UNIVERSITY OF OKLAHOMA

GRADUATE COLLEGE

MACHINE LEARNING CO-PRODUCTION IN OPERATIONAL METEOROLOGY

A DISSERTATION

SUBMITTED TO THE GRADUATE FACULTY

in partial fulfillment of the requirements for the

Degree of

DOCTOR OF PHILOSOPHY

By

DAVID HARRISON
Norman, Oklahoma
2022

Downloaded from SHAREOK. Persistent Handle URL: https://hdl.handle.net/11244/335971

MACHINE LEARNING CO-PRODUCTION IN OPERATIONAL METEOROLOGY

A DISSERTATION APPROVED FOR THE SCHOOL OF METEOROLOGY

BY THE COMMITTEE CONSISTING OF

Dr. Amy McGovern, Chair

Dr. Christopher Karstens, Co-Chair

Dr. Caleb Fulton

Dr. Jason Furtado

Dr. Jeffrey Basara

Dr. Michael Richman

© Copyright by DAVID HARRISON 2022
All Rights Reserved.

Acknowledgements

This dissertation would not have been possible without the help and guidance of my many friends and colleagues, and especially that of my co-advisers, Dr. Amy McGovern and Dr. Chris Karstens. Amy has guided my research since freshman year and unwaveringly supported my many endeavors through the intersection of computer science and meteorology. Chris was my first point of contact within the professional meteorological community and has been an invaluable source of guidance and support for both my research and the development of my professional career. I cannot thank my co-advisers enough for all the help they have given me, and I look forward to working with them on future ventures. I also wish to thank my entire committee for their time and suggestions, as well as Dr. Israel Jirak, Dr. Patrick Marsh, Matt Elliott, Bill Bunting, Dr. Russ Schneider, Dr. Burkley Gallo, Dr. David Jahn, Dr. Kenzie Krocak, Dr. Julie Demuth, and Dr. Ann Bostrom for their exceptional guidance and insight. I offer a special thank you to the forecasters at SPC for allowing me to observe their operational procedures, for providing invaluable feedback on many, many prototypes, and for giving my work a chance to make a difference in the operational community. My fellow friends and members of the OU IDEA Lab were essential for feedback on this research and frequently provided me with new ideas to improve this dissertation. In addition, I thank the many NWS forecasters, researchers, and academic members who participated in and helped facilitate the 2021 and 2022 HWT Spring Forecasting Experiments for their diligent efforts and insightful discussions, without which this research would not be possible. Finally, I must give special thanks to my parents, grandmother, and uncle. I never could have made it this far without their abundant love and support.

This dissertation was prepared in part with funding provided by NOAA/Office of Oceanic and Atmospheric Research under NOAA-University of Oklahoma Cooperative Agreements #NA16OAR4320115 and #NA21OAR4320204, U.S. Department of Commerce. Some research included herein was performed under University of Oklahoma IRB approval #13320. The scientific results and conclusions, as well as any views or opinions expressed herein, are those of the author and do not necessarily reflect the views of NOAA or the Department of Commerce.

Table of Contents

Acknowledgements **iv**

List Of Tables **viii**

List Of Figures **ix**

Abstract **xiv**

1 Introduction **1**

2 Background **10**
2.1 Machine Learning Overview 10
2.1.1 Linear Regression 12
2.1.2 Random Forests 13
2.1.3 Gradient Boosted Forests 13
2.1.4 Isotonic Regression 14
2.2 Model Evaluation and Verification 15

3 How Forecasters Evaluate Machine Learning **24**
3.1 The 2021 Spring Forecasting Experiment 24
3.2 Survey Design 26
3.3 Results and Discussion 30
3.3.1 Survey Demographics 30
3.3.2 Machine Learning vs. Traditional Forecast Products . . . 32
3.3.3 Researcher vs. Forecaster Perspectives 43
3.3.4 Collaboration in Product Development 51

4 Collaborative Co-Production **54**
4.1 Initiating Phase 56
4.2 Design Phase 58
4.3 Production Phase 60
4.4 Distribution Phase 61
4.5 Evaluation Phase 62

5 Lessons from an R2O Success **64**
5.1 Background 64
5.2 Data and Methods 69

5.2.1 The HREFCT Algorithm . . . 71
5.2.2 Calibration . . . 75
5.2.3 Instability and Reflectivity Mask . . . 78
5.3 Results and Discussion . . . 80
5.4 Operational Implementation . . . 91

6 Co-Production of a First-Guess Convective Watch Product 95
6.1 HREF-based ML guidance . . . 101
6.1.1 Feature Engineering . . . 103
6.1.2 Model Design . . . 109
6.1.3 Deriving First-Guess County-Based Watches . . . 113
6.2 SPC Severe Timing Guidance . . . 118
6.3 Results and Discussion . . . 123
6.3.1 Comparison to SPC Watches . . . 125
6.3.2 Capturing the Severe Weather Threat . . . 132

7 Results from the 2022 Hazardous Weather Testbed 142
7.1 Testbed Design . . . 143
7.2 Participant Evaluation . . . 147
7.3 Future Work . . . 154

8 Conclusions 156

Appendix A 2021 SFE Survey 161

Appendix B 2022 SFE Survey 165

Reference List 168

List Of Tables

2.1 Example of a binary contingency table. 18
2.2 Common verification metrics derived from a binary contingency table. 19

3.1 The list of factors survey respondents were asked to consider when evaluating what variables influence their decision to implement a new product in their personal forecast process. 28
3.2 The number and percent of survey respondents in each professional background. Participants who selected multiple backgrounds are counted in each associated category. 30
3.3 Bootstrapped mean importance scores μ for generic probabilistic forecast products (Q4), ML-derived probabilistic forecast products (Q7), and the differences between the two. 41
3.4 As in Table 3.3, but for the bootstrapped standard deviation (σ). 42
3.5 Bootstrapped mean importance scores for generic probabilistic forecast products (Q4) and ML-derived probabilistic forecast products (Q7) as rated by operational forecasters (μ_F) and researchers (μ_R). 50

5.1 HREF prognostic fields with the greatest ensemble mean Pearson correlation to 1-hour NLDN CG lightning flashes computed between 1 July 2017 and 1 July 2019. 72
5.2 The best thresholds (t) and weights (w) for each HREF prognostic field and forecast time interval. MU LI was excluded from the 24-hour forecast due to strong diurnal variations in the parameter. 74

6.1 Derived storm-scale and environmental fields, their optimal exceedance thresholds, and spatial masks that make up the training dataset. 106
6.2 Example of input values within the training dataset. Note that the "Class" field was removed prior to training and is only provided here for reference. 110
6.3 Optimally-tuned hyperparameters used to train a scikit-learn GBC (https://scikit-learn.org). 112
6.4 From Jirak et al. (2020): "Probabilistic inputs from the HREF and SREF to the calibrated hazard guidance." 120

List Of Figures

1.1 From Chase et al. (2022), their Fig. 1. "Clarivate Web of Science abstract results for machine learning and severe weather topics in meteorology and atmospheric sciences. Machine learning keywords searched were: linear regression, logistic regression, decision trees, random forest, gradient boosted trees, support vector machines, k-means, knearest, empirical orthogonal functions, principal component analysis and self organizing maps. Severe weather keywords searched were: tornadoes, hail, hurricanes and tropical cyclones. (a) Number of publications per year in the Meteorology/Atmospheric Science category and in the machine learning and severe weather subsets. (b) Number of machine learning and severe weather publications normalized by the total number of Meteorology/Atmospheric Science publications." . . . 5

2.1 (a) Example of a combined ROC/reliability diagram for a probabilistic forecast. The dashed line represents the line of no skill for the ROC curve and the one-to-one line for the reliability plot. This case demonstrates a forecast with good potential usefulness but poor calibration. (b) Example of a performance diagram for a probabilistic forecast. See Chapter 6 for more information about the data plotted. 20

3.1 (a) How often survey respondents utilize probabilistic forecast products as part of their typical work responsibilities. (b) How knowledgeable survey respondents are ML, including random forests, DL, and other AI techniques. 31

3.2 Mean relative importance when evaluating a generic probabilistic forecast product. Error bars represent the 95% confidence interval from 10,000 bootstrapped samples. 33

3.3 Survey Q4 responses approximated as KDE curves. Dashed vertical lines represent the mean score of each factor. Red fill represents factors with bootstrapped mean scores ≥ 4.0, yellow is used for mean scores between 3.0 - 4.0, and green indicates mean scores < 3.0. 35

3.4 Mean relative importance when evaluating an ML probabilistic forecast product. Error bars represent the 95% confidence interval from 10,000 bootstrapped samples. 37

3.5 Survey Q7 responses approximated as KDE curves. Dashed vertical lines represent the mean score of each factor. Red fill represents factors with bootstrapped mean scores $\geq$ 4.0, yellow is used for mean scores between 3.0 - 4.0, and green indicates mean scores $<$ 3.0. 38
3.6 Survey Q4 forecaster and researcher responses approximated as KDE curves. Dashed vertical lines represent the mean score of each factor. 45
3.7 Survey Q7 forecaster and researcher responses approximated as KDE curves. Dashed vertical lines represent the mean score of each factor. 46
3.8 Survey Q9 responses approximated as KDE curves. Dashed vertical lines represent the mean importance of researcher/forecaster collaboration during each phase of production. Red fill represents factors with bootstrapped mean scores $\geq$ 4.0 and yellow indicates mean scores $<$ 4.0. 52

4.1 Schematic of the collaborative co-production process. 57

5.1 (a) Mean uncalibrated reliability error of the 12z HREFCT 4-hour lightning probabilities at forecast hour 16 for the 40% probability bin. Positive values represent an over-forecast compared to NLDN observations from 13 June 2019 - 13 June 2020. (b) Mean uncalibrated reliability error of the 12z HREFCT as a function of lead time. 77
5.2 (a) HREFCT (a) 4-hour and (b) 24-hour forecasts from the 12z HREF cycle on 17 March 2021. Yellow "+" symbols indicate grid points where there was at least one CG lightning flash detected during the valid forecast period. 78
5.3 12z HREFCT 1-hour, 4-hour, and 24-hour (a) mean performance, (b) mean reliability, and (c) forecast probability frequency for 20200613 - 20210511. The shaded regions represent the 95% confidence intervals from 10,000 bootstrapped samples. 82
5.4 12z HREFCT mean spatial reliability error across six 4-hour periods. Positive values (warmer colors) represent an over-forecast and negative values (cooler colors) represent an under-forecast. The reliability error was calculated for the 40% probability bin from 20200613 - 20210511. 84
5.5 A comparison of 09z and 15z SREFCT and 12z HREFCT 4-hour (a) mean performance, (b) mean reliability, and (c) forecast probability frequency for 20200613 - 20210511. The shaded regions represent the 95% confidence intervals from 10,000 bootstrapped samples. 87

5.6 A comparison of 09z and 15z SREFCT and 12z HREFCT 4-hour (a) mean performance and (b) mean reliability error as a function of lead time between 20200613 - 20210511. The shaded regions represent the 95% confidence intervals from 1000 bootstrapped samples. Forecast lead time increases to the right for both plots. 89
5.7 (a) 12z HREFCT and (b) 09z SREFCT 4-hour calibrated thunder forecasts for 20200412 12z - 16z. Yellow "+" symbols indicate grid points where there was at least one CG lightning flash detected during the valid forecast period. 90
5.8 (a) An interactive web interface designed to easily compare HREFCT prototypes and operational SREFCT forecasts while also displaying near-real time verification. (b) HREFCT products integrated into the operational NAWIPS software. 93

6.1 Example of an (a) MCD and (b) Tornado Watch as issued by SPC on 20 May 2019. The MCD was issued at 1617z and the Tornado Watch went into effect at 1835z. 97
6.2 Verification of Tornado Watch 123 on (a) 13 April 2022 19z, (b) 13 April 2022 22z, (c) 14 April 2022 00z, and (d) 14 April 2022 02z. Counties with red outlines but no fill represent counties that were cleared from the original watch. Adjacent watches are not shown for clarity. 100
6.3 (a) Example of grid point sampling within the SPC D1 MRGL risk on 20 May 2019. Red dots indicate positive class (watch) samples while black dots represent negative (no watch) samples. (b) Grid point sampling of NMEP UH values $> 99.85\%$ of climatology on 20 May 2019. (c) Spatial distribution of sampled Tornado Watch grid points from 10 March 2018 - 31 May 2022. (d) As in (c) but for Severe Thunderstorm Watches. 108
6.4 (a) Distribution of SPC Convective Outlooks with a watch from 2009 - 2019. (b) Distribution of Tornado and Severe Thunderstorm Watches per modern SPC Convective Outlook category from 2009 - 2019. 114
6.5 GBC (a) mean performance and (b) mean reliability for 20200310 - 20210310 when not masked, masked by an SPC MRGL risk, or masked by an SPC SLGT risk area. Shaded regions represent the 95% confidence intervals from 10,000 bootstrapped samples. 116
6.6 (a) Forecast watch probabilities and (b) derived first-guess county-based watches for 20200520 23z. The blue polygons represent operational SPC Severe Thunderstorm Watch parallelograms valid for this hour. 117

