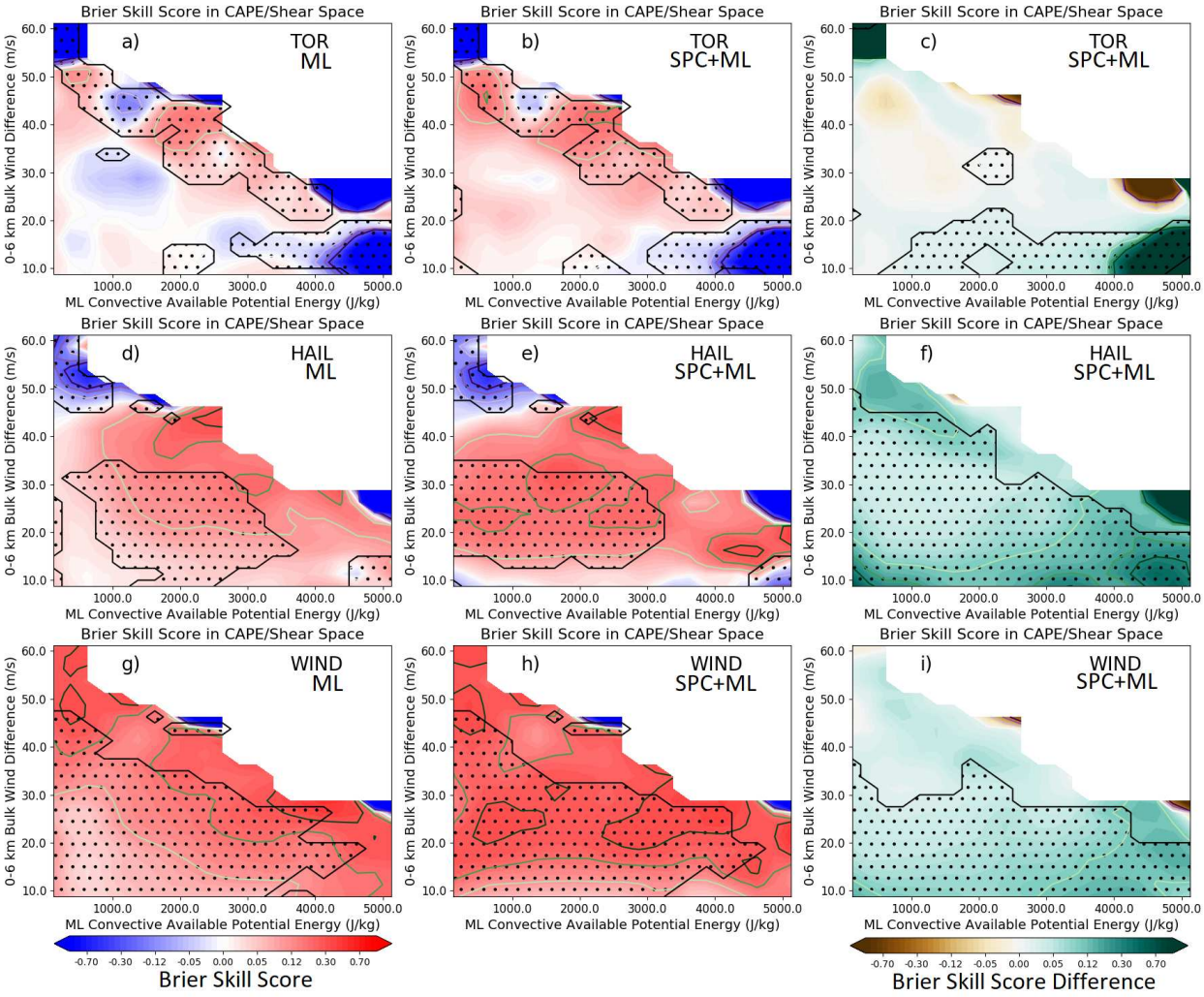


FIG. 6.17. BSS evaluation broken by CAPE versus shear parameter space for tornado, hail, and wind outlooks in panels (a)–(c), (d)–(f), and (g)–(i) as partitioned in Herman et al. (2018) and described in the chapter text. Unfilled contours replicate the filled contours at the -0.3, -0.2, -0.1, 0.1, 0.2, and 0.3 levels and are included for quantitative clarity. The left column depicts verification of the ML forecasts, the center column to the evaluation of the weighted blend of SPC and ML outlooks, and the right column presents the skill score difference between the blend and the raw interpolated SPC outlooks, with greens indicating an improvement over the SPC outlooks and browns representing loss of skill. Stippling indicates regions where the sign of the BSS or BSS difference is statistically significant with α=0.05 based on bootstrap resampling.

hail (Fig. 6.18d) as the primary risk of the day, with a 30% risk contour in addition to a significant hail contour over eastern Oklahoma, western Arkansas, and far northeastern Texas. Their wind outlook had essentially an identical outline to the severe hail one, except topping out with approximately 15% event probabilities and no significant wind contour.

The ML Day 1 outlooks did several desirable changes compared with the SPC outlooks. The tornado outlook (Fig. 6.18a) both indicates higher risk, with a maximum tornado probability over 15%; displaces the maximum to the northwest where more events were observed; and extends the probabilities farther north to at least indicate some appreciable risk in the northern cluster, albeit still lower than in the southern region. The hail (Fig. 6.18c) and wind (Fig. 6.18e) outlooks are more distinct, with higher hail probabilities to the north and west over Oklahoma, Kansas, and Nebraska and lower probabilities to the east; these changes again better collocate the high event probabilities with the observations. Compared with hail, wind probabilities maximize to the southeast over eastern Oklahoma and Arkansas. The models also had better spatial placement in the medium-range, even indicating the two primary risk areas at Day 2 (Fig. 6.18g), and encompassing the western severe weather observations when the operational outlook (Fig. 18h) did not. This was further magnified at Day 3 when only a 15% severe probability was indicated and many severe weather over the Central Plains were not encompassed by the 5% marginal contour in the operational outlook (Fig. 6.18j), while nearly every observation was encompassed by a marginal contour at Day 3 in the ML outlook (Fig. 6.18i) and severe probabilities maximized over 30%. While not all cases demonstrate this degree of success, this case study exemplifies many of the benefits consistently demonstrated by machine learning: relative spatial placement of risks, approximate risk magnitudes, and rarely missing observed events entirely.

6.5 Summary and Conclusions

RFs have been trained to generate probabilistic predictions of severe weather for Days 1–3 across CONUS with analogous predictands to SPC's convective outlooks, with tornado, hail, and wind treated separately at Day 1 and collectively for Days 2–3. Distinct RFs were trained for western, central, and eastern CONUS as partitioned in Figure 1. Inputs to the RFs came from the GEFS/R ensemble median of 12 different atmospheric fields: APCP, CAPE, CIN, PWAT, U10, V10, UV10, T2M, Q2M, MSHR, and DSHR. For the Day 1 models, three additional predictors were used: RH2M, ZLCL, and SRH. The spatiotemporal evolution of each of these fields in the vicinity of the forecast point—up to 1.5° away in any direction for some fields and up to 3° away in others, depending on the grid resolution (see Table 1)—throughout the forecast period was included in the predictor set to provide a comprehensive assessment of the simulated environmental conditions for each severe weather forecast. 3-hourly temporal resolution is used for Day 1 and 2 models, and 6-hourly resolution was used for Day 3. Each of the fifteen RFs—three regions, five predictands—was trained on nine years of forecasts spanning

12 April 2003–11 April 2012. The identified relationships between simulated model variables and observed severe weather during that period were assessed using RF FIs. The trained RFs were then run over an extended withheld test period spanning 12 April 2012–31 December 2016 and the performance of these forecasts assessed, both in isolation with a climatological reference and relative to SPC convective outlooks issued during the same period.

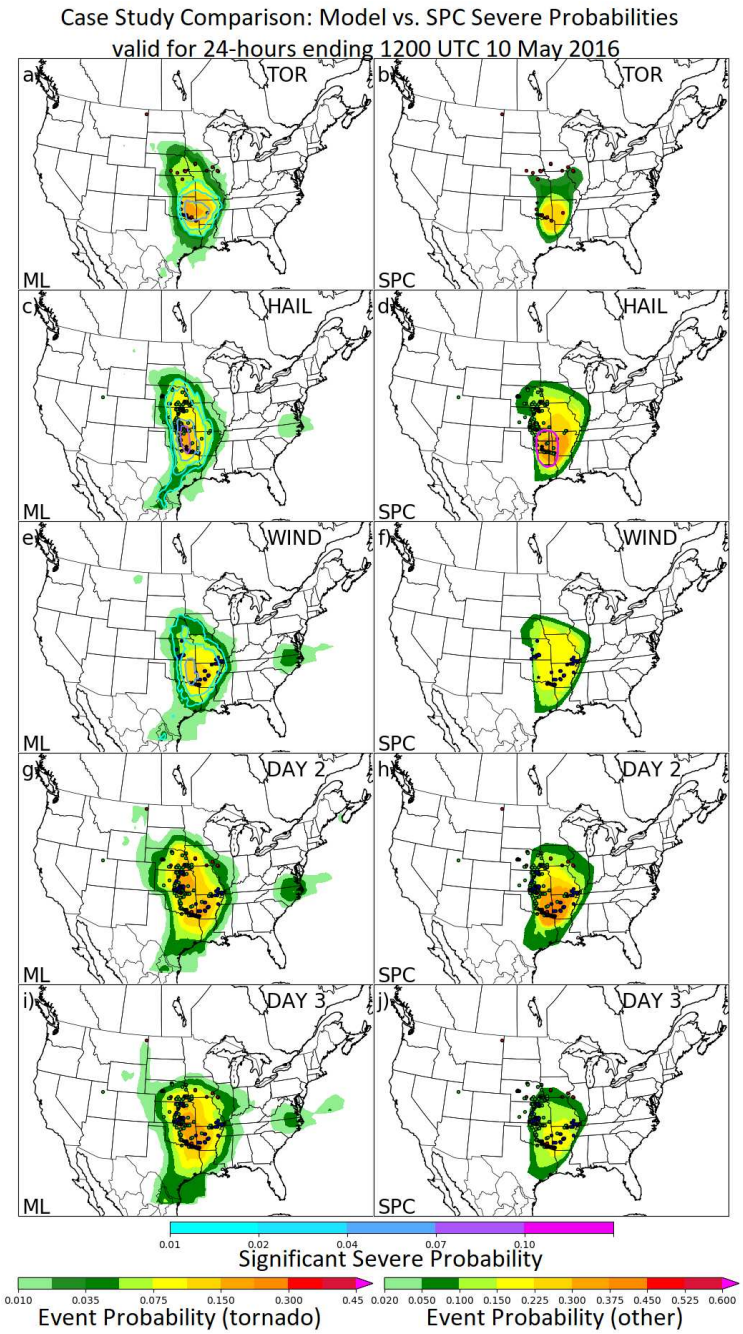


FIG. 6.18. Outlooks from the ML models and interpolated SPC contours valid for the 24-hour period ending 1200 UTC 10 May 2016 in the left and right columns, respectively. Filled contours depict severe probabilities as indicated by the corresponding colorbar on figure bottom; unfilled contours indicate significant severe probabilities for the corresponding phenomenon as applicable. Panels (a)–(b), (c)–(d), and (e)–(f) depict respectively Day 1 tornado, hail, and wind outlooks, while panels (g)–(h) and (i)–(j) show Day 2 and Day 3 outlooks issued previously for the same valid period. Severe weather reports for the period are shown with red, green, and blue circles for tornadoes, hail, and wind. Darker colored stars indicate significant severe reports for the color-corresponding phenomenon.

The statistical relationships identified by the RFs bear considerable correspondence with known physical relationships between atmospheric variables and severe weather, lending credence to the veracity of the model solutions. For example, CAPE, CIN, and wind shear—some of the most commonly used variables to characterize severe weather environments (e.g. Johns and Doswell 1992)—are consistently identified as the most predictive variables for forecasting severe weather. More nuanced identifications are made as well, including more emphasis on kinematics in tornado prediction compared with hail and wind, and additionally, wind difference over a shallower vertical layer being more predictive for tornadoes than for hail and wind. Even spatiotemporal relationships that are identified accord with physical intuition, such as meridional wind to the south of the forecast point speaking to the degree of temperature and moisture advection into the region, and upstream pressure transitioning to be over and eventually past the forecast point during the forecast period. Previously identified dynamical model biases (e.g. Wang et al. 2009, , discussed also in Chapter 4) also emerge objectively from the analysis, including the northward displacement bias of convective systems in the GEFS/R and other convection-parameterized models.

The trained models produce real-time forecasts on unseen inputs that exhibit similar spatial and quantitative character to their human-produced counterparts. In general, they produce somewhat larger regions of marginal risk equivalence and fewer incidences of moderate and high risk-equivalent outlooks. This behavior can be largely attributed to the ML-based outlooks being informed by less total real-time information—a single ensemble rather than many different models coupled with observations—and lower-resolution output than is available to the human forecaster, leading to lower confidence and higher uncertainty. Nevertheless, ML outlooks do produce across the gamut of risk categories for all lead times, and the differences in real-time forecast guidance are typically merely quantitative, rather than highlighting completely different risk areas when compared with SPC outlooks.

In terms of aggregate performance, the outlooks demonstrate impressive probabilistic forecast skill, significantly outperforming equivalent SPC outlooks at Days 2 and 3 as well as for significant severe events at Day 1, while underperforming SPC outlooks somewhat in the standard categories at Day 1. However, a weighted blend of the two outlooks statistically significantly outperformed the SPC outlooks for all phenomena and lead times, with the blend also significantly outperforming the raw ML-based outlooks at Day 1. The largest improvements came for hail and wind, with less gain seen in the tornado outlooks. Spatially, the skill gains of the blend were nearly spatial uniform, although the most gain was generally seen in the Mississippi Valley at Day 1 and the East for Day 2 and 3 with the most variability in the West owing to the low climatological frequency and small sample size. Seasonally, the largest gains at Day 1 tended to occur during the winter and spring, with the largest medium-range gains seen in the summer. Finally, the largest forecast skill improvements generally came when wind shear was relatively low, but across the spectrum of environmental CAPE.

Some limitations of this analysis should be noted. Principally, due to a combination of logistical and practical constraints, SPC outlooks are inherently limited in their probability contours, and so the human forecaster cannot issue probabilities across the entire probability spectrum like ML-models can. Some of this is partly overcome here by interpolating between SPC probability contours, which in Chapter 5 was demonstrated to yield higher probabilistic skill compared with the uninterpolated outlooks. However, some limitations remain. In particular, probabilities much above the highest risk contour, 60%, cannot be produced even with interpolation. More significantly, risk contours below the lowest risk contour—2% for tornadoes and 5% for everything else—cannot be produced at all without imposing additional assumptions about probabilities in the vicinity of but outside risk contours. Instead, all forecast probabilities outside the lowest risk contour are assumed to be zero. The ML-based outlooks frequently forecast event probabilities above 0 but below 2 or 5%, and can gain considerable probabilistic skill simply by virtue of having higher resolution in this domain of the probability space. This effect is further exacerbated for significant severe events. Here, SPC only issues a 10% risk contour, and can thus only issue 0 or 0.1 event probabilities. Forecasts above 10% do occur, but are quite rare in the ML-based outlooks, and the majority of the skill reaped in its outlooks occur from its above-climatological event probabilities that are nevertheless below 10%.

Notwithstanding these limitations, the results of this study demonstrate great promise for the application of machine learning to operational severe weather forecasting, particularly in the medium-range. Moreover, when combined with the outcomes of other studies (e.g. Herman and Schumacher

2016b, , Chapter 3), the favorable comparison with operational benchmarks across a wide range of applications suggests utility in analogous methods as a statistical post-processing tool across the broader domain of high-impact weather prediction (e.g. McGovern et al. 2017). The approach taken here is fairly simple, and based on relatively unskillful dynamical guidance compared with the current state of operational dynamical NWP. Future work that investigates use of more sophisticated pre-processing; additional physically-relevant predictors; use of additional data sources, including observations, convection-allowing guidance, and other dynamical ensembles; and more detailed and individualized treatments of the different severe weather predictands (e.g. Gagne et al. 2017) into a single synthesized machine learning-based probabilistic forecast model may yield considerable additional skill compared to what has been demonstrated here. Nevertheless, even this straightforward implementation has illustrated considerable potential benefit for using machine learning in operational severe weather forecasting, and further research in this domain is certainly warranted.

CHAPTER 7

CONCLUSIONS

This dissertation has conducted systematic efforts to better understand the best frameworks from which to approach both extreme rainfall and severe weather, diagnosing strengths and biases in existing operational forecasts and analysis tools for each forecast domain. These studies, detailed in Chapters 2 and 5, laid the foundation for subsequent studies, discussed in Chapters 3, 4, and 6, which make a concerted effort to directly improve real-time forecast quality of these high-impact weather events through ML-based SPP. Systematic investigations were conducted using over a decade of GEFS/R forecasts to train models to produce CONUS-wide probabilistic event forecasts at Days 1–3, closely mimicking existing operational products in WPC's Excessive Rainfall Outlook and SPC's Convective Outlook. Numerous sensitivity experiments were conducted to explore optimal algorithmic configuration and, more importantly, to investigate how different pieces of forecast information can be most effectively used as predictors. Examination of the internal properties of the trained models was also made. Overall, through a combination of quantitatively identifying known physical relationships and diagnosing both known and unknown recurring model biases, the post-processing models developed in this dissertation were able to add value the current state-of-the-science. In the case of extreme rainfall, the developed ML models based on the GEFS/R significantly outperformed a much better, state-of-the-science model in the ECMWF ensemble prediction system. Perhaps even more notably, for severe weather forecasts, the GEFS/R-based RFs were found to significantly outperform SPC outlooks at Days 2 and 3, and to help improve forecast skill even at Day 1. Considering the datedness of the GEFS/R compared with its higher-resolution counterparts with updated physics and numerics, this is a rather impressive feat. The relative simplicity of the approaches taken and the success across forecast problems tackled both within and outside this dissertation suggest considerable unrealized potential on other forecast problems and datasets across the gamut of weather forecasting applications. The relative simplicity in training these models further indicates a path towards steady development and implementation of these ML post-processing models in forecast operations over coming months and years.

It is worth taking a moment to consider the future role of the forecaster. Seeing such strong results from the ML algorithms developed here, the question naturally arises: will there be utility in continuing to have human weather forecasters as more of these kinds of products are developed and

implemented? Despite the demonstrated success of these algorithms and potential for even greater achievement, I would argue quite vigorously that there will, at least for the foreseeable future. Automated post-processing algorithms will always have limitations, allowing the human forecaster to add value in a number of ways. In many cases, the automated product will not be based on all information available to the human forecaster. In those circumstances, the human forecaster must be able to intelligently synthesize the post-processed model output with the external information, and be able to reconcile such input when it appears to be at odds with the statistical model forecast. Any post-processing model also imposes certain assumptions. Even in the case of a very general algorithm that imposes almost new structural assumptions on how the predictors relate to one another or to the predictand, there are still assumptions either that the training data was sufficient to provide adequate insight into the model's performance in all circumstances, or assumptions are imposed on how to extrapolate to unseen conditions. In either case, such assumptions may be invalid, and the human forecaster may add value by recognizing such circumstances and appropriately correcting for them. Third, a human forecaster often has local knowledge that an algorithm does not. In some cases, this can be providing resolution at the microclimate scale not well represented either by the dynamical or statistical model. Even for a well calibrated model, biases can still emerge at very local scales and for combinations of locality, seasonality, and weather regime; the forecaster can look for, identify, and correct for such model deficiencies where and when they do occur. A model is also only as good as it is formulated, and there can be regional differences in the utility and interpretation of the model predictand. For example, for flash flood forecasting, a model developed based on using a particular QPE source will have different biases and correspondence with actual flash flood observations in different regions of CONUS; the forecaster can be cognizant of these variations and add interpretative value in deciphering the model output. Lastly, forecast information is often highly complex, multi-faceted, and uncertain; this can make it very difficult for end users to translate raw forecast information into effective decision-making. The human forecaster plays and will continue to play a critical role in interpreting and appropriately communicating forecast information to a diverse array of end users in a way that will lead to the clearest understanding and best decision-making by those parties.

The potential avenues for future research directions are so long and abundant so as to be (metaphorically) capable of mapping a large city. The side streets and alleys are oriented in several directions. There remain many possible different predictands on which to attempt applying the SPP methodology implemented in this dissertation: different forecast problems, different datasets, different locations

and lead times. But beyond that, even the research conducted is incomplete even for the weather phenomena studied. New predictors, especially pertinent derived variables not native in model output, remain to be evaluated. More sophisticated characterizations of spatiotemporal variability of simulated fields and ensemble information remain to be explored. The pre-processing of predictors remains an unpicked yet fruitful area of future research. Chapters 3 and 4 investigated one form of pre-processing with PCA, but as noted there, PCA has structural limitations and there are many alternative approaches which may prove more effective. A more comprehensive evaluation of alternative ML algorithms presents another vast domain for future research. Like dynamical NWP models, ML algorithms are becoming increasingly sophisticated each year—to such an extent that even the techniques employed herein may be considered somewhat antiquated to some in the ML community. For example, neural networks and in particular "deep" neural networks containing many hidden layers to successively encode from the raw predictors to the final predictand afford the ability to learn optimal predictor processing automatically rather than the model developer imposing their own perceptions of those relationships directly into the algorithm by virtue of how they choose to process the raw predictors. So-called "deep learning" research is still very young when applied to NWP, but early results within NWP and in other fields suggest immense future potential continuing in this direction. This research represents an important and necessary first step in rapid advancement of NWP, but so much remains to be accomplished.