6.7 (a) SPC Severe Timing Guidance categorical forecast and (b) derived first-guess county-based watches for 20220413 22z. The red and blue polygons represent valid operational SPC Tornado and Severe Thunderstorm Watch parallelograms respectively. . . 122
6.8 Mean performance of the 12z HREF-based ML and 13z SPC Severe Timing Guidance deterministic, first-guess, county-based watch predictions for 20 March 2021 - 31 May 2022. Shaded regions denote 95% confidence intervals from 10,000 bootstrapped samples. 127
6.9 A comparison of 12z HREF-based ML and 13z Severe Timing Guidance mean conditional performance as a function of lead time between 20 March 2021 - 31 May 2022. The shaded regions represent the 95% confidence intervals from 10,000 bootstrapped samples. Forecast lead time increases to the right. 129
6.10 (a) 12z HREF-based ML and (b) 13z SPC Severe Timing Guidance FSS as a function of horizontal and temporal scales for 20 March 2021 - 31 May 2022. The shaded regions represent the 95% confidence intervals from 10,000 bootstrapped samples. . . 131
6.11 Mean performance of the 12z HREF-based ML, 13z Severe Timing Guidance, and operational SPC watch products evaluated against local storm reports for 20 March 2021 - 31 May 2022. The shaded regions represent the 95% confidence intervals from 10,000 bootstrapped samples. 134
6.12 12z HREF-based ML, 13z Severe Timing Guidance, and operational SPC watch product POD as a function of lead time for (a) all reports, (b) tornado reports, (c) wind reports, and (d) hail reports for 20 March 2021 - 31 May 2022. 136
6.13 12z HREF-based ML, 13z Severe Timing Guidance, and operational SPC watch product POD as a function of lead time for (a) all warnings, (b) TOR warnings, (c) SVR warnings for 20 March 2021 - 31 May 2022. 139

7.1 Web display presented to 2022 SFE participants while evaluating the performance of the 12z HREF-based ML and 13z SPC Severe Timing Guidance first-guess watch products. Blue polygons represent SVR warnings valid at the displayed hour, blue squares indicate wind LSRs, and green circles represent hail LSRs. . . . 145
7.2 (a) Survey Q3 and (b) Q4 responses approximated as KDE curves. Dashed vertical lines represent the mean score for each guidance product. 149

7.3 Mean performance of the 12z HREF-based ML and 13z Severe Timing Guidance watch products compared to (a) operational SPC watches and (b) LSRs during the 2022 Spring Forecasting Experiment. 151

Abstract

Machine learning, deep learning, and other artificial intelligence (AI) methods are becoming popular tools within the meteorological research community. However, despite the breadth of promising AI research and its increasing adoption within operational agencies, expert forecasters are often hesitant to fully embrace this relatively new technology. Operational forecasters have a practiced, expert insight into weather analysis and forecasting but typically lack the time, resources, or guidance for formal research and development due to the daily demands of their jobs. Conversely, many researchers have the resources, theoretical knowledge, and formal experience to solve complex meteorological challenges but may lack a full understanding of operation procedures, needs, requirements, and authority necessary to effectively bridge the research to operations (R2O) gap. To address these challenges and attempt to improve the R2O success of AI-derived products, this research investigates how operational forecasters evaluate new forecast guidance and how their perspectives about the R2O process differ from those of the research community. The results from these investigations are then used to derive a collaborative co-production framework intended to optimize the R2O process while improving researcher-forecaster communication throughout the development cycle. Finally, the benefit of this collaborative co-production framework is demonstrated by applying modern AI techniques in tandem with the expert knowledge of Storm Prediction Center forecasters to develop two new forecast products designed to predict lightning hazards and emulate county-based Severe Thunderstorm and Tornado Watches that dynamically evolve with the predicted time and location of the severe weather threat.

Chapter 1

Introduction

The phrase "research to operations" (R2O) is widely used within the meteorological community to describe the transfer of new ideas and technologies into an operational working environment. Within the academic and public research sectors, this often means the transition of a new product, algorithm, or technique from a prototype or experimental research phase into a form that is routinely used and supported operationally by the National Oceanic and Atmospheric Administration (NOAA) or one of its line offices such as the National Weather Service (NWS). To aid in R2O transitions, NOAA has implemented a hierarchy of nine Readiness Levels (RLs; NOAA 2022) intended to provide a consistent, systematic assessment of the maturity of ongoing research and development. In this hierarchy, RL 1 represents a project in the basic or theoretical research stage, while a project in RL 9 has been deployed and is used routinely in operations. These RLs often serve as a standardized project template for those in the research community and provide iterative milestones for funding agencies to assess development progress.

The benefits of collaboration between the research and operational communities during the R2O process have long been documented in the scientific literature. Operational forecasters have a practiced, expert insight into weather analysis and forecasting but typically lack the time, resources, or guidance for formal research and development (Doswell 1986; Auciello and Lavoie 1993; Kain

et al. 2003). Conversely, many researchers have the resources, theoretical knowledge, and formal experience to solve complex meteorological challenges but lack an understanding of operation procedures, needs, requirements, and authority necessary to effectively bridge the R2O gap (Auciello and Lavoie 1993; Serafin et al. 2002). Sustained collaboration between researchers and operational forecasters, then, serves as the most viable strategy to bridge this gap while offering the potential to further a better understanding and improved prediction of atmospheric processes (Kain et al. 2003) via ongoing multi-disciplinary knowledge transfer between the research and operational communities. However, despite these apparent benefits, the R2O process is often perceived to be a unidirectional interface between the research and operational communities. For example, researchers abiding by the NOAA RL milestones may not necessarily interact with their intended end users and stakeholders (e.g., operational forecasters) until late in the development process. Indeed, the first milestone that explicitly requires interaction with potential end users is RL 6, at which point the project is expected to demonstrate a functioning prototype in a formal testbed or other relevant environment. This structure potentially disincentivizes two-way communication between researchers and their end users during earlier stages of development and unintentionally limits opportunity for collaboration. Such lack of collaboration during the development process may come as a detriment to the value, usability, and adoption of the final product in an operational environment (Doswell et al. 1981; Auciello and Lavoie 1993; Serafin et al. 2002; Kain et al. 2003).

Deal and Hoffman (2010a) chronicle the many challenges facing researchers and operational forecasters as they navigate the R2O process. New technologies and products proposed for operational implementation must derive from the

forefront of modern research and be relevant to the immediate needs of the end users. However, development, testing, and buy-in by operational agencies can take years - sometimes decades - to complete, by which point the new technology may no longer be new or relevant. To further complicate matters, stakeholder needs and requirements are vulnerable to change throughout the development process as the operational context, expertise, and managerial priorities evolve. These challenges can leave researchers trying to meet a moving target that ultimately increases the complexity, time, and cost of development. Frese and Sauter (2003) and Hoffman et al. (2009) identify several elements common to successful R2O transitions, including accommodation to changing requirements, management buy-in, and communication among executives, managers, developers, suppliers, and end users. While these recommendations are valuable as general guidelines, Deal and Hoffman (2010a) argue that unstructured communication alone may not be sufficient to completely solve the challenges inherent with designing products for operations. Researchers should instead actively engage their end users in detailed collaboration to learn their operational needs, desires, and procedures. Additionally, end users who more actively engage in the development process may be more aware of the assumptions and limitations of the new tools and technologies being produced. Per Deal and Hoffman (2010a), successful R2O transitions "tend to be those in which the technology developers had a deep understanding of the nature of the user's work." To formalize this collaborative process, Hoffman et al. (2010) and Deal and Hoffman (2010b) introduce and demonstrate a cyclic model for collaborative co-development which they call the Practitioner's Cycles. This model and the concept of collaborative co-development is discussed further in Chapter 4.

The consequences of limited researcher-forecaster collaboration are perhaps most apparent in the meteorology community's recent advances in machine learning research. The ever-increasing volume and quality of data from high-resolution numerical weather prediction (NWP) models, ground-based observational systems, and Earth-orbiting satellites has made the field of meteorology an ideal target for the application of artificial intelligence research (McGovern et al. 2017). Indeed, machine learning (ML), deep learning (DL), and other artificial intelligence (AI) methods are becoming more popular tools among the meteorological research community. There are numerous studies in the literature (e.g., Gagne et al. 2014, 2017; Lagerquist et al. 2017, 2019; McGovern et al. 2017; Burke et al. 2020; Hill et al. 2020; Loken et al. 2020; Zhou et al. 2020; Shield and Houston 2022; van Straaten et al. 2022; Yang et al. 2022) that showcase the potential of AI techniques for nowcasting and forecasting severe and high-impact weather. Others have developed ML and DL methods to automate the detection of meteorological features such as synoptic-scale fronts (Lagerquist et al. 2019; Justin et al. 2022; Niebler et al. 2022), convection (Haberlie and Ashley 2018a,b; Cintineo et al. 2020b), atmospheric rivers (Chapman et al. 2019; Muszynski et al. 2019), and extratropical cyclones (Kumler-Bonfanti et al. 2020) from ground- and space-based observational systems. As of this writing, the number of formal meteorology publications that mention machine learning, deep learning, or artificial intelligence is increasing at a rate greater than that of most traditional meteorology topics. In particular, there has been recent exponential growth in the number of papers utilizing tree-based or (convolutional) neural network ML techniques, and this trend is expected to continue for the foreseeable future (Fig. 1.1; Chase et al. 2022).

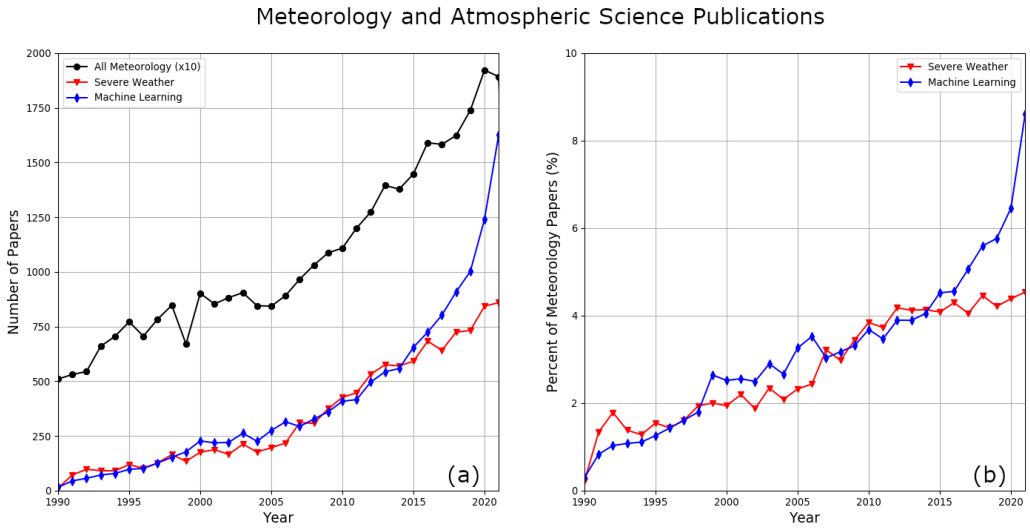


Figure 1.1: From Chase et al. (2022), their Fig. 1. "Clarivate Web of Science abstract results for machine learning and severe weather topics in meteorology and atmospheric sciences. Machine learning keywords searched were: linear regression, logistic regression, decision trees, random forest, gradient boosted trees, support vector machines, k-means, knearest, empirical orthogonal functions, principal component analysis and self organizing maps. Severe weather keywords searched were: tornadoes, hail, hurricanes and tropical cyclones. (a) Number of publications per year in the Meteorology/Atmospheric Science category and in the machine learning and severe weather subsets. (b) Number of machine learning and severe weather publications normalized by the total number of Meteorology/Atmospheric Science publications."

This saturation of AI applications within modern meteorological research has not gone unnoticed by the public-sector operational community. In 2020, NOAA published the *NOAA Artificial Intelligence Strategy* (NOAA 2020) which lays out a series of specific goals to dramatically expand the application of AI in every NOAA mission area. The strategy proposes to accomplish this by improving the efficiency, effectiveness, and coordination of AI development and usage across the agency. As part of this strategy, a variety of experimental ML-derived nowcast and forecast products have been evaluated by NWS forecasters within testbed settings (e.g., Calhoun et al. 2021; Clark et al. 2021). Many of these products, such as Hill et al. (2020), Loken et al. (2020), and Schumacher et al. (2021), have even progressed beyond testbed evaluation and are now being experimentally assessed and operationally implemented at NWS weather forecast offices (WFOs) and national centers. However, despite NOAA's recent emphasis on AI, the breadth of promising AI research, and its increasing adoption within operational agencies, expert forecasters are often hesitant to fully embrace this relatively new technology (McGovern et al. 2019, 2022).