The results presented herein suggest considerable breadth to the scope of potential applications of ML methods towards effective post-processing. There are a few "highways"—areas where considerable and highly valuable research innovations remain to be made. One such highway concerns the use of intelligent post-processing to integrate disparate forecasts and tools. At present, there are a plethora of different models used for a variety of scenarios and use cases. Spatially, dynamical models span from global to continental to regional and in some cases even smaller scales. Temporally, some models are used heavily for hour-ahead forecasts, others for the day-ahead time frame, more for the medium-range, and different models still for subseasonal to seasonal forecasting. From the statistical modeling standpoint, different products exist for certain sensible weather elements and weather regimes, and as noted in Chapter 1, different products, methods, and tools exist for the relatively quiescent versus rare, high-impact scenarios. In most cases, we have a host of models giving often conflicting guidance, produced by different centers or just as independent simulations of the same model in the case of an ensemble. This reality makes certain practical sense; after all, different forecasts warrant varied

approaches to tackling the task, and a combination of computational constraints, chaos, and different underlying governing processes result in the most practical, effective choices varying for different forecast applications. Each approach also exhibits certain flaws and biases, and different approaches can give a more accurate and complete perspective of the weather situation than the use of a single "best" model. Nevertheless, space, time, magnitude—these are all *continuous* phenomena that have been discretized many times. One of the advantages of SPP, and in particular ML methods such as those employed here, is its ability to synthesize many different pieces of predictive information and distill them into a single, cohesive product using a unified framework. This has been demonstrated to some extent in this dissertation, with the developed models assimilating an ensemble of forecasts with climatological information pertinent to the forecast task. But this barely scratches the surface of the potential use of ML SPP in integrating NWP. It can be used to synthesize inputs from more models, to link products at different spatial and temporal scales, and connect the benign with the extreme to produce a more robust, streamlined, and interpretable analysis.

The benefits of conducting these additional explorations would not be simply limited to more consistent, skillful, and manageable operational, but this can also be used as a powerful tool for scientific inquiry. One justification for the abundance of different models in use is that each model has different strengths and weaknesses, and different kinds of information are more and less useful at various timescales. Present observations—station reports, radar, satellite—may be very predictive of very near-term forecasts but essentially useless for long-term prediction. Conversely, teleconnections and related indicators may be useful for speaking to average conditions over a broad spatial and temporal domain at seasonal time scales, but worthless at predicting the location of a severe storm two hours ahead. Similarly, one microphysical parameterization can perform very well in certain weather regimes while performing much inferior to the same model with a competing parameterization in other circumstances. ML post-processing can be leveraged to scientifically investigate both the predictive ability of particular facets of forecast information specific to weather conditions or forecast task. But even more innovatively, it can be used to examine the *interaction* of different sources of forecast information with the predictand. What does it mean when the model forecast portrays certain conditions but the observations depict something different? When a dynamical simulations suggests one global flow pattern but teleconnections indicate another? How about when the GFS depicts a storm in one location and the ECMWF places it somewhere else? Or when the Morrison microphysics model produces a major

rainstorm and the Thompson model does not? Every operational forecaster face these sorts of questions routinely, and must make determinations based on his or her knowledge and experiences. But the role of *joint* information such as this has been explored only very preliminarily to date, and much of it is limited to case studies as opposed to more comprehensive analysis. ML provides an objective, methodical approach to investigate these important practical questions, which can in turn be used to improve parameterizations and alleviate model biases.

Another highway concerns the interpretability of ML models. One of the biggest obstacles to the proliferation of ML both in operational forecasting as well as scientific inquiry remains the perception that these models are "black boxes". One of the most powerful aspects of ML and related algorithms is the ability to quickly and accurately analyze the massive datasets available in the meteorological community in a very general way, without imposing many assumptions about the underlying data. However, by the same token, this same flexibility afforded in analysis can make the process of understanding the findings generated all the more difficult to distill. That does not by any means make human interpretation and understanding impossible. Though by no means comprehensive, chapters 4 and 6 have taken first steps in investigating how ML forecast models can be used to gain both statistical and dynamical insights about a forecast problem. However, there remains significant unexplored knowledge encoded in the trained models, including how the model behaves in specific meteorological regimes and how different predictors interact with one another. With additional effort, such valuable insights may be reaped by scrutiny of these and other ML models. It should also be noted that some of the motivation for selecting RFs for this work came from the relative ease of interpretation. Other ML algorithms, such as neural networks and support vector machines, are even more challenging to interpret and distill. For operational forecasting, this lack of interpretability leaves the human forecaster feeling as if they cannot add any additional value beyond what is supplied by the ML model, unable to diagnose what factors and to what extent each factor has been accounted for in the model prediction. Research is ongoing to improve interpretation of ML and deep learning models; such gains could sharpen an already powerful analysis tool while simultaneously producing more tractable, usable forecasts.

A third highway relates to the often neglected "post-post-processing" step of the forecast pipeline. In this dissertation and many other parallel advancements, the NWP community is making considerable gains in the quality of automated forecast guidance through SPP. However, ultimately the most important outcome of weather forecasting is improved decision making, and one lagging area of research relates to the science and mathematics of effective translation of forecast information into decisions.

Decision making is a highly complex, multi-faceted, and multi-disciplinary process. It is very much a psychological and sociological consideration; making advances in decision science will require parallel advances in understanding how decision makers, be they the broader public or a more targeted audience, weigh various outcomes, priorities, and objectives to formulate an actionable decision. But it is also a mathematical one, and decisions are ultimately an integration of the quantified full range of possible weather outcomes (probability density function) coupled with the objective results of those outcomes. The science in all facets of this area is still young, and much remains to be research on producing, processing, and communicating forecast information in manners that result in the most effective, actionable decisions.

The final highway lies on the shadier side of town, and warrants brief additional prefacing. There are two primary, related limitations to ML-based SPP as employed in this dissertation. The first is the difficulty in incorporating new data into the post-processing as it becomes available. Advances and new information are constantly being introduced to the forecast process, including new models and analysis datasets. It is certainly desirable to be able to incorporate this process into the forecast pipeline, as is done in the human forecast process. However, current supervised learning practices use a single, consistent predictor set for training an SPP model; this requires the predictor to be generable not just for the present forecast, but through the entire training period. In the case of new predictive information becoming available mid-training period, current methods generally require an unappealing choice between not using such information or truncating the training period. Algorithmic innovations that would more directly and robustly handle missing predictors in the training set would be enormously helpful in developing skillful, long-lasting post-processing models. The second, related problem points to a prominent schism in the NWP community. One of the most persistent, recurring challenges remains the conflict between dynamical model developers and the post-processing community. The enormous advances have been made as a community over the last several decades, much remains insufficiently understood about the physics of certain atmospheric processes. These processes are being actively researched, and when new knowledge about how the system being modeled behaves is acquired, it is natural to want to incorporate that knowledge into the model. Additionally, as noted previously, many small-scale physical processes simply cannot be represented in current operational forecast models due to insufficient model resolution. Advances in computation power continue to permit increases in model resolution and changes to improve model numerics, reducing

numeric errors and allowing more accurate representation of smaller-scale processes. It is natural to want to incorporate these improvements as soon as they become practical possibilities.

However, such changes directly conflict with current best practices for effective SPP. Such changes can fundamentally alter the performance and bias characteristics of the model. If a skillful post-processing algorithm diagnoses a recurring model bias in its past history and develops an objective correction for it, applying the post-processing to that biased model will alleviate that bias. But then when information is discovered to help explain *why* that bias is occurring, the model developer may very reasonably be inclined to revise the model to alleviate that bias directly. The post-processing algorithm is unaware, however, that this bias has been removed from the model and applies the same corrections now to the detriment of the final forecast. With a frequently updated and revised model, the bias characteristics of that model are also constantly changing. This makes it extremely difficult for contemporary post-processing approaches to accurately identify biases of the *current* model version from a historical record of the model; instead, the post-processing community benefits from the model being left static for an extended period of time so that a post-processing algorithm has an opportunity to learn the errors and biases of the model and make corrections itself. In some senses, the same is true for the human forecaster: they know how to interpret and correct for deficiencies for models they've been looking at for a long time, while they cannot apply their expertise and experience in the same way when introduced to a new model.

This creates a fundamental disconnect between the optimal actions for the dynamical model developers versus those for a SPP model developer. Advances in forecast quality can surely come most rapidly when the entire NWP community is working collaboratively rather than purely competitively. The current paradigm is not particularly conducive to such a collaborative environment between the communities. In current operational practice, one of three avenues is generally pursued, either by design or happenstance. In the first, a model is updated frequently as more computation power becomes available, new parameterizations are developed, and biases are identified. Consequently, only limited post-processing can be performed to correct the latest biases in the model. This is likely the most common of the three; one prominent example following such a model is the High Resolution Rapid Refresh (HRRR). In the second approach, the model remains static in the same configuration for an extended period; the model becomes antiquated as newer models are introduced, but its biases can be corrected for statistically through SPP. One notable historical example of this is the Nested Grid Model (NGM), introduced in 1987; a MOS system was developed to post-process the NGM which was

heavily used in forecast operations until 2009, despite the introduction of more modern models in the Eta and AVN/MRF models in the 1990s and later the NAM and GFS, respectively, earlier in the 2000s. The third approach gets the best of both worlds: introduce revisions to the dynamical model as they are available, but provide a new record of reforecasts of historical cases each time such an upgrade is made. This allows dynamical model developers to make improvements to the model, but also allows for the development of robust SPP algorithms to be developed and tailored to the particular model. The long record of reforecasts generated for the GEFS/R was necessary for the creation of the models developed in this dissertation, and like approaches would likely not have been as effective with a constantly updating model. For example, Herman and Schumacher (2016b) found substantially degraded performance using only a 9 month training period compared with one spanning 3 years. This problem becomes further amplified with increasingly extreme, high-impact events, as even more cases must be collected in order to acquire an adequately large sampling of historical instances of the extreme event. However, this approach has one major limitation: producing these reforecasts can be extraordinarily computationally expensive and time consuming. Resources are often not afforded for this purpose, making the generation of a reforecast dataset infeasible.

A paradigm shift may be necessary to mend this gap. One approach may be to move away from the "offline" model framework used in this dissertation and commonly used in SPP more generally, whereby a model is trained from a set of historical cases to produce a static model that generates predictions for new, unseen cases. In an alternative "online" framework, the model itself updates as new forecasts are supplied and issued. Although not perfect, over time, the model will adapt to changes in underlying bias characteristics of the supplied predictors. This approach offers some clear advantages, but still either suffers from the effects resulting from predictors with inconsistent biases or from the adverse consequences of a shortened training set. Taking the "online" approach one radical step farther, another possibility is to incorporate ML *directly* into a dynamical model via the parameterizations or even the dynamical core. Many aspects about dynamical model formulation remain uncertain, either because of insufficient knowledge about a physical process, a tradeoff in numerical errors from simplification or truncation, or a combination. Recognizing this, there has been a proliferation of stochastic physics schemes in recent years, using random plausible values rather than a single, arbitrarily decreed one. However, this could be taken a step farther. The forecasts produced given certain combinations of values for uncertain parameters could be recorded, with the forecast *outcomes* documented as well. Using ML, the model could learn over time which combination of parameter values verify best under

different meteorological regimes, in essence adapting parameterization and other structural simplifications to be the least damaging for the specific present meteorological conditions. By continually updating the distribution of parameter values from which to sample as new forecasts are produced, the model can dynamically correct for biases as the model encounters new or different weather regimes, but also adapt organically to other model changes as those are made, without any abrupt changes in performances characteristics.

The path is long. The possibilities are many. The future is bright.

REFERENCES

Abbe, C., 1901: The physical basis of long-range weather forecasts. *Mon. Wea. Rev.*, **29**, 551–561.