The apparent difficulty ML research has navigating the R2O process is often attributed to a perceived inherent distrust that forecasters and other domain experts have of "black boxes" (Ding 2018; Sejnowski 2018; McGovern et al. 2019, 2020). Most modern ML models learn to solve classification or regression tasks by training and optimizing a potentially complex combination of mathematical weights, thresholds, and nonlinear cost functions. As such, it is often difficult to determine how these models reach a solution from their given input or if the relationships learned by the models are physically realistic. This general lack of transparency and interpretability has earned ML models the

dubious association with seemingly magic "black boxes." Anecdotal observations, informal conversations, and formal research (e.g., Hoffman et al. 2013, 2017; Harrison 2018; Karstens et al. 2018) support the notion that operational forecasters are hesitant to trust output from an automated model or algorithm that they don't understand or lack familiarity with. As a result, there has been a recent surge of formal study within the computer science and meteorological research communities to attempt to make the inner workings of ML more transparent and physically interpretable (e.g., Herman and Schumacher 2016; Olah et al. 2017; Lipton 2018; McGovern et al. 2019, 2020; Molnar 2020; Toms et al. 2020). Although these efforts have been shown to generally improve forecaster trust and may help ease the transition into operations (Cains et al. 2022), the interpretation and explanation of ML products typically occurs near the end of the development cycle when training potential end users. These methods alone do not necessarily increase communication or collaboration between researchers and forecasters during the development process, but merely improve the forecasters' understanding of the product they are being asked to use. As such, the application of ML interpretation during the development cycle, while important, *might be indicative of symptoms of a larger problem within the meteorological research community.*

Consider what it truly means for ML models to be regarded as "black boxes" and how that perception compares to more traditional products or "anchors" utilized within the operational community. Do all operational forecasters understand the inner workings of NWP used in everyday forecasting procedures? Are they familiar with the data assimilation techniques, initial and boundary conditions, microphysics schemes, and dynamical cores of every member of the convection-allowing ensembles that help guide their short-term forecast

decisions? In the context of public-sector operations, nonlinear physical and statistical models of any type are difficult to interpret and could be considered "black boxes" in the same way as ML models (Herman and Schumacher 2016; Boukabara et al. 2019). This discrepancy warrants deeper investigation to understand the dilemma impeding ML success in the R2O process. Does the operational community truly evaluate ML products with more scrutiny, or are there deeper concerns perhaps resulting from insufficient researcher-forecaster collaboration during the development cycle?

I hypothesize that many AI-driven products struggle to transition to the operational sector at least in part because the new products fail to sufficiently meet the needs and requirements of their intended end users. Furthermore, I hypothesize that actively increasing communication and collaboration between researchers and forecasters will improve the R2O success of new ML technologies. To assess these hypotheses, this dissertation investigates how operational forecasters evaluate ML guidance and how their perspectives about the R2O process differ from those of the research community (Chapter 3). The results from this investigation are then used to derive a collaborative co-production framework based on Hoffman et al. (2010)'s Practitioner's Cycles to optimize the R2O process while improving researcher-forecaster communication throughout the development process (Chapter 4). The potential benefit of this collaborative co-production is revealed by detailing the design and successful operational implementation of a new linear regression lightning forecast product at the National Weather Service's Storm Prediction Center (SPC; Chapter 5). Additionally, collaborative co-production principles are demonstrated by applying modern ML techniques in tandem with the expert domain knowledge of SPC forecasters to develop a new forecast product designed to emulate county-based

Severe Thunderstorm and Tornado Watches that dynamically evolve with the predicted time and location of the severe weather threat (Chapter 6). Finally, this new watch forecast guidance was presented and tested during the 2022 Hazardous Weather Testbed Spring Forecasting Experiment, and the results from that experiment are provided (Chapter 7).

Chapter 2

Background

The research presented in this dissertation largely focuses on the collaborative application of machine learning techniques to solve complex meteorological problems. However, ML is a broad field of study and includes a large variety of methods such as random forests, convolutional neural networks, and other AI techniques. To simplify future discussion, this chapter introduces the basic concepts of ML, the ML methods utilized in this dissertation, and common metrics for evaluating ML performance.

2.1 Machine Learning Overview

The origins of ML can be traced back to the 1950s when Arthur Samuel formally defined ML as a field of study that provides learning capability to computers without being explicitly programmed (Samuel 1959; Alzubi et al. 2018). At the fundamental level, these computational algorithms are designed to emulate or surpass human intelligence by extracting information from large, multidimensional datasets, using that information to derive complex relationships, and generalizing those relationships to unseen tasks. ML modeling techniques have proven particularly successful in environmental and atmospheric sciences where the complexity of data often does not align with the idealized assumptions required by more traditional statistical methods (Breiman 2001b).

Most modern ML algorithms fall under the category of supervised or unsupervised learning. Supervised learning methods rely on the user to manually oversee many aspects of the learning process, including feature selection, training criteria, and verification methods (Russell and Norvig 2010; Mercer et al. 2021). These types of algorithms are typically supplied with a collection of curated input features that correspond to a predefined set of labels or solutions. Supervised models then learn and optimize relationships between the input features and solutions to maximize predictive performance on an independent dataset. Conversely, unsupervised learning methods are not provided with a predefined set of solutions, but rather learn patterns in the training data to estimate their own solutions. Self-organizing maps (Nowotarski and Jensen 2013), k-means clustering (Wilks 2011), and kernel principal component analysis (Schölkopf et al. 1998) are a few examples of common unsupervised learning algorithms. The research presented in this dissertation primarily utilizes supervised learning techniques, and so the following discussion will omit unsupervised methods.

Supervised ML techniques can be further subdivided into classification and regression models. As the name suggests, classification models are designed to estimate the probabilities that a given sample belongs to one or more predefined classes. In scenarios where a deterministic solution is desired, classification models can be optimized or tuned to convert these probabilities into a single class prediction. Regression models, on the other hand, generally output solutions that fall within a continuous range of numerical values. For example, a classification model might be trained to predict if it is going to rain (a binary yes/no solution) while a regression model could be used to predict how much rain will fall (a continuous range of values). Many supervised ML algorithms

can be applied to both regression and classification tasks (Géron 2017), including linear regression, random forest, and gradient boosting models which are applied later in this dissertation. These ML algorithms are briefly described in the following subsections.

2.1.1 Linear Regression

Linear regression is a method that falls at the intersection of traditional statistical analysis and modern ML modeling. The most basic form of linear regression defines the dependence of a predictand to one or more predictor variables (Maulud and Abdulazeez 2020) via the equation

$$\hat{y} = \sum_{i=0}^{n} w_i x_i \tag{2.1}$$

where w_i is a learned weight, x_i is a predictor variable, and n is the total number of predictor variables. The weights w_i are typically fit to minimize the residual summed square error (RSS) loss function

$$RSS = \sum_{j=0}^{N} (y_j - \hat{y}_j)^2. \tag{2.2}$$

Here, y_j is the true value, $\hat{y}_j$ is the predicted value, and N is the number of samples in the dataset. Basic linear regression models have the advantage of simplicity and computational efficiency, meaning that they are fast to train and easily interpretable. However, they can also be sensitive to outliers or noise in the dataset which can cause the models to overfit or fail to converge on a solution. These limitations of linear regression can be addressed in part by applying techniques to regulate the loss function such that the range of possible coefficients is constrained. Common regulation methods include ridge regression (Hoerl and Kennard 1970), Lasso regression (Tibshirani 1996), and elastic nets (Zou and Hastie 2005).

2.1.2 Random Forests

A random forest (Breiman 2001a) is a collection of decision trees (Breiman 1984) that work as an ensemble to estimate the solution to a regression or classification task. Each decision tree in a random forest is trained on a random subset of the training dataset sampled by bootstrapping with replacement. Only a small random subset of the total training variables are evaluated for splitting at each node within each tree, and this forces tree nodes to split along the best variable in the subset rather than along the best overall variable. As such, some trees in the forest are grown from suboptimal features within a dataset, resulting in greater tree diversity across the ensemble. Once a sufficient number of trees have been grown from the input data, the final prediction from the forest is the mean of the predicted values from all individual trees. By averaging the ensemble results, random forests are able to produce a smoother range of predicted values with lower variance than a single decision tree (Strobl et al. 2008). Additionally, random forests have been shown to exhibit a lower model bias compared to other tree-building techniques (Géron 2017).

2.1.3 Gradient Boosted Forests

Boosting within ML is commonly defined as any ensemble method that combines multiple weak learners to produce a strong learner (Géron 2017). In most cases, boosting methods iteratively train a model such that each subsequent iteration attempts to correct the errors of the previous version (Freund et al. 1996; Freund and Schapire 1997; Drucker 1997). Gradient boosting (Breiman 1997; Friedman 2001, 2002) applies this technique by first building and training a weak decision tree on the training dataset to solve a regression or classification

task. The gradient boosting algorithm then sequentially adds new predictors to the ensemble such that each new predictor is trained on the residual errors of the previous iteration (Géron 2017). Once a specified number of predictors have been trained, the final prediction of the gradient boosting model is determined by taking the sum of the predictions from all predictors in the ensemble. The iterative nature of the training process utilized by gradient boosting algorithms means that only one tree within the ensemble can be grown at a time. As a result, gradient boosting is often slower to train than other prominent ML algorithms. However, studies such as McGovern et al. (2017) and Gagne et al. (2009) have suggested that gradient boosting models may be more robust than other tree-based ML methods and better able to generalize to noisy data.

2.1.4 Isotonic Regression

When attempting to solve classification tasks, ML algorithms output probabilities that a given sample belongs to one or more predefined classes. However, the predictions produced by classification models are often overconfident and underdispersive, resulting in probability distributions biased towards the extremes of 0 and 1. To achieve probabilities that are more statistically reliable, a calibration technique known as isotonic regression can be applied. Isotonic regression (Dykstra and Robertson 1982) is a statistical inference (Barlow 1972) that finds a non-decreasing approximation of a function while minimizing the mean squared error of the training data. Isotonic regressions do not make any assumptions about the linearity of the target function, but do require that each point of the function be greater than or equal to the previous point (non-decreasing). In practice, isotonic regressions are typically trained on the output

of another ML classification and applied as a calibration so that the predicted class probabilities more closely match the true observed frequency.

2.2 Model Evaluation and Verification

When designing a new ML model (or any forecast product) for operational implementation, it is typically desirable, if not crucial, to supply potential end users with comparative metrics that describe the product's performance or goodness. Murphy (1993) explains that the goodness of a forecast is dependent on that forecast's consistency, quality, and value. In the context of automated algorithms and ML-derived products, consistency generally refers to how well a forecast aligns with the prior knowledge and experience of the expert user. This consistency, or lack thereof, is what often associates ML products with "black boxes" as described in Chapter 1. If an algorithm or model is difficult to interpret and the underlying logic that produces the output is opaque, domain end users cannot know if the predictions are consistent with their own expertise (Chase et al. 2022).

A forecast is said to have value if it provides a benefit to the user, and a forecast's quality can be measured by how well the prediction corresponds to observations (Murphy 1993). It is important to note, however, that forecast quality is not the same as forecast value. Consider a case in which an NWP model predicts heavy rain at a particular location. In this example, heavy rain does indeed occur within the region but not at the exact location predicted by the model. By many metrics, this forecast would be considered of poor quality; however, the prediction could still be valuable to a forecaster attempting to issue a regional rainfall forecast. The subjective nature of consistency and value

make these aspects of forecast goodness difficult to evaluate with generalizable metrics. As such, forecast quality is often the primary standard used to evaluate, verify, and compare modern forecast products. The importance and practical implications of product verification will be discussed further in Chapter 3, but first it is necessary to understand how forecast quality is evaluated and what metrics are appropriate for different applications.

According to Murphy (1993) and Wilks (2011), there are nine primary aspects of a forecast that contribute to its quality. Those attributes are:

1. Bias - the correspondence between the mean forecast and mean observation. In other words, bias is the average ratio of the number of forecast events to the number of observed events. A forecast with more predicted events than observations is known as an *overforecast* while a forecast with fewer predicted events than observations is an *underforecast*.

2. Association - the strength of the linear relationship between forecast and observation pairs. Association is often equivalent to the linear correlation coefficient between forecasts and observations.

3. Accuracy - the level of agreement between forecast and observation pairs, or how "correct" a forecast is. Disagreement between forecast and observation pairs is the forecast *error* and is inversely proportional to accuracy.

4. Skill - the accuracy of a forecast relative to the accuracy of a standard reference. The standard reference is often an unskilled forecast derived from random chance or climatology.

5. Reliability - the average agreement between conditional mean observations and the conditioning forecasts. Reliability is often determined by stratifying forecasts into different ranges or categories.

6. Resolution - the difference between the conditional mean observation and unconditional mean observation averaged across all forecasts. A forecast has resolution if it is able to sort observed events into frequency distributions that are different from each other.

7. Sharpness - the variability of the forecast distribution. For example, a smoothed climatology would typically exhibit low forecast variability and thus would have low sharpness.

8. Discrimination - the ability of the forecast to discriminate between different types of observations.

9. Uncertainty - the variability of the observation distribution.

These attributes together represent the joint probability distribution between forecasts and observations and provide a coherent framework for the verification process. Readers are referred to Murphy (1993) for more detail about the probability distribution subsets represented by each aspect of forecast quality.

A comprehensive depiction of the nine attributes of forecast quality requires the application of multiple metrics and evaluation techniques. In many instances, a given forecast can be reduced to a dichotomous prediction of whether or not an event will occur. For example, in non-deterministic frameworks this might mean selecting one or more thresholds to discriminate between "yes" and "no" predictions. The quality of discrete binary forecasts can then be evaluated

		Observed	
Forecast		**Yes**	**No**
	Yes	True Positive (TP)	False Positive (FP)
	No	False Negative (FN)	True Negative (TN)

Table 2.1: Example of a binary contingency table.

using a contingency table (Table 2.1; Wilks 2011), a 2x2 matrix that represents all possible forecast/observation combinations within the joint probability distribution. Attributes of the contingency table consist of true positives (TP), false positives (FP), false negatives (FN), and true negatives (TN). These values can be combined into a variety of metrics that together present a comprehensive depiction of the forecast quality. Common verification metrics derived from a binary contingency table are provided by equations 2.3–2.9 in Table 2.2.

Each score in Table 2.2 provides critical information about an aspect of forecast performance, but they can also be misleading when used incorrectly. For example, accuracy depicts the total fraction of correct predictions, where a higher accuracy equates to a more correct forecast. However, most forecast applications within the atmospheric sciences focus on predicting rare, high-impact events. As such, there is often a large class imbalance between the target rare event and the much more common null event. Because accuracy only depicts the total number of correct forecasts, a model or algorithm can achieve a high accuracy simply by always predicting the majority (null) class. Similarly, probability of detection (POD) can be artificially improved by predicting more "yes" events, while false alarm ratio (FAR) and success ratio (SR) can be improved

Accuracy	$\frac{TP+TN}{TP+FN+TN+FP}$	(2.3)
Probability of Detection (POD)	$\frac{TP}{TP+FN}$	(2.4)
Probability of False Detection (POFD)	$\frac{FP}{TN+FP}$	(2.5)
False Alarm Ratio (FAR)	$\frac{FP}{TP+FP}$	(2.6)
Success Ratio (SR)	$\frac{TP}{TP+FP}$	(2.7)
Bias	$\frac{TP+FP}{TP+FN}$	(2.8)
Critical Success Index (CSI)	$\frac{TP}{TP+FP+FN}$	(2.9)

Table 2.2: Common verification metrics derived from a binary contingency table.

with more "no" events. Therefore, it is important to consider all metrics together when evaluating the quality of a forecast. Critical success index (CSI) (Gilbert 1884) is often considered a better single-metric measure of a forecast's performance as it combines information about TP, FP, and FN into a composite score that equally penalizes misses and false alarms. However, CSI by itself does not distinguish the source of forecast error and is sensitive to the climatological frequency of the target event.

Additional information about a forecast's quality can be derived from combinations of the above metrics. For instance, a probabilistic forecast's ability to discriminate between two alternative outcomes can be evaluated by plotting the relationship between the forecast's POD and probability of false detection (POFD). This relationship, known as the relative operating characteristic (ROC; Mason 1982), is found by plotting POD as a function of POFD using a sequence of increasing thresholds to transform the probabilistic forecast into a series of discrete binary predictions. The ROC diagram is always constructed such that the lowest threshold results in a POD of 1 and a POFD of 0, while the highest threshold has a POD of 0 and a POFD of 1. The integrated area under

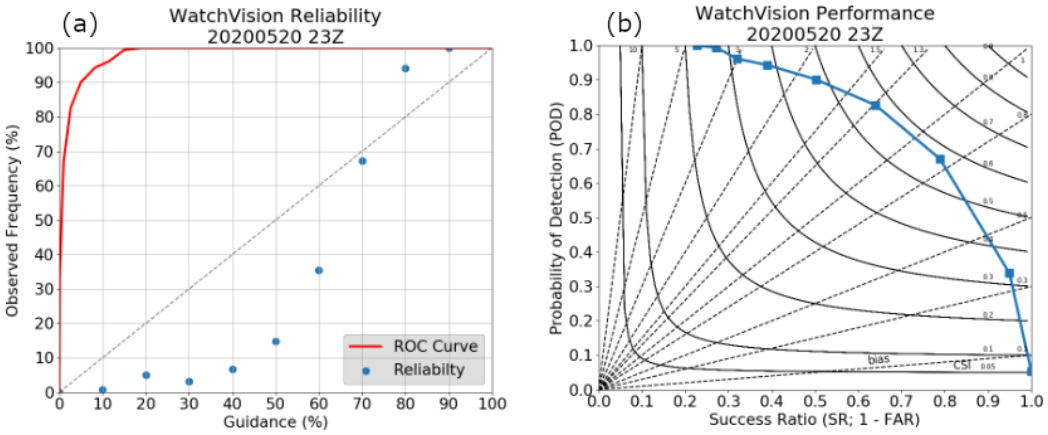


Figure 2.1: (a) Example of a combined ROC/reliability diagram for a probabilistic forecast. The dashed line represents the line of no skill for the ROC curve and the one-to-one line for the reliability plot. This case demonstrates a forecast with good potential usefulness but poor calibration. (b) Example of a performance diagram for a probabilistic forecast. See Chapter 6 for more information about the data plotted.

the ROC curve (AUC) is used to score the relative skill of the forecast across all thresholds, such that a score of 1 represents a perfect forecast and a score of 0.5 equates the forecast to random chance. ROC curves are not sensitive to forecast bias, so a poorly calibrated prediction may still exhibit a favorable AUC. As such, ROC curves are often considered a measure of the *potential* usefulness of a forecast, but should be paired with other metrics particularly when calibration is important. To this end, ROC diagrams are often combined with reliability diagrams, which plot the relative observed frequency of an event against the relative forecast frequency (Zadrozny and Elkan 2002). An example of a combined ROC/reliability diagram is shown in Fig. 2.1a.

A variation of the ROC diagram, known as a performance diagram (Roebber 2009; Fig. 2.1b) can be used to assess many aspects of a forecast's quality in a single plot. By replacing POFD with SR on the x-axis, the performance diagram takes advantage of the mathematical relationship between POD, FAR, SR, CSI, and bias to explicitly depict all five metrics simultaneously. In general, a forecast's quality improves as POD, SR, CSI, and bias approach 1 (and FAR approaches 0). Therefore, higher quality forecasts lie toward the upper-right corner of the performance diagram. In addition, the relative skill of a forecast can be assessed by plotting a reference forecast or climatology on the performance diagram and comparing the forecasts' relative positions in the parameter space. Note that each score contained within a performance diagram ignores TN predictions. As such, a performance diagram is most appropriate for assessing the prediction of rare events where TN predictions are trivial.

Each of the verification metrics discussed thus far rely on a direct mapping of forecast-observation pairs to assess forecast quality. While these scores provide considerable information about the overall performance of a forecast, they do not explicitly diagnose the sources or nature of forecast error. Consider again the example where a heavy rainfall forecast is spatially offset from where the heavy rain was observed. In a spatial forecast product, metrics such as those listed in Table 2.2 require that the location of the predicted and observed events coincide within a margin of error determined by the spatial and temporal resolution of the forecast. This means that forecasts which correctly predict an event but slightly miss the time or location are penalized equally to forecasts that do not predict the event at all. To achieve a complete depiction of forecast performance, it is necessary to consider the value of the forecast in addition to its quality. As previously stated, forecast value is generally subjective in nature and can be

difficult to directly quantify via traditional metrics. That said, a separate class of metrics, known as diagnostic verification (Murphy et al. 1989), can be used to evaluate sources of error and estimate how these errors impact the value of a forecast.

One such diagnostic metric, the fractional skill score (FSS; Roberts and Lean 2008), can be used to assess how forecast skill varies with spatial scale. This metric directly compares the fractional coverage of predicted and observed events within a spatial window of user-specified size to assess how well the forecast corresponds to observations. The FSS can then be calculated as:

$$FSS = 1 - \frac{\frac{1}{N}\sum_N (P_f - P_o)^2}{\frac{1}{N}[\sum_N P_f^2 + \sum_N P_o^2]} \tag{2.10}$$

where P_f is the forecast fraction, P_o is the observed fraction, and N is the number of spatial windows within the domain. An FSS value of 1 indicates a perfect match between forecasts and observations, while a score of 0 represents a complete mismatch. Forecast events that do not occur and observed events that weren't forecast always result in an FSS of 0. By applying a neighborhood window to the forecast and observations, FSS does not penalize forecasts for small spatial displacement from the observed event. This window can also be applied along the temporal axis to assess error resulting from discrepancies in the predicted and observed timing of an event. Other diagnostic verification methods include fuzzy logic (Damrath 2004), multi-scale statistical organization (Zepeda-Arce et al. 2000), and spatial multi-event contingency tables (Atger 2001).

How appropriate a verification metric is for evaluating a particular forecast can depend on many factors including the type of forecast, the climatology of the target event, the question(s) being addressed by the verification, and the

target audience. When developing a product for operational implementation, many of these factors are dictated by the needs and conventions of the end users. As such, it is important to have a good understanding of these needs throughout the development cycle.

Chapter 3

How Forecasters Evaluate Machine Learning

To better identify and understand sources of forecaster hesitancy precluding the wider adoption and success of AI products within NWS operations, it is necessary to first understand the decision-making process by which end users determine whether a new product or system satisfies their needs. For example, how do NWS forecasters evaluate new products and technologies? What factors influence their decision to trust and implement those products in their daily procedures? Do these factors differ between ML-derived products and products designed via more traditional (i.e., non-ML) methods? To address these questions, a structured survey was presented to operational forecasters and researchers at the 2021 Hazardous Weather Testbed Spring Forecasting Experiment to elicit first-hand perspectives about the challenges facing AI-derived products in an operational environment.

3.1 The 2021 Spring Forecasting Experiment

The Spring Forecasting Experiment (SFE) is an annual program conducted as part of NOAA's Hazardous Weather Testbed (HWT) during which participants investigate and evaluate a variety of NWP models, convection-allowing models (CAMs), ML- and traditionally-derived products, and other forecast guidance for the the prediction of severe and high-impact weather (Kain et al. 2003; Gallo

et al. 2017; Clark et al. 2022). The SFE is jointly managed by NOAA's National Severe Storms Laboratory (NSSL) and SPC and is traditionally held within the HWT facility - a pysical space within the National Weather Center (NWC) in Norman, Oklahoma, that hosts a combined forecast and research laboratory. Week-long activities performed during the SFE are designed to emulate conditions within SPC operations, and experimental products are tested on live weather data across the contiguous United States (CONUS) to assess performance in varied circumstances. By bringing researchers, developers, and NWS forecasters together in one large experiment, SFEs provide an opportunity for systematic in-person collaboration and feedback.

The 2021 SFE was somewhat unique in that COVID-19-related restrictions precluded a traditional in-person experiment (Clark et al. 2021, 2022). Instead, the SFE was held virtually via the Google Meet video-communication service, and facilitators remotely guided participants through online web-based interfaces to help them assess and evaluate experimental products. As described by Clark et al. (2022), science-based discussions and collaborations can be difficult in a virtual environment. However, this new virtual format also unbound the number of participants from the size constraints of the physical HWT facility. As such, the 2021 SFE was able to host 133 invited forecasters, researchers, and students over a five week period from 3 May 2021 - 4 June 2021. This was the largest SFE in the program's history at the time, and participants hailed from a variety of NWS WFOs, NOAA research laboratories, universities, cooperative institutes, and international agencies. The increased participation and diversity of the 2021 SFE made the experiment attendees an ideal sample from which to elicit perspectives about ML in operational environments.

3.2 Survey Design

Attendees of the 2021 SFE were polled via an online form designed and distributed using the Qualtrics survey software. The virtual survey was composed of ten questions, including one multiple choice, five matrix tables, and four open responses. Question 1 (Q1) asked participants to select their professional background from a list of choices including "Operational forecaster," "Researcher," "Academic faculty/staff," "Student," and "Other". Multiple selections were permitted so that a participant could identify in more than one background. Those that selected "Other" were asked to specify their professional background in an open response field. This question was designed to assess the professional demographic of the survey participants, and the responses were later used to stratify survey results as described in section 3.3.3. Q2 further expanded on participant demographics by asking survey takers to indicate how often they utilize probabilistic forecast products as part of their work-related duties. For the purposes of this survey, probabilistic forecast products were defined as any probabilistic forecast derived from a model or ensemble that might represent the ensemble's inherent uncertainty and spread or otherwise extract information not explicitly provided by the original model. The question text also included links to various operational probabilistic forecasts (i.e., Jirak et al. 2014; Cintineo et al. 2020a; Harrison et al. 2022) to serve as examples of the types of products intended for this query. Participants responded by choosing the best frequency from a 5-point Likert scale (Jebb et al. 2021) with labels of "Never", "Once a week", "2-3 times a week", "4-6 times a week", and "Daily." The final demographic question, Q6, asked respondents to rate their general knowledge and understanding of ML, including random forests, DL, and other AI techniques.