Adams-Selin, R. D. and C. L. Ziegler, 2016: Forecasting hail using a one-dimensional hail growth model within WRF. *Mon. Wea. Rev.*, **144**, 4919–4939.

Agee, E. and S. Childs, 2014: Adjustments in tornado counts, F-scale intensity, and path width for assessing significant tornado destruction. *J. Appl. Meteor. Climatol.*, **53**, 1494–1505, doi: 10.1175/JAMC-D-13-0235.1.

AghaKouchak, A., A. Behrangi, S. Sorooshian, K. Hsu, and E. Amitai, 2011: Evaluation of satellite-retrieved extreme precipitation rates across the central United States. *J. Geophys. Res.*, **116**.

Agresti, A. and B. A. Coull, 1998: Approximate is better than âĂIJexactâĂİ for interval estimation of binomial proportions. *The American Statistician*, **52**, 119–126, doi: 10.1080/00031305.1998.10480550.

Ahijevych, D., J. O. Pinto, J. K. Williams, and M. Steiner, 2016: Probabilistic forecasts of mesoscale convective system initiation using the random forest data mining technique. *Wea. Forecasting*, **31**, 581–599.

Alessandrini, S., L. D. Monache, C. M. Rozoff, and W. E. Lewis, 2018: Probabilistic prediction of tropical cyclone intensity with an analog ensemble. *Mon. Wea. Rev.*, **in press**.

Alvarez, F. M., 2014: Statistical calibration of extended-range probabilistic tornado forecasts with a reforecast dataset. Ph.D. thesis, SAINT LOUIS UNIVERSITY, 210 pp.

Anderson, C. J., C. K. Wikle, Q. Zhou, and J. A. Royle, 2007: Population influences on tornado reports in the United States. *Wea. Forecasting*, **22**, 571–579, doi: 10.1175/WAF997.1.

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

Anthony, R. W. and P. W. Leftwich, Jr., 1992: Trends in severe local storm watch verification at the National Severe Storms Forecast Center. *Wea. Forecasting*, **7**, 613–622, doi: 10.1175/1520-0434(1992)007<0613:TISLSW>2.0.CO;2.

Antolik, M. S., 2000: An overview of the National Weather Service's centralized statistical quantitative precipitation forecasts. *J. Hydrol.*, **239 (1)**, 306–337.

Applequist, S., G. E. Gahrs, R. L. Pfeffer, and X.-F. Niu, 2002: Comparison of methodologies for probabilistic quantitative precipitation forecasting*. *Wea. Forecasting*, **17**, 783–799.

Ashley, S. T. and W. S. Ashley, 2008: Flood fatalities in the United States. *J. Appl. Meteor. Climatol.*, **47**, 805–818.

Baars, J. A. and C. F. Mass, 2005: Performance of National Weather Service forecasts compared to operational, consensus, and weighted model output statistics. *Wea. Forecasting*, **20 (6)**, 1034–1047.

Baggett, C. F., K. M. Nardi, S. J. Childs, S. N. Zito, E. A. Barnes, and E. D. Maloney, 2018: Skillful five week forecasts of tornado and hail activity. *J. Geophys. Res.*, **submitted**.

Baldwin, M. E. and J. S. Kain, 2006: Sensitivity of several performance measures to displacement error, bias, and event frequency. *Wea. Forecasting*, **21**, 636–648, doi: 10.1175/WAF933.1.

Barnes, L. R., E. C. Gruntfest, M. H. Hayden, D. M. Schultz, and C. Benight, 2007: False alarms and close calls: A conceptual model of warning accuracy. *Wea. Forecasting*, **22**, 1140–1147, doi: 10.1175/WAF1031.1.

Barthold, F. E., T. E. Workoff, B. A. Cosgrove, J. J. Gourley, D. R. Novak, and K. M. Mahoney, 2015: Improving flash flood forecasts: The HMT-WPC flash flood and intense rainfall experiment. *Bull. Amer. Meteor. Soc.*, **96**, 1859–1866.

Bentzien, S. and P. Friederichs, 2012: Generating and calibrating probabilistic quantitative precipitation forecasts from the high-resolution NWP model COSMO-DE. *Wea. Forecasting*, **27**, 988–1002.

Bermingham, A. and A. Smeaton, 2011: On using Twitter to monitor political sentiment and predict election results. *Proceedings of the Workshop on Sentiment Analysis where AI meets Psychology (SAAIP 2011)*, 2–10.

Bermowitz, R. J., 1975: An application of model output statistics to forecasting quantitative precipitation. *Mon. Wea. Rev.*, **103 (2)**, 149–153.

Bieringer, P. and P. S. Ray, 1996: A comparison of tornado warning lead times with and without NEXRAD Doppler radar. *Wea. Forecasting*, **11**, 47–52, doi: 10.1175/1520-0434(1996)011<0047:ACOTWL>2.0.CO;2.

Bolton, D., 1980: The computation of equivalent potential temperature. *Mon. Wea. Rev.*, **108**, 1046–1053.

Bonnin, G. M., D. Martin, B. Lin, T. Parzybok, M. Yekta, and D. Riley, 2006: Precipitation-frequency atlas of the United States. *NOAA Atlas 14*, **2**.

Bonnin, G. M., D. Todd, B. Lin, T. Parzybok, M. Yekta, and D. Riley, 2004: Precipitation-frequency atlas of the United States. *NOAA Atlas 14*, **1**.

Bothwell, P., J. Hart, and R. Thompson, 2002: An integrated three-dimensional objective analysis scheme in use at the Storm Prediction Center. *Preprints, 21st Conf. on Severe Local Storms, San Antonio, TX, Amer. Meteor. Soc.*, J117–J120.

Bougeault, P., et al., 2010: The thorpex interactive grand global ensemble. *Bull. Amer. Meteor. Soc.*, **91 (8)**, 1059–1072.

Breiman, L., 2001: Random forests. *Mach. Learn.*, **45**, 5–32, doi: 10.1023/A:1010933404324.

Bremnes, J. B., 2004: Probabilistic forecasts of precipitation in terms of quantiles using NWP model output. *Mon. Wea. Rev.*, **132**, 338–347.

Brier, G. W., 1950: Verification of forecasts expressed in terms of probability. *Mon. Wea. Rev.*, **78**, 1–3, doi: 10.1175/1520-0493(1950)078<0001:VOFEIT>2.0.CO;2.

Brimelow, J. C., G. W. Reuter, R. Goodson, and T. W. Krauss, 2006: Spatial forecasts of maximum hail size using prognostic model soundings and HAILCAST. *Wea. Forecasting*, **21**, 206–219, doi: 10.1175/WAF915.1.

Brocca, L., F. Melone, and T. Moramarco, 2008: On the estimation of antecedent wetness conditions in rainfall–runoff modelling. *Hydrological Processes*, **22**, 629–642.

Bröcker, J. and L. A. Smith, 2007: Increasing the reliability of reliability diagrams. *Wea. Forecasting*, **22**, 651–661, doi: 10.1175/WAF993.1.

Brodley, C. E. and P. E. Utgoff, 1995: Multivariate decision trees. *Machine learning*, **19 (1)**, 45–77.

Brooks, H. E., C. A. Doswell III, and M. P. Kay, 2003: Climatological estimates of local daily tornado probability for the united states. *Wea. Forecasting*, **18 (4)**, 626–640.

Brooks, H. E. and D. J. Stensrud, 2000: Climatology of heavy rain events in the United States from hourly precipitation observations. *Mon. Wea. Rev.*, **128**, 1194–1201.

Brotzge, J., S. Erickson, and H. Brooks, 2011: A 5-yr climatology of tornado false alarms. *Wea. Forecasting*, **26**, 534–544, doi: 10.1175/WAF-D-10-05004.1.

Buizza, R., A. Hollingsworth, F. Lalaurette, and A. Ghelli, 1999: Probabilistic predictions of precipitation using the ecmwf ensemble prediction system. *Wea. Forecasting*, **14 (2)**, 168–189.

Calianno, M., I. Ruin, and J. J. Gourley, 2013: Supplementing flash flood reports with impact classifications. *J. Hydrol.*, **477**, 1–16.

Cao, L.-J. and F. E. H. Tay, 2003: Support vector machine with adaptive parameters in financial time series forecasting. *IEEE Transactions on neural networks*, **14**, 1506–1518.

Carpenter, T., J. Sperfslage, K. Georgakakos, T. Sweeney, and D. Fread, 1999: National threshold runoff estimation utilizing GIS in support of operational flash flood warning systems. *J. Hydrology*, **224**, 21–44.

Castillo, V., A. Gomez-Plaza, and M. Martınez-Mena, 2003: The role of antecedent soil water content in the runoff response of semiarid catchments: a simulation approach. *J. Hydrology*, **284**, 114–130.

Childs, C., 2004: Interpolating surfaces in ArcGIS spatial analyst. *ArcUser*, **3235**, 569.

Clark, A. J., W. A. Gallus Jr, and T.-C. Chen, 2007: Comparison of the diurnal precipitation cycle in convection-resolving and non-convection-resolving mesoscale models. *Mon. Wea. Rev.*, **135**, 3456–3473.

Clark, A. J., W. A. Gallus Jr, and M. L. Weisman, 2010: Neighborhood-based verification of precipitation forecasts from convection-allowing NCAR WRF model simulations and the operational NAM. *Wea. Forecasting*, **25**, 1495–1509.

Clark, R. A., J. J. Gourley, Z. L. Flamig, Y. Hong, and E. Clark, 2014: Conus-wide evaluation of national weather service flash flood guidance products. *Wea. Forecasting*, **29**, 377–392.

Collobert, R., J. Weston, L. Bottou, M. Karlen, K. Kavukcuoglu, and P. Kuksa, 2011: Natural language processing (almost) from scratch. *Journal of Machine Learning Research*, **12**, 2493–2537.

Cortes, C. and V. Vapnik, 1995: Support-vector networks. *Mach. Learn.*, **20 (3)**, 273–297, doi: 10.1007/BF00994018.

Davies, J. M. and R. H. Johns, 1993: Some wind and instability parameters associated with strong and violent tornadoes: 1. wind shear and helicity. *The Tornado: Its Structure, Dynamics, Prediction, and Hazards*, 573–582.

Davis, C. A., K. W. Manning, R. E. Carbone, S. B. Trier, and J. D. Tuttle, 2003: Coherence of warm-season continental rainfall in numerical weather prediction models. *Mon. Wea. Rev.*, **131**, 2667–2679.

Davis, J. M. and M. D. Parker, 2014: Radar climatology of tornadic and nontornadic vortices in high-shear, low-cape environments in the mid-Atlantic and southeastern United States. *Wea. Forecasting*, **29**, 828–853, doi: 10.1175/WAF-D-13-00127.1.

Davis, R. S., 2001: Flash flood forecast and detection methods. *Severe Convective Storms*, Springer, 481–525.

Dean, A. R., R. S. Schneider, R. L. Thompson, J. Hart, and P. D. Bothwell, 2009: The conditional risk of severe convection estimated from archived NWS/Storm Prediction Center mesoscale objective analyses: Potential uses in support of forecast operations and verification. *Preprints, 23rd Conf. on Weather Analysis and Forecasting/19th Conf. on Numerical Weather Prediction, Omaha, NE, Amer. Meteor. Soc.*, 6A.5.

Delrieu, G., et al., 2005: The catastrophic flash-flood event of 8–9 September 2002 in the Gard Region, France: a first case study for the cévennes–vivarais mediterranean hydrometeorological observatory. *J. Hydrometeor.*, **6 (1)**, 34–52.

DeMaria, M. and J. Kaplan, 1994: A statistical hurricane intensity prediction scheme (SHIPS) for the Atlantic basin. *Wea. Forecasting*, **9**, 209–220.

Doswell, C. A., III, 2007: Small sample size and data quality issues illustrated using tornado occurrence data. *Electronic J. Severe Storms Meteor.*, **2 (5)**.

Doswell, C. A., III, H. E. Brooks, and R. A. Maddox, 1996: Flash flood forecasting: An ingredients-based methodology. *Wea. Forecasting*, **11**, 560–581.

Doswell, C. A., III, R. Davies-Jones, and D. L. Keller, 1990: On summary measures of skill in rare event forecasting based on contingency tables. *Wea. Forecasting*, **5**, 576–585, doi: 10.1175/1520-0434(1990)005<0576:OSMOSI>2.0.CO;2.

Doswell, C. A., III, S. J. Weiss, and R. H. Johns, 1993: Tornado forecasting: A review. *The Tornado: Its Structure, Dynamics, Prediction, and Hazards, Geophys. Monogr.*, No. 79, Amer. Geophys. Union., 557–571.

Doswell III, C. A., 2004: Weather forecasting by humans: Heuristics and decision making. *Wea. Forecasting*, **19**, 1115–1126.

Doswell III, C. A. and D. M. Schultz, 2006: On the use of indices and parameters in forecasting severe storms. *Electronic J. Severe Storms Meteor.*, **1**, URL http://www.ejssm.org/ojs/index.php/ejssm/article/viewArticle/11.

Drobot, S. and D. J. Parker, 2007: Advances and challenges in flash flood warnings. *Environ. Hazards*, **7**, 173–178, doi: 10.1016/j.envhaz.2007.09.001.

Duda, J. D. and W. A. Gallus, 2013: The impact of large-scale forcing on skill of simulated convective initiation and upscale evolution with convection-allowing grid spacings in the WRF*. *Wea. Forecasting*, **28**, 994–1018.

Durran, D. R., 2010: *Numerical methods for fluid dynamics: With applications to geophysics*, Vol. 32. Springer Science & Business Media.

Eckel, F. A., 2003: Effective mesoscale, short-range ensemble forecasting. Ph.D. thesis, University of Washington.

Edwards, R. and G. W. Carbin, 2016: Estimated convective winds: Reliability and effects on severe-storm climatology. *Preprints, 28th Conf. on Severe Local Storms, Portland, OR, Amer. Meteor. Soc., 14B.6*, [Available online at https://ams.confex.com/ams/28SLS/webprogram/Manuscript/Paper300279/sls-estd.pdf].

Edwards, R., G. W. Carbin, and S. F. Corfidi, 2015: Overview of the Storm Prediction Center. *Preprints, 13th History Symp., Phoenix, AZ, Amer. Meteor. Soc., 1.1*, [Available online at http://www.spc.noaa.gov/publications/edwards/spc-over.pdf].

Elsner, J. B. and H. M. Widen, 2014: Predicting spring tornado activity in the central Great Plains by 1 March. *Mon. Wea. Rev.*, **142**, 259–267.

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

Ferree, J., 2009: National change of the hail criteria for severe storms from 3/4 inch to 1 inch beginning January 5, 2010. National Weather Service, [Available online at http://www.nws.noaa.gov/oneinchhail/docs/One_Inch_Hail.pdf].

Ferree, J. T., J. Looney, and K. Waters, 2006: NOAA/National Weather ServicesâĂŹ storm-based warnings. *Extended Abstracts, 23rd Conference on Severe Local Storms*, St. Louis, MO, Amer. Meteor. Soc., [Available online at https://ams.confex.com/ams/23SLS/techprogram/paper_115513.htm].

Friedman, J. H., 1997: On bias, variance, 0/1 loss, and the curse-of-dimensionality. *Data mining and knowledge discovery*, **1 (1)**, 55–77, doi: https://doi.org/10.1023/A:1009778005914, URL

https://link.springer.com/article/10.1023/A:1009778005914.

Friedman, J. H., 2001: Greedy function approximation: a gradient boosting machine. *Ann. Stat.*, 1189–1232, doi: 10.1214/aos/1013203451.

Fritsch, J. M. and R. Carbone, 2004: Improving quantitative precipitation forecasts in the warm season: A USWRP research and development strategy. *Bull. Amer. Meteor. Soc.*, **85 (7)**, 955–965, doi: 10.1175/BAMS-85-7-955.

Fulton, R. A., J. P. Breidenbach, D.-J. Seo, D. A. Miller, and T. OâĂŹBannon, 1998: The WSR-88D rainfall algorithm. *Wea. Forecasting*, **13 (2)**, 377–395.

Gagne, D. J., A. McGovern, J. B. Basara, and R. A. Brown, 2012: Tornadic supercell environments analyzed using surface and reanalysis data: a spatiotemporal relational data-mining approach. *J. Appl. Meteor. Climatol.*, **51**, 2203–2217.

Gagne, D. J., A. McGovern, and J. Brotzge, 2009: Classification of convective areas using decision trees. *J. Atmos. Oceanic Technol.*, **26**, 1341–1353.

Gagne, D. J., A. McGovern, S. E. Haupt, R. A. Sobash, J. K. Williams, and M. Xue, 2017: Storm-based probabilistic hail forecasting with machine learning applied to convection-allowing ensembles. *Wea. Forecasting*, **32**, 1819–1840.

Gagne, D. J., A. McGovern, and M. Xue, 2014: Machine learning enhancement of storm-scale ensemble probabilistic quantitative precipitation forecasts. *Wea. Forecasting*, **29**, 1024–1043, doi: https://doi.org/10.1175/WAF-D-13-00108.1.

Gagne, D. J., II, A. McGovern, J. Brotzge, M. Coniglio, J. Correia Jr, and M. Xue, 2015: Day-ahead hail prediction integrating machine learning with storm-scale numerical weather models. *AAAI*, 3954–3960.

Gagne, D. J. I., 2016: Coupling data science techniques and numerical weather prediction models for high-impact weather prediction. Ph.D. thesis, University of Oklahoma, URL https://shareok.org/handle/11244/44917.

Gallo, B. T., A. J. Clark, B. T. Smith, R. L. Thompson, I. Jirak, and S. R. Dembek, 2018: Blended probabilistic tornado forecasts: Combining climatological frequencies with NSSL–WRF ensemble forecasts. *Wea. Forecasting*, **33**, 443–460.

Gallus, W. A., Jr., N. A. Snook, and E. V. Johnson, 2008: Spring and summer severe weather reports over the Midwest as a function of convective mode: A preliminary study. *Wea. Forecasting*, **23**, 101–113, doi: 10.1175/2007WAF2006120.1.

Gandin, L. S. and A. H. Murphy, 1992: Equitable skill scores for categorical forecasts. *Mon. Wea. Rev.*, **120**, 361–370.

Gensini, V. A. and W. S. Ashley, 2011: Climatology of potentially severe convective environments from the North American Regional Reanalysis. *Electronic J. Severe Storms Meteor.*, **6 (8)**.

Gensini, V. A., T. L. Mote, and H. E. Brooks, 2014: Severe-thunderstorm reanalysis environments and collocated radiosonde observations. *J. Appl. Meteor. Climatol.*, **53**, 742–751, doi: https://doi.org/10.1175/JAMC-D-13-0263.1.

Geurts, P., D. Ernst, and L. Wehenkel, 2006: Extremely randomized trees. *Machine learning*, **63 (1)**, 3–42.

Glahn, H. R. and D. A. Lowry, 1972: The use of model output statistics (MOS) in objective weather forecasting. *J. Appl. Meteor.*, **11 (8)**, 1203–1211, doi: 10.1175/1520-0450(1972)011<1203:TUOMOS>2.0.CO;2.

Gneiting, T., A. E. Raftery, A. H. Westveld III, and T. Goldman, 2005: Calibrated probabilistic forecasting using ensemble model output statistics and minimum CRPS estimation. *Mon. Wea. Rev.*, **133**, 1098–1118.

Gochis, D., et al., 2015: The great Colorado flood of September 2013. *Bull. Amer. Meteor. Soc.*, **96 (9)**, 1461–1487.

Gourley, J. J., J. M. Erlingis, Y. Hong, and E. B. Wells, 2012: Evaluation of tools used for monitoring and forecasting flash floods in the United States. *Wea. Forecasting*, **27**, 158–173.

Gourley, J. J., et al., 2013: A unified flash flood database across the United States. *Bull. Amer. Meteor. Soc.*, **94**, 799–805, doi: 10.1175/BAMS-D-12-00198.1.

Gourley, J. J., et al., 2017: The FLASH project: Improving the tools for flash flood monitoring and prediction across the United States. *Bull. Amer. Meteor. Soc.*, **98**, 361–372.

Graham, R. A. and R. H. Grumm, 2010: Utilizing normalized anomalies to assess synoptic-scale weather events in the western united states. *Wea. Forecasting*, **25 (2)**, 428–445.

Grams, J. S., W. A. Gallus Jr, S. E. Koch, L. S. Wharton, A. Loughe, and E. E. Ebert, 2006: The use of a modified Ebert-McBride technique to evaluate mesoscale model QPF as a function of convective system morphology during IHOP 2002. *Wea. Forecasting*, **21**, 288–306.

Hagedorn, R., T. M. Hamill, and J. S. Whitaker, 2008: Probabilistic forecast calibration using ECMWF and GFS ensemble reforecasts. Part I: Two-meter temperatures. *Mon. Wea. Rev.*, **136 (7)**, 2608–2619.

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

Hall, T., H. E. Brooks, and C. A. Doswell III, 1999: Precipitation forecasting using a neural network. *Wea. Forecasting*, **14**, 338–345.

Hamill, T. M., 2017: Changes in the systematic errors of global reforecasts due to an evolving data assimilation system. *Mon. Wea. Rev.*, **145**, 2479–2485.

Hamill, T. M., G. T. Bates, J. S. Whitaker, D. R. Murray, M. Fiorino, T. J. Galarneau Jr, Y. Zhu, and W. Lapenta, 2013: NOAA's second-generation global medium-range ensemble reforecast dataset. *Bull. Amer. Meteor. Soc.*, **94**, 1553–1565, doi: 10.1175/BAMS-D-12-00014.1.

Hamill, T. M., R. Hagedorn, and J. S. Whitaker, 2008: Probabilistic forecast calibration using ECMWF and GFS ensemble reforecasts. part II: Precipitation. *Mon. Wea. Rev.*, **136**, 2620–2632.

Hamill, T. M. and J. Juras, 2006: Measuring forecast skill: Is it real skill or is it the varying climatology? *Quart. J. Roy. Meteor. Soc.*, **132**, 2905–2924.

Hamill, T. M., M. Scheuerer, and G. T. Bates, 2015: Analog probabilistic precipitation forecasts using GEFS reforecasts and climatology-calibrated precipitation analyses. *Mon. Wea. Rev.*, **143**, 3300–3309.