As before, participants answered by selecting from a 5-point Likert scale ranging from "Not knowledgeable at all" to "Extremely knowledgeable."

The main body of the survey focused on identifying what factors respondents consider most important when deciding whether to implement a new forecast product as part of their daily procedures. To this end, Q4 requested participants first consider some probabilistic forecast product they have utilized in the past to serve as a point of reference. Respondents were then presented with a list of ten factors in randomized order and asked to answer the question, *"When evaluating how useful that product might be to your personal forecasting process, how important are each of the following factors?"* Each factor was independently assessed on a 5-point Likert scale ranging from "Not at all important" to "Extremely important," and participants were asked to describe any additional relevant factors in an open response Q5. The ten factors included in this survey (Table 3.1) were selected based on input from SPC forecasters, SPC management, and academic social scientists to represent various aspects of forecast consistency, quality, and value as described in Chapter 2. For example, the "statistical verification of a product" is a direct measure of forecast quality, while "how closely the product aligns with human-generated output" is representative of forecast consistency.

The next set of questions (Q7-8) once again asked respondents to evaluate the importance of the ten factors when determining how useful a new product might be for their personal forecast process. However, this time the participants were specifically told that the hypothetical product in consideration was the direct result of a ML model. These somewhat repetitive questions were included to help identify any differences in how the respondents evaluate a ML product

(A)	The statistical verification of the product
(B)	Previous experience evaluating experimental versions of the product
(C)	How closely the probabilistic output aligns with human-generated forecasts
(D)	Knowledge of how the probabilistic output is derived
(E)	How closely the variables used as inputs to the product align with traditional meteorological knowledge
(F)	Use by other experts in the field
(G)	Timeliness and availability of the product
(H)	Previous experience with the developers of the product
(I)	Knowledge of the product's limitations and failure conditions
(J)	Performance of the product in case studies

Table 3.1: The list of factors survey respondents were asked to consider when evaluating what variables influence their decision to implement a new product in their personal forecast process.

compared to more traditional products. When designing this section of the survey, it was hypothesized that any discrepancies in the results between Q4-5 and Q7-8 might represent a change in evaluation priorities specific to ML products, and that these discrepancies may partially explain the apparent hesitancy of forecasters to adopt new ML products operationally. As before, an open response field was provided for participants to describe any relevant factors not included in the question matrix.

Finally (Q9), survey participants were asked to indicate their subjective perspectives on how important it is for researchers and developers to collaborate with operational forecasters during each phase of a product's development cycle (i.e., exploratory research; initial design and planning; technical and logistical development; product testing; and publication, training, and outreach). The importance of collaboration at each stage of development was assessed independently via a 5-point Likert scale with labels ranging from "Not very important" to "Extremely important." This last question was included to assess interest in multidisciplinary collaboration during the R2O process. Respondents were also encouraged to provide any additional comments they might have regarding researcher/forecaster collaboration in an open response field (Q10).

The Qualtrics survey was introduced to the 2021 SFE attendees during the morning of the first experiment day of each week immediately following introductions. The survey was voluntary and attendees were informed that all responses would be deidentified prior to analysis. Those that agreed to participate completed the survey virtually on their personal devices and were permitted to review their responses prior to submission. This study was approved by the University of Oklahoma Institutional Review Board. A copy of the survey is provided for reference in Appendix A.

Background	Number of respondents	Percent of respondents
Operational forecaster	36	34%
Researcher	38	35%
Academic faculty/staff	16	15%
Students	10	9%
Other	7	7%

Table 3.2: The number and percent of survey respondents in each professional background. Participants who selected multiple backgrounds are counted in each associated category.

3.3 Results and Discussion

3.3.1 Survey Demographics

Of the 133 attendees of the 2021 SFE, 92 voluntarily completed the survey for a 69% response rate. Respondents consisted of 36 operational forecasters, 38 researchers, 16 academic faculty/staff, 10 students, and 7 individuals who identified in an "other" professional background (Table 3.2). Thirteen respondents selected multiple backgrounds, including 4 that identified as both a researcher and an operational forecaster. The remaining mixed backgrounds consisted of combinations of researchers, academic faculty/staff, and students. Individuals who did not fit into the provided categories specified their professional backgrounds as model developers, program managers, and private sector employees.

Survey participants were generally found to be familiar with the concept of probabilistic forecast products, but utilization of those products in a formal capacity varied (Fig. 3.1a). About 33% of all respondents reported that they apply probabilistic forecast products as part of their typical work duties on

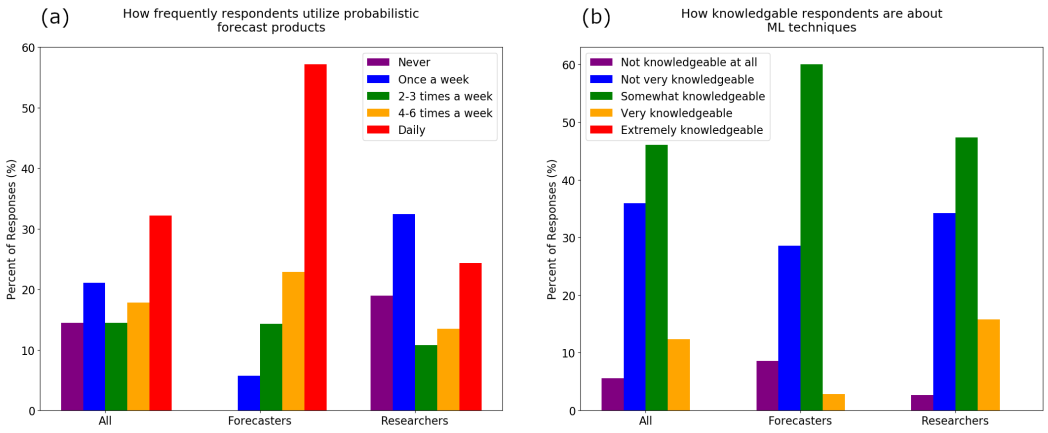


Figure 3.1: (a) How often survey respondents utilize probabilistic forecast products as part of their typical work responsibilities. (b) How knowledgeable survey respondents are ML, including random forests, DL, and other AI techniques.

a daily basis, 18% indicated they use the products 4-6 times per week, 15% selected 2-3 times per week, and 21% said once per week. Fifteen percent of all participants stated that they never use probabilistic forecast products as part of their typical work duties. Perhaps unsurprisingly, those who identified as operational forecasters reported the most frequent utilization, with about 95% stating they use probabilistic forecast products more than once per week and 58% on a daily basis. Conversely, researchers' responses were much more varied, with about 25% reporting they use the products on a daily basis and 51% indicating they use the products at most once per week. Note that these results do not necessarily reflect how much experience participants have with probabilistic forecasts as respondents may still use such products outside of their typical work-related duties. Instead, this question was intended to assess how familiar respondents are with the formal application of probabilistic forecast products within a structured professional environment. This will be discussed further in section 3.3.3.

Participants generally described themselves as somewhat or not very knowledgeable about ML techniques, with about 82% of all respondents falling in these categories (Fig. 3.1b). Approximately 12% indicated they were very knowledgeable of ML, while 6% of participants said they were not knowledgeable at all. These results were relatively similar between researchers and forecasters as well, though researchers did generally express slightly more familiarity with ML practices overall. Of the 38 respondents who identified as a researcher, about 15% claimed to be very knowledgeable of ML, 48% were somewhat knowledgeable, 34% were not very knowledgeable, and 3% were not knowledgeable at all. In comparison, only 3% of forecasters indicated they were very knowledgeable about ML techniques, 60% were somewhat knowledgeable, 29% were not very knowledgeable, and 8% indicated that they were not knowledgeable at all. Notably, no survey respondent reported to be extremely knowledgeable about ML, DL, or other AI techniques.

3.3.2 Machine Learning vs. Traditional Forecast Products

Participant responses to Q4 and Q7 were processed and evaluated to identify what factors most influence an end user's decision to trust and implement a new product as part of their personal forecast process. To aid in this evaluation, a kernel density estimation (KDE; Wilks 2011) was applied to the data. The resulting KDE curves are analogous to histograms and approximate the discrete survey responses as a linear combination of nonparametric Gaussian probability density functions. These curves more accurately estimate the underlying distributions of the survey responses than traditional histograms which are sensitive to the selected bin intervals and end points and often demonstrate greater variance for small sample sizes (Potvin et al. 2019). The KDE curves

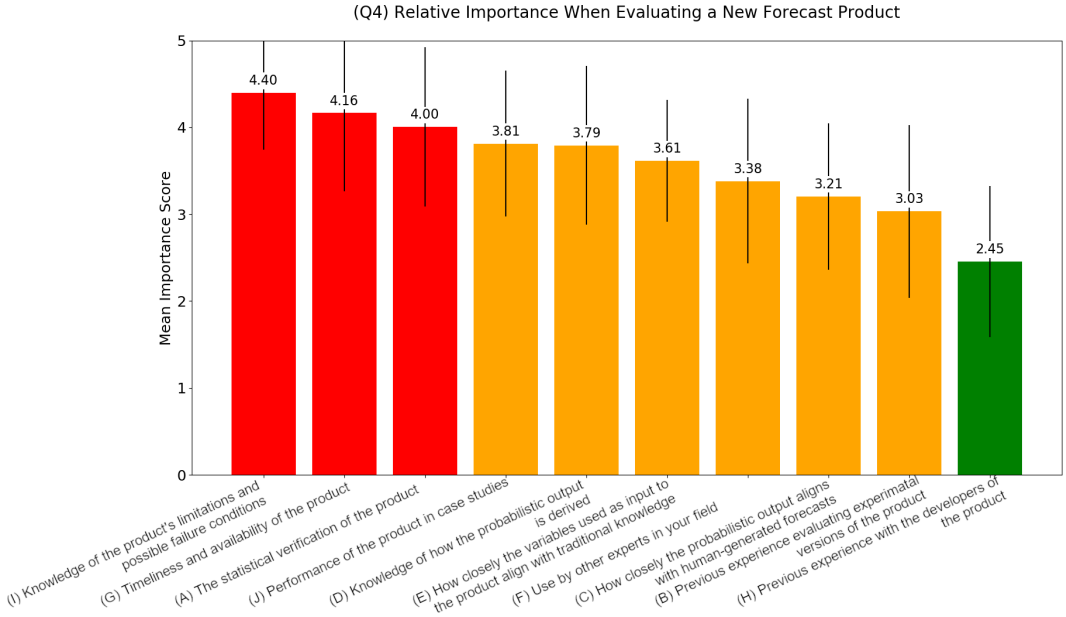


Figure 3.2: Mean relative importance when evaluating a generic probabilistic forecast product. Error bars represent the 95% confidence interval from 10,000 bootstrapped samples.

presented here represent the relative frequency of survey responses across the provided 5-point Likert scale and are useful for identifying, interpreting, and comparing response variance. To best present the data, a gaussian kernel with a bandwidth of 0.35 was utilized to smooth the KDE curves. Additionally, the mean score and variance of each factor was computed from 10,000 bootstrapped samples.

Survey participants rated knowing the limitations and possible failure conditions of a probabilistic forecast product as the most important factor when evaluating that product for operational application. This consideration achieved a bootstrapped mean score of 4.40 out of the possible 5.00 across all survey respondents (Fig. 3.2), with 92% of participants citing the factor as “very” or

"extremely" important. The next most important factors according to the survey results were the timeliness and reliable availability of the product (4.17), the statistical verification of the product (4.02), how well the product performed in case studies (3.81), and knowledge of how the probabilistic output is derived (3.78). How closely the product inputs align with traditional meteorological knowledge was rated 3.61, and the product's use by other experts in the field scored 3.36 on average. Finally, how closely the probabilistic output aligns with human-generated products (3.20), the user's previous experience evaluating experimental versions of the product (3.02), and the user's previous experience with the developers of the product (2.45) were rated as the least important factors to the participants' decision-making process. Other important factors commonly mentioned by participants in an optional open response prompt (Q5) include the product's ease of access, how well end users understand the product output, and how well the product is tuned to the end users' specific needs.