Hamill, T. M. and J. S. Whitaker, 2006: Probabilistic quantitative precipitation forecasts based on reforecast analogs: Theory and application. *Mon. Wea. Rev.*, **134 (11)**, 3209–3229.

Hamill, T. M., J. S. Whitaker, and X. Wei, 2004: Ensemble reforecasting: Improving medium-range forecast skill using retrospective forecasts. *Mon. Wea. Rev.*, **132 (6)**, 1434–1447.

Hapuarachchi, H., Q. Wang, and T. Pagano, 2011: A review of advances in flash flood forecasting. *Hydrological Processes*, **25**, 2771–2784.

Hart, J. A. and A. E. Cohen, 2016: The challenge of forecasting significant tornadoes from June to October using convective parameters. *Wea. Forecasting*, **31**, 2075–2084, doi: 10.1175/WAF-D-16-0005.1.

Herman, G. R., 2018: Forecasting severe weather with random forests. *Mon. Wea. Rev.*, **in revisions**.

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

Herman, G. R. and R. S. Schumacher, 2016a: Extreme precipitation in models: An evaluation. *Wea. Forecasting*, **31**, 1853–1879, doi: 10.1175/WAF-D-16-0093.1.

Herman, G. R. and R. S. Schumacher, 2016b: Using reforecasts to improve forecasting of fog and visibility for aviation. *Wea. Forecasting*, **31**, 467–482.

Herman, G. R. and R. S. Schumacher, 2018a: Dendrology in numerical weather prediction: What random forests and logistic regression tell us about forecasting extreme precipitation. *Mon. Wea. Rev.*, **146**, 1785–1812, doi: 10.1175/MWR-D-17-0307.1.

Herman, G. R. and R. S. Schumacher, 2018b: Flash flood verification: pondering precipitation proxies. *J. Hydrometeor.*, **in press**, doi: 10.1175/JHM-D-18-0092.1.

Herman, G. R. and R. S. Schumacher, 2018c: Money doesn't grow on trees, but forecasts do: Forecasting extreme precipitation with random forests. *Mon. Wea. Rev.*, **146**, 1571–1600, doi: 10.1175/MWR-D-17-0250.1.

Hershfield, D. M., 1961: Technical paper no. 40: Rainfall frequency atlas of the United States. *Weather Bureau, Department of Commerce, Washington, DC*.

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

Hitchens, N. M. and H. E. Brooks, 2013: Preliminary investigation of the contribution of supercell thunderstorms to the climatology of heavy and extreme precipitation in the United States. *Atmospheric research*, **123**, 206–210.

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

Hitchens, N. M. and H. E. Brooks, 2017: Determining criteria for missed events to evaluate significant severe convective outlooks. *Wea. Forecasting*, **32**, 1321–1328, doi: 10.1175/WAF-D-16-0170.1.

Hitchens, N. M., H. E. Brooks, and R. S. Schumacher, 2013: Spatial and temporal characteristics of heavy hourly rainfall in the United States. *Mon. Wea. Rev.*, **141**, 4564–4575.

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

Hong, T., P. Pinson, S. Fan, H. Zareipour, A. Troccoli, and R. J. Hyndman, 2016: Probabilistic energy forecasting: Global energy forecasting competition 2014 and beyond. *International Journal of Forecasting*, **32**, 896–913.

Hope, J. R. and C. J. Neumann, 1970: An operational technique for relating the movement of existing tropical cyclones to past tracks. *Mon. Wea. Rev.*, **98 (12)**, 925–933.

Hou, D., et al., 2014: Climatology-calibrated precipitation analysis at fine scales: Statistical adjustment of stage IV toward CPC gauge-based analysis. *J. Hydrometeor.*, **15**, 2542–2557.

Igel, A. L., M. R. Igel, and S. C. van den Heever, 2015: Make it a double? sobering results from simulations using single-moment microphysics schemes. *J. Atmos. Sci.*, **72**, 910–925.

Jacks, E., 2014: Service change notice 14-42. National Weather Service, Fire and Public Weather Services Branch, [Available online at http://www.nws.noaa.gov/os/notification/scn14-42day1-3outlooks_cca.htm].

Jacks, E., J. B. Bower, V. J. Dagostaro, J. P. Dallavalle, M. C. Erickson, and J. C. Su, 1990: New NGM-based MOS guidance for maximum/minimum temperature, probability of precipitation, cloud amount, and surface wind. *Wea. Forecasting*, **5 (1)**, 128–138.

Jarvinen, B. R. and C. J. Neumann, 1979: Statistical forecasts of tropical cyclone intensity for the North Atlantic basin.

Johns, R. H. and C. A. Doswell, III, 1992: Severe local storms forecasting. *Wea. Forecasting*, **7**, 588–612.

Jolliffe, I. T. and D. B. Stephenson, 2003: *Forecast verification: a practitioner's guide in atmospheric science*. John Wiley & Sons.

Jones, T. A., D. Cecil, and M. DeMaria, 2006: Passive-microwave-enhanced statistical hurricane intensity prediction scheme. *Wea. Forecasting*, **21**, 613–635.

Junker, N. W., M. J. Brennan, F. Pereira, M. J. Bodner, and R. H. Grumm, 2009: Assessing the potential for rare precipitation events with standardized anomalies and ensemble guidance at the hydrometeorological prediction center. *Bull. Amer. Meteor. Soc.*, **90 (4)**, 445–453.

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., et al., 2008: Some practical considerations regarding horizontal resolution in the first generation of operational convection-allowing NWP. *Wea. Forecasting*, **23**, 931–952.

Kalnay, E., 2003: *Atmospheric modeling, data assimilation and predictability*. Cambridge university press.

Karstens, C. D., et al., 2018: Development of a human-machine mix for forecasting severe convective events. *Wea. Forecasting*, **in press**.

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

Khain, A., et al., 2015: Representation of microphysical processes in cloud-resolving models: Spectral (bin) microphysics versus bulk parameterization. *Rev. Geophys.*, **53**, 247–322.

Klein, W. H., B. M. Lewis, and I. Enger, 1959: Objective prediction of five-day mean temperatures during winter. *J. Meteor.*, **16 (6)**, 672–682.

Krocak, M. J. and H. E. Brooks, 2018: Climatological estimates of hourly tornado probability for the United States. *Wea. Forecasting*, **33**, 59–69.

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

Kunkel, K. E., D. R. Easterling, D. A. Kristovich, B. Gleason, L. Stoecker, and R. Smith, 2012: Meteorological causes of the secular variations in observed extreme precipitation events for the conterminous United States. *J. Hydrometeor.*, **13**, 1131–1141.

Lackmann, G., 2011: *Midlatitude Synoptic Meteorology*. American Meteorological Society, 360 pp.

Lackmann, G. M., 2013: The south-central US flood of May 2010: Present and future. **26 (13)**, 4688–4709.

Lagerquist, R., A. McGovern, and T. Smith, 2017: Machine learning for real-time prediction of damaging straight-line convective wind. *Wea. Forecasting*, **32**, 2175–2193.

Larranaga, P., et al., 2006: Machine learning in bioinformatics. *Briefings in bioinformatics*, 86–112.

Lean, H. W., P. A. Clark, M. Dixon, N. M. Roberts, A. Fitch, R. Forbes, and C. Halliwell, 2008: Characteristics of high-resolution versions of the Met Office Unified Model for forecasting convection over the United Kingdom. *Mon. Wea. Rev.*, **136**, 3408–3424.

Lin, Y. and K. E. Mitchell, 2005: The NCEP Stage II/IV hourly precipitation analyses: Development and applications. *19th Conf. Hydrology, American Meteorological Society, San Diego, CA, USA*, Citeseer.

Liu, H., V. Chandrasekar, and G. Xu, 2001: An adaptive neural network scheme for radar rainfall estimation from WSR-88D observations. *J. Appl. Meteor.*, **40 (11)**, 2038–2050.

Lorenz, E. N., 1956: Empirical orthogonal functions and statistical weather prediction.

Lorenz, E. N., 1963: Deterministic nonperiodic flow. *J. Atmos. Sci.*, **20 (2)**, 130–141.

Loridan, T., R. P. Crompton, and E. Dubossarsky, 2017: A machine learning approach to modeling tropical cyclone wind field uncertainty. *Mon. Wea. Rev.*, **145**, 3203–3221.

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

Marjerison, R. D., M. T. Walter, P. J. Sullivan, and S. J. Colucci, 2016: Does population affect the location of flash flood reports? *J. Appl. Meteor. Climatol.*, **55**, 1953–1963.

Markowski, P. and Y. Richardson, 2010: *Mesoscale Meteorology in Midlatitudes*. John Wiley & Sons, 424 pp.

Marzban, C., 1998: Scalar measures of performance in rare-event situations. *Wea. Forecasting*, **13**, 753–763.

Marzban, C. and G. J. Stumpf, 1996: A neural network for tornado prediction based on Doppler radar-derived attributes. *J. Appl. Meteor.*, **35**, 617–626.

Marzban, C. and G. J. Stumpf, 1998: A neural network for damaging wind prediction. *Wea. Forecasting*, **13**, 151–163.

Marzban, C. and A. Witt, 2001: A Bayesian neural network for severe-hail size prediction. *Wea. Forecasting*, **16**, 600–610.

McGovern, A., K. L. Elmore, D. J. Gagne, S. E. Haupt, C. D. Karstens, R. Lagerquist, T. Smith, and J. K. Williams, 2017: Using artificial intelligence to improve real-time decision making for high-impact weather. *Bull. Amer. Meteor. Soc.*, **98**, 2073–2090, doi: 10.1175/BAMS-D-16-0123.1.

McGovern, A., D. J. Gagne, J. K. Williams, R. A. Brown, and J. B. Basara, 2014: Enhancing understanding and improving prediction of severe weather through spatiotemporal relational learning. *Machine Learning*, **95**, 27–50, doi: 10.1007/s10994-013-5343-x.

McGovern, A., D. John Gagne, N. Troutman, R. A. Brown, J. Basara, and J. K. Williams, 2011: Using spatiotemporal relational random forests to improve our understanding of severe weather processes. *Statistical Analysis and Data Mining: The ASA Data Science Journal*, **4**, 407–429.

Meierdiercks, K. L., J. A. Smith, M. L. Baeck, and A. J. Miller, 2010: Analyses of urban drainage network structure and its impact on hydrologic response. *J. Amer. Water Resources Association*, **46 (5)**, 932–943.

Mercer, A. E., C. M. Shafer, C. A. Doswell III, L. M. Leslie, and M. B. Richman, 2012: Synoptic composites of tornadic and nontornadic outbreaks. *Mon. Wea. Rev.*, **140**, 2590–2608.

Mesinger, F., et al., 2006: North American regional reanalysis. *Bull. Amer. Meteor. Soc.*, **87**, 343–360, doi: 10.1175/BAMS-87-3-343.

Miller, J., R. Frederick, and R. Tracey, 1973: NOAA Atlas 2. *Precipitation-frequency atlas of the western United States*, **3**.

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 (529)**, 73–119.

Murphy, A. H., 1973: A new vector partition of the probability score. *J. Appl. Meteor.*, **12 (4)**, 595–600.

Murphy, A. H. and R. L. Winkler, 1977: Reliability of subjective probability forecasts of precipitation and temperature. *Appl. Statistics*, **26**, 41–47, doi: 10.2307/2346866.

Murphy, K. P., 2012: *Machine learning: a probabilistic perspective*. MIT press.

Nelson, B. R., O. P. Prat, D.-J. Seo, and E. Habib, 2016: Assessment and implications of NCEP Stage IV quantitative precipitation estimates for product intercomparisons. *Wea. Forecasting*, **31**, 371–394.

Neumann, C. J., 1972: *An alternate to the HURRAN (hurricane analog) tropical cyclone forecast system*. NOAA.

Nielsen, E. R., G. R. Herman, R. C. Tournay, J. M. Peters, and R. S. Schumacher, 2015: Double impact: When both tornadoes and flash floods threaten the same place at the same time. *Wea. Forecasting*, **30**, 1673–1693, doi: 10.1175/WAF-D-15-0084.1.

Nielsen, E. R. and R. S. Schumacher, 2016: Using convection-allowing ensembles to understand the predictability of an extreme rainfall event. *Mon. Wea. Rev.*, **144**, 3651–3676.

Nielsen, E. R. and R. S. Schumacher, 2018: Dynamical insights into extreme short-term precipitation associated with supercells and mesovortices. *J. Atmos. Sci.*, Submitted, doi: 10.1175/JAS-D-18-0385.1.

North, G. R., T. L. Bell, R. F. Cahalan, and F. J. Moeng, 1982: Sampling errors in the estimation of empirical orthogonal functions. *Mon. Wea. Rev.*, **110 (7)**, 699–706.

Novak, D. R., C. Bailey, K. F. Brill, P. Burke, W. A. Hogsett, R. Rausch, and M. Schichtel, 2014: Precipitation and temperature forecast performance at the Weather Prediction Center. *Wea. Forecasting*, **29**, 489–504.

Ntelekos, A. A., K. P. Georgakakos, and W. F. Krajewski, 2006: On the uncertainties of flash flood guidance: Toward probabilistic forecasting of flash floods. *J. Hydrometeor.*, **7**, 896–915.

NWS, 2012: Technical implementation notice 11-53. National Weather Service, Headquarters, [Available online at `http://www.nws.noaa.gov/os/notification/tin11-53ruc_rapaae.htm`].

NWS, 2017a: Service change notice 17-100. National Centers for Environmental Prediction, Weather Prediction Center, [Available online at http://www.nws.noaa.gov/os/notification/scn17-100wpc_excessive_rainfall.htm].

NWS, 2017b: Service change notice 17-100. National Centers for Environmental Prediction, Weather Prediction Center, [Available online at http://www.nws.noaa.gov/os/notification/scn17-100wpc_excessive_rainfall.htm].

NWS, 2017c: Summary of natural hazard statistics in the United States. National Weather Service, Office of Climate, Weather, and Water Services, [Available online at http://www.nws.noaa.gov/om/hazstats.shtml].

NWS, 2018: Summary of natural hazard statistics in the United States. National Weather Service, Office of Climate, Weather, and Water Services, [Available online at http://www.nws.noaa.gov/om/hazstats.shtml].

Ogden, F., H. Sharif, S. Senarath, J. Smith, M. Baeck, and J. Richardson, 2000: Hydrologic analysis of the Fort Collins, Colorado, flash flood of 1997. *J. Hydrology*, **228**, 82–100.

Orf, L., R. Wilhelmson, B. Lee, C. Finley, and A. Houston, 2017: Evolution of a long-track violent tornado within a simulated supercell. *Bull. Amer. Meteor. Soc.*, **98**, 45–68.

Parker, M. D., 2014: Composite VORTEX2 supercell environments from near-storm soundings. *Mon. Wea. Rev.*, **142**, 508–529.

Pedregosa, F., et al., 2011: Scikit-learn: Machine learning in Python. *J. Mach. Learn. Res*, **12**, 2825–2830.

Pelosi, A., H. Medina, J. Van den Bergh, S. Vannitsem, and G. B. Chirico, 2017: Adaptive kalman filtering for postprocessing ensemble numerical weather predictions. *Mon. Wea. Rev.*, **145 (12)**, 4837–4854.

Perica, S., S. Pavlovic, M. S. Laurent, C. Trypaluk, D. Unruh, D. Martin, and O. Wilhite, 2015: Precipitation-frequency atlas of the United States. *NOAA Atlas 14*, **10**.

Perica, S., et al., 2011: Precipitation-frequency atlas of the United States. *NOAA Atlas 14*, **6**.

Perica, S., et al., 2013: Precipitation-frequency atlas of the United States. *NOAA Atlas 14*, **9**.

Peters, J. M. and R. S. Schumacher, 2014: Objective categorization of heavy-rain-producing MCS synoptic types by rotated principal component analysis. *Mon. Wea. Rev.*, **142**, 1716–1737.

Pielke, R. A., M. W. Downton, and J. B. Miller, 2002: *Flood damage in the United States, 1926-2000: a reanalysis of National Weather Service estimates*. University Corporation for Atmospheric Research Boulder, CO.

Pinto, J. O., J. A. Grim, and M. Steiner, 2015: Assessment of the High-Resolution Rapid Refresh modelâĂŹs ability to predict mesoscale convective systems using object-based evaluation. *Wea. Forecasting*, **30 (4)**, 892–913.

Polger, P. D., B. S. Goldsmith, R. C. Przywarty, and J. R. Bocchieri, 1994: National Weather Service warning performance based on the WSR-88D. *Bull. Amer. Meteor. Soc.*, **75**, 203–214, doi:

10.1175/1520-0477(1994)075<0203:NWSWPB>2.0.CO;2.

Quinlan, J. R., 1986: Induction of decision trees. *Machine learning*, **1 (1)**, 81–106.

Ralph, F. M., P. J. Neiman, G. N. Kiladis, K. Weickmann, and D. W. Reynolds, 2010: A multi-scale observational case study of a Pacific atmospheric river exhibiting tropical-extratropical connections and a mesoscale frontal wave. *Mon. Wea. Rev.*, **139**, 1169–1189.

Ramsay, H. A. and C. A. Doswell, III, 2005: A sensitivity study of hodograph-based methods for estimating supercell motion. *Wea. Forecasting*, **20**, 954–970.

Ramshaw, J. D., 1985: Conservative rezoning algorithm for generalized two-dimensional meshes. *Journal of Computational Physics*, **59**, 193–199.

Reed, S., J. Schaake, and Z. Zhang, 2007: A distributed hydrologic model and threshold frequency-based method for flash flood forecasting at ungauged locations. *J. Hydrology*, **337**, 402–420.

Richardson, L. F., 2007: *Weather prediction by numerical process*. Cambridge University Press.

Richman, M. B., 1986: Rotation of principal components. *International Journal of Climatology*, **6**, 293–335.

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

Roebber, P. J., 2013: Using evolutionary programming to generate skillful extreme value probabilistic forecasts. *Mon. Wea. Rev.*, **141**, 3170–3185.

Roebber, P. J., 2015: Adaptive evolutionary programming. *Mon. Wea. Rev.*, **143 (5)**, 1497–1505.

Romps, D. M., 2017: Exact expression for the lifting condensation level. *J. Atmos. Sci.*, **74**, 3891–3900.

Ross, D. A., J. Lim, R.-S. Lin, and M.-H. Yang, 2008: Incremental learning for robust visual tracking. *International Journal of Computer Vision*, **77 (1-3)**, 125–141.

Rosten, E. and T. Drummond, 2006: Machine learning for high-speed corner detection. *Computer Vision–ECCV*, 430–443.

Rothfusz, L., C. Karstens, and D. Hilderbrand, 2014: Forecasting a continuum of environmental threats: Exploring next-generation forecasting of high impact weather. *Eos, Trans. Amer. Geophys. Union*, **95**, 325–326.

Rozas-Larraondo, P., I. Inza, and J. A. Lozano, 2014: A method for wind speed forecasting in airports based on nonparametric regression. *Wea. Forecasting*, **29**, 1332–1342.

Rutz, J. J., W. J. Steenburgh, and F. M. Ralph, 2014: Climatological characteristics of atmospheric rivers and their inland penetration over the western United States. *Mon. Wea. Rev.*, **142**, 905–921.

Scheuerer, M. and T. M. Hamill, 2015: Statistical postprocessing of ensemble precipitation forecasts by fitting censored, shifted gamma distributions*. *Mon. Wea. Rev.*, **143 (11)**, 4578–4596.

Schmidt, J. A., A. Anderson, and J. Paul, 2007: Spatially-variable, physically-derived flash flood guidance. *AMS 21st Conference on Hydrology, San Antonio, TX B*, Vol. 6.

Schneider, R. S. and A. R. Dean, 2008: A comprehensive 5-year severe storm environment climatology for the continental United States. *Preprints, 24th Conf. on Severe Local Storms, Savannah, GA, Amer. Meteor. Soc.*, 16A.4.

Schroeder, A. J., et al., 2016: The development of a flash flood severity index. *J. Hydrol.*, **541**, 523–532, doi: 10.1016/j.jhydrol.2016.04.005.

Schumacher, R. S., 2017: Heavy rainfall and flash flooding. Interactive Factory, URL http://naturalhazardscience.oxfordre.com/view/10.1093/acrefore/9780199389407.001.0001/acrefore-9780199389407-e-132.

Schumacher, R. S., A. J. Clark, M. Xue, and F. Kong, 2013: Factors influencing the development and maintenance of nocturnal heavy-rain-producing convective systems in a storm-scale ensemble. *Mon. Wea. Rev.*, **141 (8)**, 2778–2801.

Schumacher, R. S. and R. H. Johnson, 2005: Organization and environmental properties of extreme-rain-producing mesoscale convective systems. *Mon. Wea. Rev.*, **133**, 961–976.

Schumacher, R. S. and R. H. Johnson, 2006: Characteristics of US extreme rain events during 1999-2003. *Wea. Forecasting*, **21**, 69–85.

Schumacher, R. S. and R. H. Johnson, 2008: Mesoscale processes contributing to extreme rainfall in a midlatitude warm-season flash flood. *Mon. Wea. Rev.*, **136 (10)**, 3964–3986.

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

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

Shlens, J., 2014: A tutorial on principal component analysis. *arXiv preprint arXiv:1404.1100*.

Simmons, K. M. and D. Sutter, 2005: WSR-88D radar, tornado warnings, and tornado casualties. *Wea. Forecasting*, **20**, 301–310, doi: 10.1175/WAF857.1.

Simmons, K. M. and D. Sutter, 2008: Tornado warnings, lead times, and tornado casualties: An empirical investigation. *Wea. Forecasting*, **23**, 246–258, doi: 10.1175/2007WAF2006027.1.

Sloughter, J. M. L., A. E. Raftery, T. Gneiting, and C. Fraley, 2007: Probabilistic quantitative precipitation forecasting using Bayesian model averaging. *Mon. Wea. Rev.*, **135**, 3209–3220.

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

Smith, J. A., M. L. Baeck, Y. Zhang, and C. A. Doswell III, 2001: Extreme rainfall and flooding from supercell thunderstorms. *J. Hydrometeor.*, **2**, 469–489.