Considerable variability was noted in the survey responses, and this was particularly observed in the lower-ranking factors (Fig. 3.3). For example, the importance of a user's previous experience evaluating experimental versions of a product saw the greatest disagreement among all survey participants, with a bootstrapped standard deviation of 0.99. The KDE curve for that factor reveals nearly equal response rates of "not very important," "somewhat important," and "very important," representing a wide range of views among respondents. How often the product is used by other experts in the field saw the second highest standard deviation of 0.95, while the statistical verification of the product had a standard deviation of 0.92. Conversely, having knowledge of a product's limitations and possible failure conditions saw the most respondent agreement with a bootstrapped standard deviation of 0.66. How well the product inputs

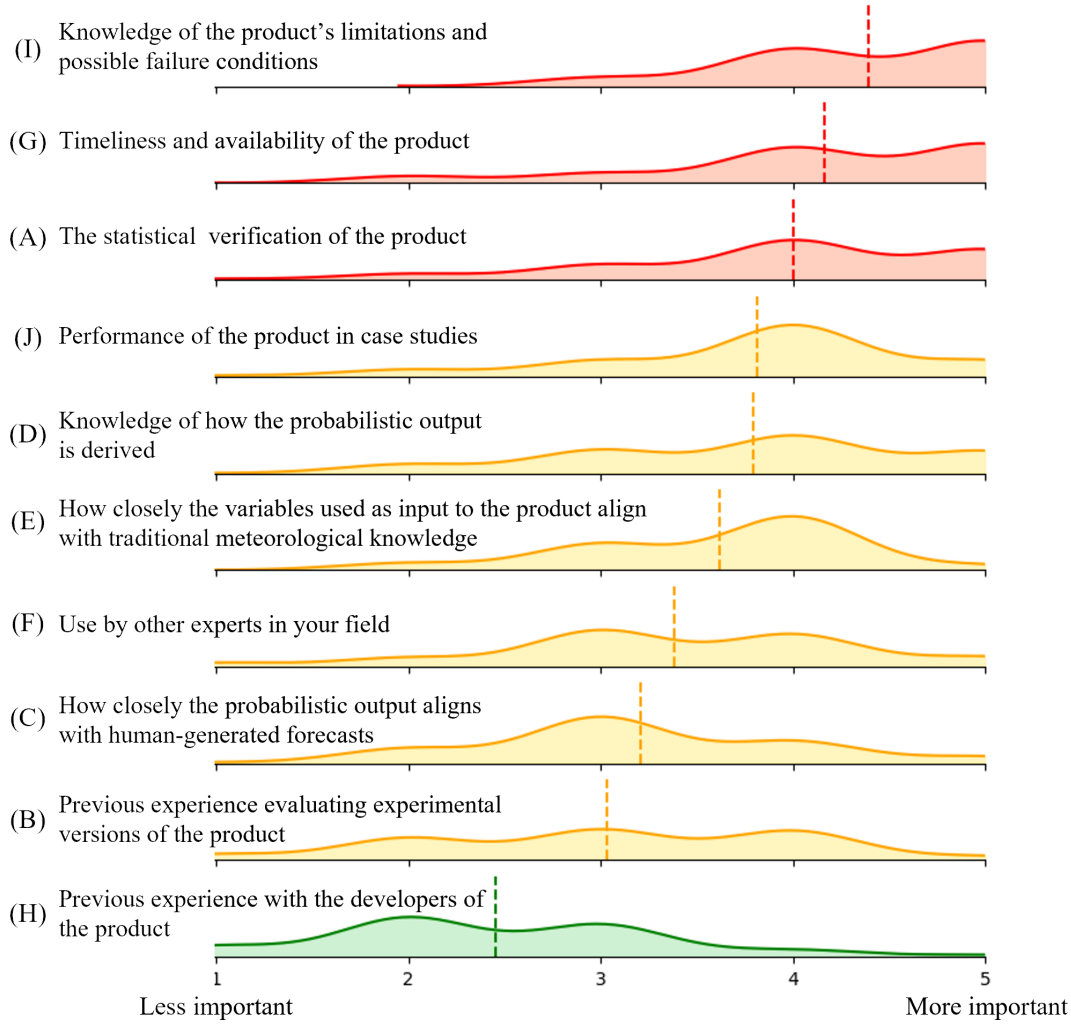


Figure 3.3: Survey Q4 responses approximated as KDE curves. Dashed vertical lines represent the mean score of each factor. Red fill represents factors with bootstrapped mean scores ≥ 4.0, yellow is used for mean scores between 3.0 - 4.0, and green indicates mean scores < 3.0.

align with traditional meteorological knowledge had the second lowest standard deviation of 0.70, largely stemming from consistent scores of "somewhat" or "very" important.

It is notable that the third highest ranked factor also exhibited the third greatest variability among all participant responses. About 78% of respondents rated a product's statistical verification as "very" or "extremely" important, 15% said it is "somewhat" important, 6% responded with "not very" important, and one operational forecaster stated that it is "not at all" important to their decision-making process. That forecaster explained their reasoning in the open response Q5, commenting, "A product's statistical verification does not necessarily translate to its usefulness on the forecast desk. I don't care if it has been shown to be 75% accurate over the entire CONUS over the last 5 years. That's great, but what I really need to know is: Should I trust this product right now over my forecast area for the time period in question?" This insight indirectly invokes the relationship between a forecast's quality and value as described in Chapter 2. A forecast product may exhibit excellent statistical verification, but it may not necessarily be considered a good forecast if it does not provide a benefit to the end user in a specific situation. As described by Murphy (1993), the relationship between a forecast's quality and value is inherently nonlinear and subjective, varying from situation to situation and user to user. In this instance, the survey respondent placed greater emphasis on forecast value while other participants applied equal or greater weighting to forecast quality. This discrepancy potentially explains some of the variability observed in the survey results. Other sources of variability among participant responses are explored in section 3.3.3.

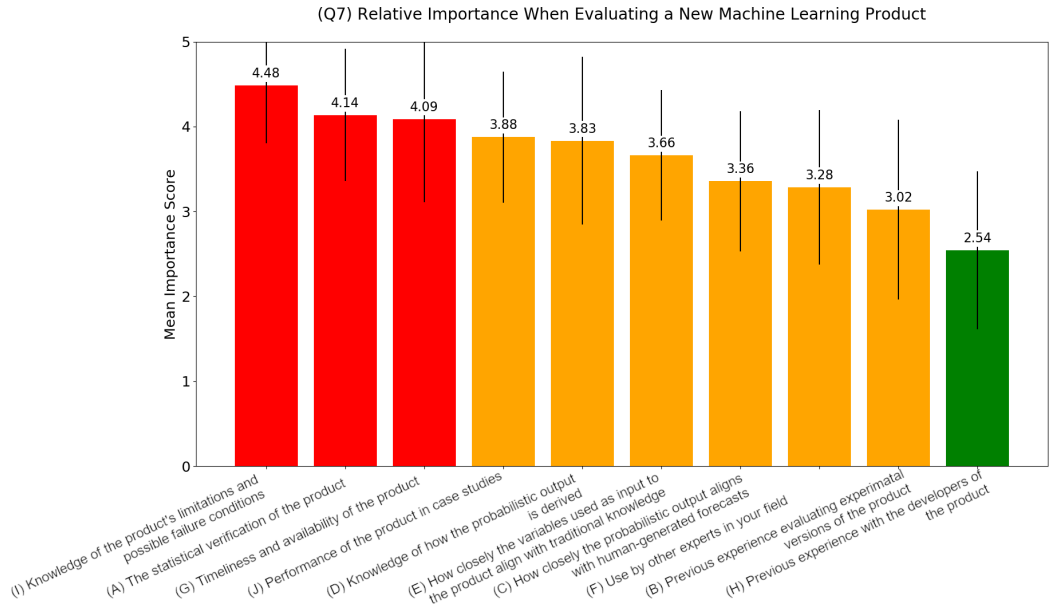


Figure 3.4: Mean relative importance when evaluating an ML probabilistic forecast product. Error bars represent the 95% confidence interval from 10,000 bootstrapped samples.

Participants overall showed little discrimination in how they ranked factors between traditional probabilistic forecast products (Q4) and products specifically derived from ML methods (Q7). The indicated importance of the ten variables was nearly identical for ML products as those previously described, with knowledge of the product's limitations and possible failure conditions once again earning the highest bootstrapped mean score of 4.49 (Fig. 3.4). In fact, there were only two changes to the order of importance compared to those shown in Fig. 3.2. First, the statistical verification of the product moved up from the third most important factor to the second most important, jumping the timeliness and availability of the product with scores of 4.15 and 4.10 respectively. Second, how often the product is used by other experts in the field fell to the

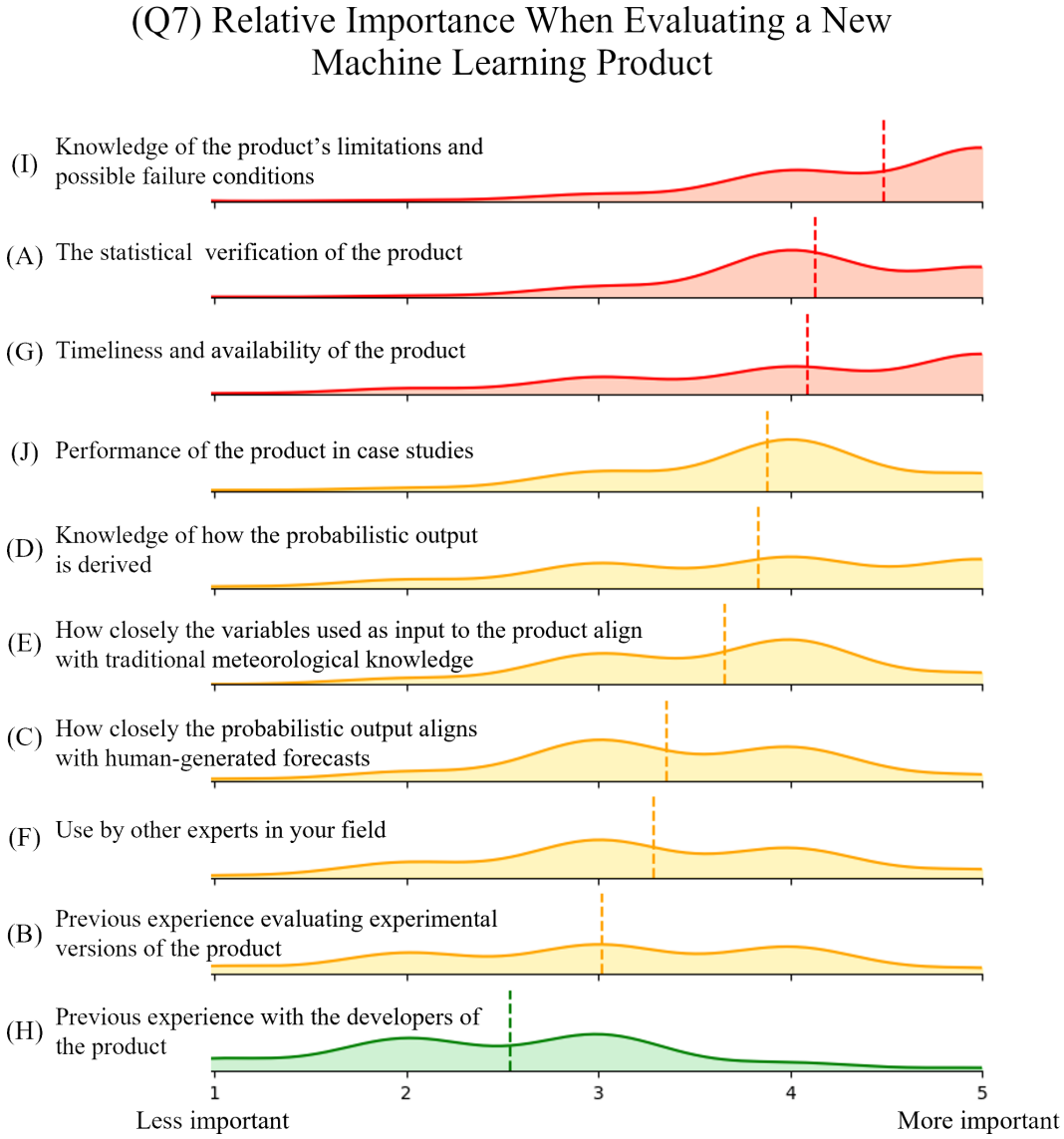


Figure 3.5: Survey Q7 responses approximated as KDE curves. Dashed vertical lines represent the mean score of each factor. Red fill represents factors with bootstrapped mean scores ≥ 4.0, yellow is used for mean scores between 3.0 - 4.0, and green indicates mean scores < 3.0.

third least important factor (3.26) after being overtaken by how closely the probabilistic output aligns with human-generated forecasts (3.35). However, these differences fall well within the bootstrapped 95% confidence intervals of each factor, making the statistical relevance of the changes dubious at best. More notable differences were observed in the variability of survey responses between generic probabilistic forecast products and ML-derived products (Fig. 3.5). As before, the importance of a user's previous experience evaluating experimental versions of a product saw the greatest respondent disagreement, with a standard deviation of 1.06. This time, however, knowledge of how the probabilistic output is derived had the second greatest variability with a standard deviation of 0.99, followed by the timeliness and availability of the product at 0.98. Knowledge of the ML product's limitations and failure conditions once again had the most respondent agreement with a standard deviation of 0.68, followed by how closely the product inputs align with traditional meteorological knowledge at 0.77. Comparisons of the mean scores and standard deviations between generic and ML-derived probabilistic forecast products are provided in Tables 3.3 and 3.4.