Smith, J. A., A. J. Miller, M. L. Baeck, P. A. Nelson, G. T. Fisher, and K. L. Meierdiercks, 2005: Extraordinary flood response of a small urban watershed to short-duration convective rainfall. *J. Hydrometeor.*, **6**, 599–617.

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, doi: 10.1175/WAF-D-10-05046.1.

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, doi: 10.1175/WAF-D-16-0073.1.

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, doi: 10.1175/WAF-D-15-0138.1.

SPC, cited 2017a: SPC convective outlooks. [Available online at http://www.spc.noaa.gov/cgi-bin-spc/getacrange.pl].

SPC, cited 2017b: SVRGIS (updated: 15 May 2017). [Available online at http://www.spc.noaa.gov/gis/svrgis/].

SPC, cited 2017c: Severe weather climatology (1982–2011). [Available online at http://www.spc.noaa.gov/new/SVRclimo/climo.php?parm=anySvr].

Stephenson, D., B. Casati, C. Ferro, and C. Wilson, 2008: The extreme dependency score: A non-vanishing measure for forecasts of rare events. *Meteorol. Appl.*, **15**, 41–50, doi: 10.1002/met.53.

Stevenson, S. N. and R. S. Schumacher, 2014: A 10-year survey of extreme rainfall events in the central and eastern United States using gridded multisensor precipitation analyses. *Mon. Wea. Rev.*, **142**, 3147–3162.

Stough, S., E. Leitman, J. Peters, and J. Correia Jr., 2010: The role of Storm Prediction Center products in decision making leading up to severe weather events. 14 pp., [Available online at http://www.spc.noaa.gov/publications/leitman/stough.pdf].

Strobl, C., A.-L. Boulesteix, T. Kneib, T. Augustin, and A. Zeileis, 2008: Conditional variable importance for random forests. *BMC bioinformatics*, **9**, 307.

Strobl, C., A.-L. Boulesteix, A. Zeileis, and T. Hothorn, 2007: Bias in random forest variable importance measures: Illustrations, sources and a solution. *BMC bioinformatics*, **8**, 25.

Surcel, M., I. Zawadzki, and M. Yau, 2016: The case-to-case variability of the predictability of precipitation by a storm-scale ensemble forecasting system. *Mon. Wea. Rev.*, **144**, 193–212, doi: 10.1175/MWR-D-15-0232.1.

Sweeney, T. L., 1992: Modernized areal flash flood guidance.

Taub, L., 2004: *Ancient meteorology*. Routledge.

Thompson, D. W. and J. M. Wallace, 1998: The Arctic Oscillation signature in the wintertime geopotential height and temperature fields. *Geophys. Res. Lett.*, **25**, 1297–1300.

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

Tippett, M. K., A. H. Sobel, and S. J. Camargo, 2012: Association of US tornado occurrence with monthly environmental parameters. *Geophys. Res. Lett.*, **39**, L02 801.

Trapp, R. J., D. M. Wheatley, N. T. Atkins, R. W. Przybylinski, and R. Wolf, 2006: Buyer beware: Some words of caution on the use of severe wind reports in postevent assessment and research. *Wea. Forecasting*, **21**, 408–415, doi: 10.1175/WAF925.1.

Vaughan, M. T., B. H. Tang, and L. F. Bosart, 2017: Climatology and analysis of high-impact, low predictive skill severe weather events in the northeast United States. *Wea. Forecasting*, **32**, 1903–1919, doi: 10.1175/WAF-D-17-0044.1.

Verbout, S. M., H. E. Brooks, L. M. Leslie, and D. M. Schultz, 2006: Evolution of the U.S. tornado database: 1954–2003. *Wea. Forecasting*, **21**, 86–93, doi: 10.1175/WAF910.1.

Verlinden, K. L. and D. R. Bright, 2017: Using the second-generation GEFS reforecasts to predict ceiling, visibility, and aviation flight category. *Wea. Forecasting*, **32**, 1765–1780.

Versini, P.-A., E. Gaume, and H. Andrieu, 2010: Application of a distributed hydrological model to the design of a road inundation warning system for flash flood prone areas. *Natural Hazards and Earth System Sciences*, **10**, 805.

Vescio, M. D. and R. L. Thompson, 2001: Subjective tornado probability forecasts in severe weather watches. *Wea. Forecasting*, **16**, 192–195, doi: 10.1175/1520-0434(2001)016<0192:FSFSTP>2.0.CO;2.

Villarini, G., W. F. Krajewski, A. A. Ntelekos, K. P. Georgakakos, and J. A. Smith, 2010: Towards probabilistic forecasting of flash floods: The combined effects of uncertainty in radar-rainfall and flash flood guidance. *J. Hydrol.*, **394**, 275–284.

Vislocky, R. L. and J. M. Fritsch, 1997: Performance of an advanced MOS system in the 1996-97 National Collegiate Weather Forecasting Contest. *Bull. Amer. Meteor. Soc.*, **78 (12)**, 2851.

Wang, S.-Y., T.-C. Chen, and S. E. Taylor, 2009: Evaluations of NAM forecasts on midtropospheric perturbation-induced convective storms over the US northern plains. *Wea. Forecasting*, **24**, 1309–1333.

Waters, K., et al., 2005: Polygon weather warnings: a new approach for the National Weather Service. *Extended Abstracts, 21st International Conference on Interactive Information Processing Systems (IIPS) for Meteorology, Oceanography, and Hydrology*, Phoenix, AZ, Amer. Meteor. Soc., [Available online at https://ams.confex.com/ams/Annual2005/techprogram/paper_86326.htm].

Wei, M., Z. Toth, R. Wobus, and Y. Zhu, 2008: Initial perturbations based on the ensemble transform (ET) technique in the NCEP global operational forecast system. *Tellus A*, **60**, 62–79.

Weijs, S. V. and N. Van De Giesen, 2011: Accounting for observational uncertainty in forecast verification: An information-theoretical view on forecasts, observations, and truth. *Mon. Wea. Rev.*, **139**, 2156–2162.

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.

Welles, E., S. Sorooshian, G. Carter, and B. Olsen, 2007: Hydrologic verification: A call for action and collaboration. *Bull. Amer. Meteor. Soc.*, **88**, 503–511.

Whan, K. and M. Schmeits, 2018: Comparing area-probability forecasts of (extreme) local precipitation using parametric and machine learning statistical post-processing methods. *Mon. Wea. Rev.*, **submitted**.

Wheeler, M. C. and H. H. Hendon, 2004: An all-season real-time multivariate MJO index: Development of an index for monitoring and prediction. *Mon. Wea. Rev.*, **132**, 1917–1932.

Wick, G. A., P. J. Neiman, F. M. Ralph, and T. M. Hamill, 2013: Evaluation of forecasts of the water vapor signature of atmospheric rivers in operational numerical weather prediction models. *Wea. Forecasting*, **28**, 1337–1352.

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

Williams, J. K., 2014: Using random forests to diagnose aviation turbulence. *Machine learning*, **95**, 51–70.

Willis, E. P. and W. H. Hooke, 2006: Cleveland abbe and american meteorology, 1871–1901. *Bull. Amer. Meteor. Soc.*, **87**, 315–326.

Wilson, J. W. and R. D. Roberts, 2006: Summary of convective storm initiation and evolution during IHOP: Observational and modeling perspective. *Mon. Wea. Rev.*, **134**, 23–47.

Wolff, A., 2013: Simulation of pavement surface runoff using the depth-averaged shallow water equations. Ph.D. thesis, University of Stuttgart.

Wolter, K., J. K. Eischeid, L. Cheng, and M. Hoerling, 2016: What history tells us about 2015 us daily rainfall extremes. *Bull. Amer. Meteor. Soc.*, **97 (12)**, S9–S13.

Wood, E. F., 1976: An analysis of the effects of parameter uncertainty in deterministic hydrologic models. *Water Resources Research*, **12**, 925–932.

Yatheendradas, S., T. Wagener, H. Gupta, C. Unkrich, D. Goodrich, M. Schaffner, and A. Stewart, 2008: Understanding uncertainty in distributed flash flood forecasting for semiarid regions. *Water Resources Research*, **44**.

Zeng, J. and W. Qiao, 2011: Support vector machine-based short-term wind power forecasting. *Power Systems Conference and Exposition (PSCE), 2011 IEEE/PES*, IEEE, 1–8, doi:

10.1109/PSCE.2011.5772573, URL https://ieeexplore.ieee.org/abstract/document/5772573/.

Zhang, F., N. Bei, R. Rotunno, C. Snyder, and C. C. Epifanio, 2007: Mesoscale predictability of moist baroclinic waves: Convection-permitting experiments and multistage error growth dynamics. *J. Atmos. Sci.*, **64**, 3579–3594.

Zhang, F., C. Snyder, and R. Rotunno, 2003: Effects of moist convection on mesoscale predictability. *J. Atmos. Sci.*, **60**, 1173–1185, doi: 10.1175/1520-0469(2003)060<1173:EOMCOM>2.0.CO;2.

Zhang, J., et al., 2016: Multi-Radar Multi-Sensor (MRMS) quantitative precipitation estimation: Initial operating capabilities. *Bull. Amer. Meteor. Soc.*, **97**, 621–638.

Zou, H., T. Hastie, and R. Tibshirani, 2006: Sparse principal component analysis. *Journal of computational and graphical statistics*, **15**, 265–286.

Acronym	Full Name
AI	Accumulation interval
AMS	American meteorological society
APCP	Accumulated precipitation
ARI	Average recurrence interval
BS	Brier score
BSS	Brier skill score
CAPE	Convective available potential energy
CCPA	Climatology Calibrated Precipitation Analysis
CIN	Convective inhibition
CONUS	Conterminous United States
CSI	Critical success index
CWA	County warning area
ECMWF	European Centre for Medium-Range Weather Forecasts
ETS	Equitable threat score
FB	Frequency bias
FFG	Flash flood guidance
FFR	Flash flood report
FFW	Flash flood warning
FT	Fixed threshold
GEFS/R	Second Generation Global Ensemble Forecast System Reforecast
GIS	Geographic information system
HRAP	Hydrologic Rainfall Analysis Project
IEM	Iowa Environmental Mesonet
LCL	Lifted condensation level
LR	Logistic regression
LSR	Local storm report
ML	Machine learning
MRMS	Multi-Radar Multi-Sensor QPE Analysis

MSLP	Mean sea-level pressure
NCEP	National Centers for Environmental Prediction
NWP	Numerical weather prediction
NWS	National Weather Service
PC	Principal component
PCA	Principal component analysis
PD	Performance diagram
POD	Probability of detection
PWAT	Precipitable water
Q	Specific humidity
QPE	Quantitative precipitation estimate
QPF	Quantitative precipitation forecast
RF	Random Forest
RFC	River forecast center
RH	Relative humidity
RPSS	Rank probability skill score
RP SPC	Storm Prediction Center
SPP	Statistical post-processing
SR	Success ratio
SRH	Storm relative helicity
ST4	Stage IV Precipitation Analysis
T	Temperature
U	Zonal component of wind
V	Meridional component of wind
WFO	Weather forecast office
WPC	Weather Prediction Center

TABLE A1. List of all acronyms or abbreviations used in this dissertation and their spelled out meanings.