These results appear to largely refute the earlier hypothesis that discrepancies between Q4 and Q7 might represent a change in evaluation priorities specific to ML products, and that these discrepancies may partially explain the apparent hesitancy of forecasters to adopt new ML products operationally. Instead, the survey responses suggest that the respondents on average do not consciously evaluate ML-derived forecast products any differently than they do products derived from more traditional methods. This is further supported by the direct comments of the respondents, such as the private sector employee

who stated, "I would treat a machine learning-produced product pretty similarly to any other probabilistic product." However, there is a caveat to these conclusions. Respondents were made aware of the nature of this survey prior to their participation, and the limited number of questions make it apparent that Q4 and Q7 are intended to be compared. As such, this design may introduce the possibility of acquiescence and desirability bias where respondents consciously or subconsciously provide the "desired" responses rather than their true opinions. Because participants were allowed to review their answers prior to submitting, it is possible that some respondents may have adjusted their ratings to be consistent between Q4 and Q7 to avoid the appearance of bias for or against ML-derived products. Regardless of these potential response biases, the survey results provide useful insight into the decision-making process of the end users and generally align with the results of past and ongoing studies (e.g., Doswell 2004; Hoffman et al. 2017; Cains et al. 2022).

	Factor	$\mu(Q4)$	$\mu(Q7)$	$\mu(Q7) - \mu(Q4)$
(I)	Knowledge of the product's limitations and failure conditions	4.40	4.49	+0.09
(G)	Timeliness and availability of the product	4.17	4.10	-0.07
(A)	The statistical verification of the product	4.02	4.15	+0.13
(J)	Performance of the product in case studies	3.81	3.88	+0.07
(D)	Knowledge of how the probabilistic output is derived	3.78	3.82	+0.04
(E)	How closely the variables used as inputs to the product align with traditional meteorological knowledge	3.61	3.65	+0.04
(F)	Use by other experts in the field	3.36	3.26	-0.10
(C)	How closely the probabilistic output aligns with human-generated forecasts	3.20	3.35	+0.15
(B)	Previous experience evaluating experimental versions of the product	3.02	3.02	0.00
(H)	Previous experience with the developers of the product	2.45	2.53	+0.08

Table 3.3: Bootstrapped mean importance scores μ for generic probabilistic forecast products (Q4), ML-derived probabilistic forecast products (Q7), and the differences between the two.

	Factor	$\sigma(Q4)$	$\sigma(Q7)$	$\sigma(Q7)-\sigma(Q4)$
(I)	Knowledge of the product's limitations and failure conditions	0.66	0.68	+0.02
(G)	Timeliness and availability of the product	0.90	0.98	+0.08
(A)	The statistical verification of the product	0.92	0.78	-0.14
(J)	Performance of the product in case studies	0.84	0.77	-0.07
(D)	Knowledge of how the probabilistic output is derived	0.91	0.99	+0.08
(E)	How closely the variables used as inputs to the product align with traditional meteorological knowledge	0.70	0.77	+0.07
(F)	Use by other experts in the field	0.95	0.91	-0.94
(C)	How closely the probabilistic output aligns with human-generated forecasts	0.84	0.83	-0.01
(B)	Previous experience evaluating experimental versions of the product	0.99	1.06	+0.07
(H)	Previous experience with the developers of the product	0.87	0.93	+0.06

Table 3.4: As in Table 3.3, but for the bootstrapped standard deviation (σ).

3.3.3 Researcher vs. Forecaster Perspectives

To further explore the observed variability in survey responses, all results were stratified by professional background and compared. This strategy revealed notable differences between the responses of individuals who identified as researchers and those who identified as operational forecasters as shown in Fig. 3.6 and 3.7. Forecasters, for example, rated the timeliness and availability of a generic probabilistic forecast product as the most influential consideration when evaluating the usefulness of that product. This factor was given a mean score of 4.64 and surpassed the rated importance of knowledge about a product's limitations and failure conditions at 4.50. Forecasters were largely in agreement about the importance of a product's timeliness and availability as well, with the factor earning a near unanimous rating of "very" or "extremely" important and a standard deviation of 0.54. Conversely, researchers rated the factor as their fourth most important consideration on average, with a mean score of 3.97 and a standard deviation of 1.0. One forecaster explained the importance of product availability among other factors, commenting, "A good training/overview of the product, good visualization tools, and easy, reliable access are all important. These things help me to begin dabbling in the new product during real-time operations without compromising my attention to other things. Once I can get to that stage of somewhat routine experimentation with the product, it has a chance to impress me with performance in individual cases, which is actually more important to me than the statistical verification." Another operational forecaster noted, "I also look at how the product has performed over time in real-world conditions based on looking at it informally during active weather, especially during atypical or unusual situations." These anecdotes mirror the recommendations of Hoffman et al. (2013), who suggest that end users require

time to assess an automated product's reliability, validity, utility, robustness, and false alarm before trust in that product can be earned. As such, *having timely and reliable access to a new probabilistic forecast product enables the assessment of the other factors included within this survey.*

Other notable differences between researcher and forecaster responses were observed for lower ranking factors as well. For example, forecasters rated how closely a product's input variables align with traditional meteorological knowledge with a mean score of 3.72 and standard deviation of 0.93. In comparison, researchers rated that factor a 3.08 on average with a standard deviation of 1.02. Forecasters also tended to place more value on how often a product is used by other experts in the field, with a mean score of 3.5 compared to researchers' score of 3.0.

Forecaster responses exhibited only marginal discrepancies between generic probabilistic forecast products (Q4) and ML-derived products (Q7), with no change in the assessed order of importance (Fig. 3.7). The bootstrapped mean score of a product's timeliness and availability increased slightly from 4.64 to 4.72, while how closely a product's inputs align with traditional meteorological knowledge decreased from 3.72 to 3.56. Factors such as a product's statistical reliability and the performance of a product in case studies achieved the same average scores on both questions. Despite these relatively consistent scores, many forecaster respondents provided additional context about their thoughts regarding ML in the optional open response Q8. For example, one operational forecaster stated, "Some knowledge of the inner-workings of the machine model algorithm would be helpful. Often, I feel that ML algorithms are somewhat of a black box. Knowing that the guidance is based upon factors that hold meteorological significance would improve my confidence in using the output."

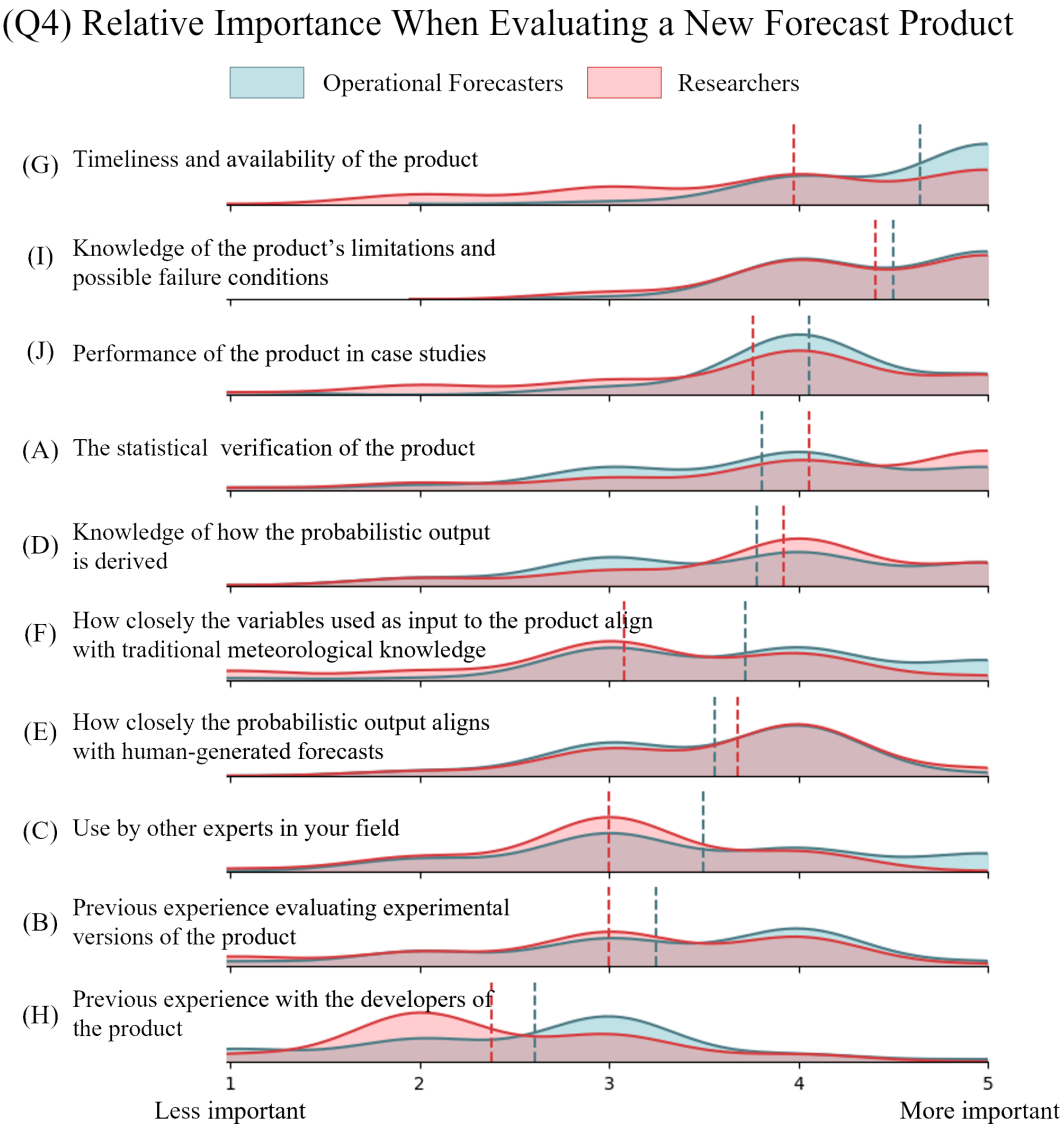


Figure 3.6: Survey Q4 forecaster and researcher responses approximated as KDE curves. Dashed vertical lines represent the mean score of each factor.

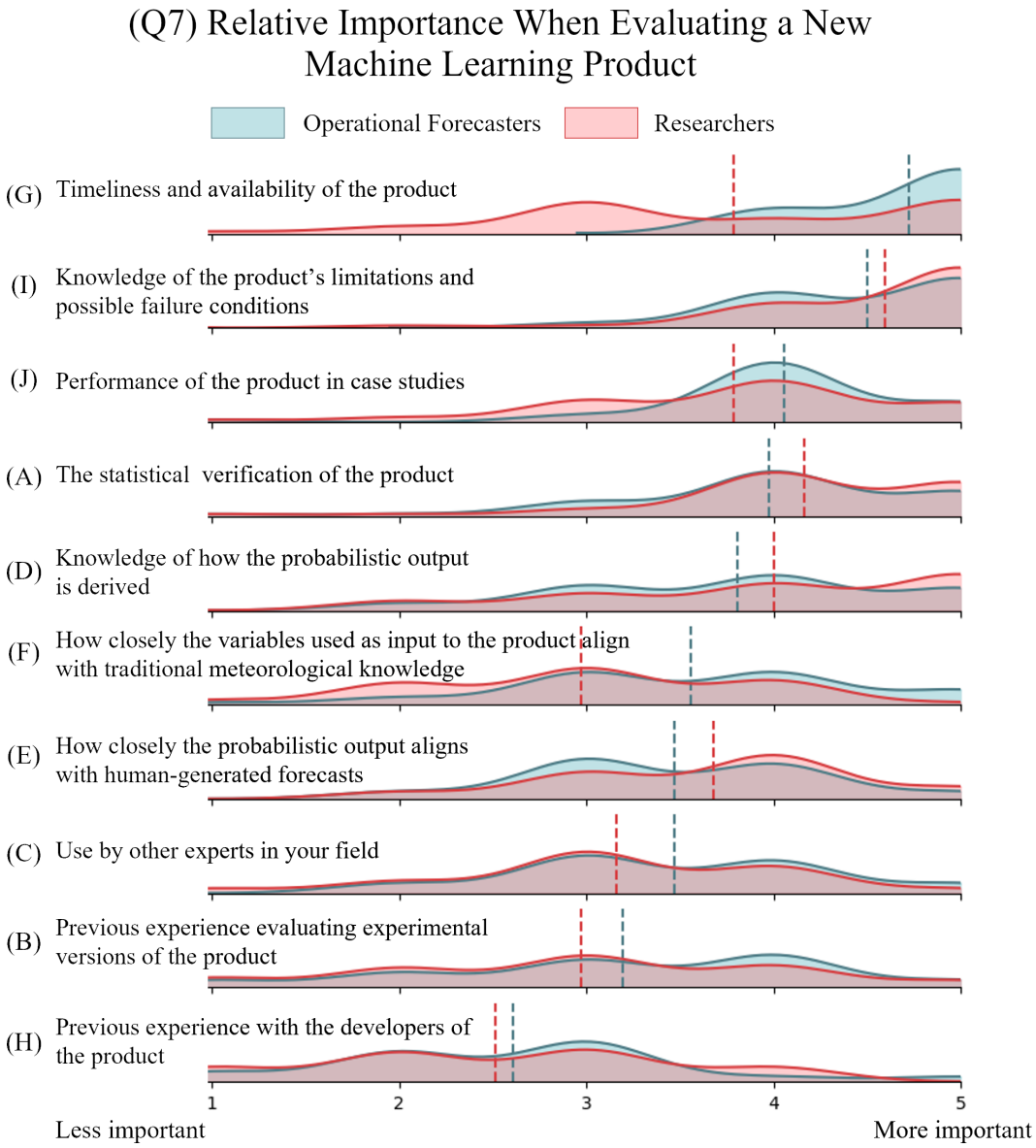


Figure 3.7: Survey Q7 forecaster and researcher responses approximated as KDE curves. Dashed vertical lines represent the mean score of each factor.

These insights were echoed by another forecaster with, "Most of my evaluation would be based on how well the ML product does in real time. The main thing I know about ML is that it has to be given the right problem to solve to be the most effective, so I would be very interested in how the ML was done." Finally, a third forecaster contextualized their responses with an anecdote: "I like to know what was the main goal of the developers, and does that goal directly correspond to a product or service that I provide as a forecaster? [. . .] I began using a ML-based tool to support national flash flood outlooks, but the tool was based on rainfall recurrence intervals and did not account for wet soil. This was a key limitation I needed to know."

In contrast to the consistent forecaster ratings, the mean importance scores assessed from researcher responses saw more considerable changes between generic forecast products and ML-derived products (Fig. 3.7). For instance, the mean importance of how closely a product's input variables align with traditional meteorological knowledge fell from 3.08 to 2.97, while how often the product is used by other experts in the field increased from 3.00 to 3.16. Additionally, the indicated importance of a product's timeliness and availability decreased from 3.97 to 3.78, and the importance of the statistical verification of a product increased from 4.41 to 4.59. Overall, researcher responses between Q4 and Q7 changed by a mean absolute difference of 0.10, while forecaster responses exhibited a mean absolute difference of 0.06. However, these differences were found to be well within the bootstrapped 95% confidence intervals of each factor and thus should only be interpreted as a trend in the data rather than a statistically significant conclusion. A comparison of researcher and forecaster responses for Q4 and Q7 are provided in Table 3.5. Notably, researcher respondents tended to assign greater importance to factors related to the understanding of how an

ML product works than they did for generic probabilistic forecast products. This is further supported by the comments of researcher participants, with one stating, "I would like to know which type of machine learning is utilized and why because each type has its strengths and weaknesses." Another researcher indicated that "information on the training data used to derive the ML product and any biases that may be introduced from that training data" would be an important consideration as they evaluated the usefulness of the product.

Overall, survey participants who identified as operational forecasters tended to place more importance on factors that represent a forecast's consistency and value, while researcher responses generally emphasized considerations of forecast quality. These results potentially reflect the different skills, experience, and needs of the two professions and showcase the varying perspectives contained within the diverse meteorological community. Professional researchers are often intimately familiar with the demands of peer review and the need for objective metrics and other statistics by which new products are typically assessed. Conversely, operational forecasters are most concerned about their next forecast and using the tools at their disposal to provide the best service to their partners and end users. A forecast product has no intrinsic value on its own, but rather gains value by influencing the decisions of those who use it (Murphy 1993). As such, the long-term statistical verification of a forecast product does not matter in an operational setting if the product is unable to benefit the forecast process. These differing perspectives can represent a healthy balance of theoretical and operational expertise within the meteorological community (Doswell et al. 1981); however, they can also be a source of misunderstanding, confusion, and conflict between researchers and forecasters when transitioning a new product through R2O processes (Deal and Hoffman 2010a).

The meteorological research community generally drives the development, testing, and peer review of new ML forecast products that are eventually transitioned into NWS operations. However, the results of this survey suggest that those same researchers may evaluate ML products with different priorities than those of the intended end users. If these differing perspectives remain unchecked during development, researchers and developers may utlimately produce a finished product that is considered a success by their standards but completely fails to meet the needs of their intended end users. As such, I offer the hypothesis that forecasters are not necessarily hesitant to adopt new ML products because they evaluate ML with more scrutiny than other methods. Rather, *ML products that struggle in the R2O transition may not provide a tangible benefit or otherwise fail to meet the needs of their intended end users.* To repeat a line from Chapter 1, sustained collaboration serves as the most viable strategy to bridge the R2O gap by ensuring new products address a real operational need, satisfy operational requirements, and are presented in a way that is accessible by forecasters (Kain et al. 2003). Therefore, these survey results support the calls of Doswell et al. (1981), Auciello and Lavoie (1993), Serafin et al. (2002), Kain et al. (2003) and others for increased collaboration between the research and operational communities so that we may better apply our diverse expertise toward our shared scientific goals.

	Factor	Q4			Q7		
		μ_F	μ_R	$\mu_F - \mu_R$	μ_F	μ_R	$\mu_F - \mu_R$
(G)	Timeliness and availability of the product	4.64	3.97	+0.67	4.72	3.78	+0.94
(I)	Knowledge of the product's limitations and failure conditions	4.50	4.41	+0.09	4.50	4.59	-0.09
(J)	Performance of the product in case studies	4.06	3.76	+0.30	4.06	3.78	+0.28
(A)	The statistical verification of the product	3.81	4.05	-0.24	3.97	4.16	-0.19
(D)	Knowledge of how the probabilistic output is derived	3.78	3.92	-0.14	3.81	4.00	-0.19
(F)	Use by other experts in the field	3.72	3.08	+0.64	3.56	2.97	+0.59
(E)	How closely the variables used as inputs to the product align with traditional meteorological knowledge	3.56	3.68	-0.12	3.47	3.68	-0.21
(C)	How closely the probabilistic output aligns with human-generated forecasts	3.50	3.00	+0.50	3.47	3.16	+0.31
(B)	Previous experience evaluating experimental versions of the product	3.25	3.00	+0.25	3.19	2.97	+0.22
(H)	Previous experience with the developers of the product	2.61	2.38	+0.23	2.61	2.51	+0.10

Table 3.5: Bootstrapped mean importance scores for generic probabilistic forecast products (Q4) and ML-derived probabilistic forecast products (Q7) as rated by operational forecasters (μ_F) and researchers (μ_R).

3.3.4 Collaboration in Product Development

The survey results presented thus far support the need for increased collaboration between researchers and operational forecasters during the development and implementation of new forecast products. To further expand upon these conclusions, survey participants were asked to indicate how important they believe it is for researchers and developers to collaborate with forecasters during incremental stages of product development. Survey responses to this question were strongly positive among all participants, with the product testing phase achieving the highest bootstrapped mean score of 4.72 (Fig. 3.8). Respondents overwhelmingly rated collaboration at this stage as "very" or "extremely" important, with 99% of responses falling in these categories. Collaboration during publication, training, and outreach was given the second highest rating of 4.52, followed by collaboration during the initial design and planning phase at 3.94. Collaboration during exploratory research and during technical and logistical development tied for the lowest rating of 3.47. Notably, these responses were similar for both forecasters and researchers, indicating a strong interest in collaboration from both parties.

Forecasters in particular were very vocal in their support for increased collaboration, with one participant noting, "Collaboration throughout the development process is critical to providing something that meets the users' needs and finding issues early and correcting them prior to operationalization. [...] Loss of confidence in a product is difficult to win back." Another researcher/forecaster commented, "Some exploratory, visionary work could be done without forecasters, but to ensure utility of the product forecasters should be involved throughout development and testing. The training aspect is also critical so that forecasters understand strengths and weaknesses, and do not form their own vague

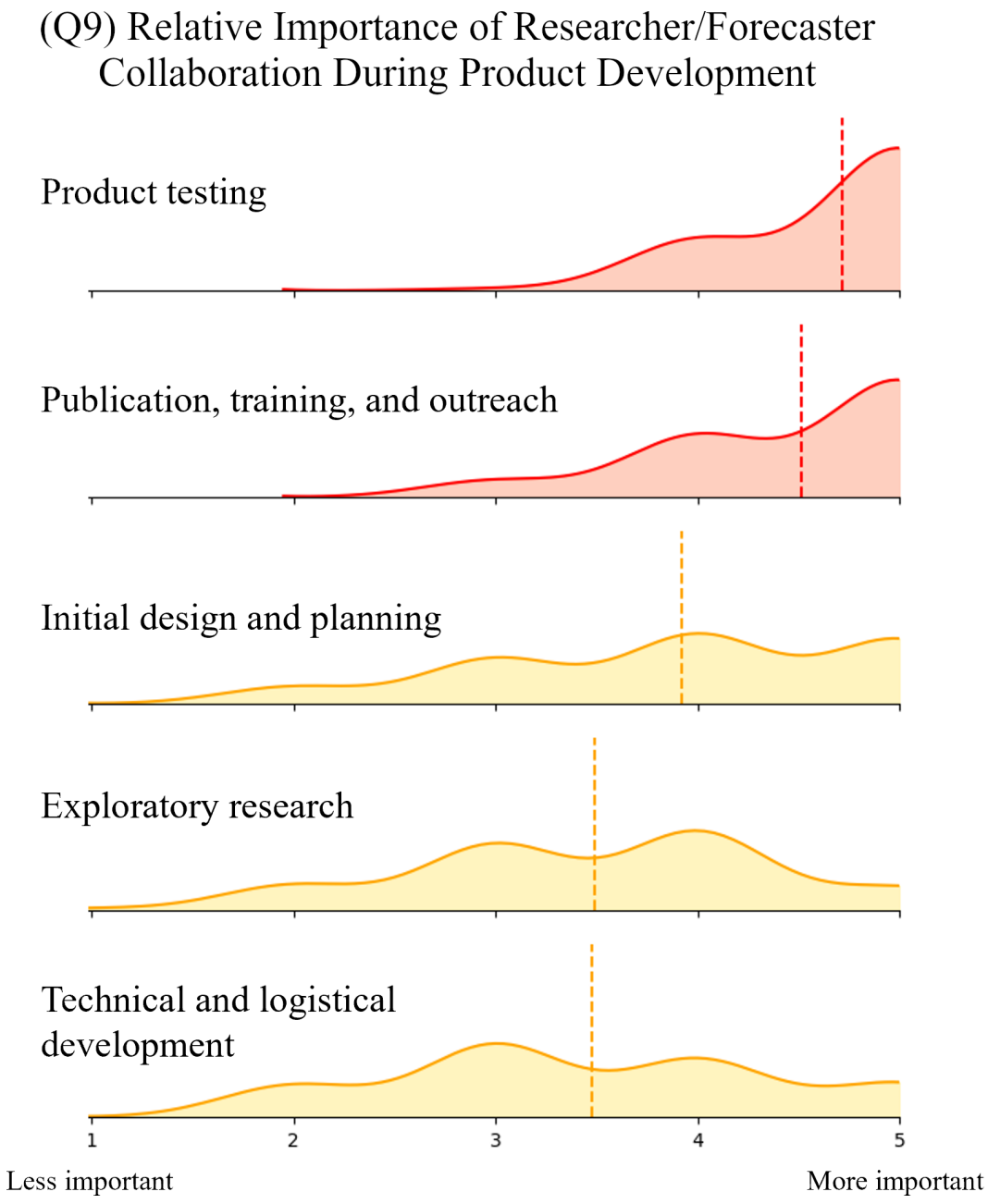


Figure 3.8: Survey Q9 responses approximated as KDE curves. Dashed vertical lines represent the mean importance of researcher/forecaster collaboration during each phase of production. Red fill represents factors with bootstrapped mean scores ≥ 4.0 and yellow indicates mean scores < 4.0.

impressions of what the product is designed to do." This sentiment was echoed by a researcher who suggested, "I believe it is quite important for developers and evaluators/forecasters to collaborate on a new forecast product well before the official evaluations are sent out. Offering forecasters the opportunity to give input at the earlier stages of development would be beneficial for all of those involved in the model evaluation/upgrade/implementation." Finally, a third forecaster expressed strong support for early collaboration, suggesting, "Early interaction with operational forecasters is extremely important because it offers insight to where we need additional forecast guidance support and the types of information, time-scale and data resolution that would be most beneficial on a daily basis. This can help steer the ship and make sure whatever new approaches are being designed will actually be useful. So early collaboration with operational forecasters is essential."

Such strong support for researcher/forecaster collaboration is encouraging, but the question remains how best to achieve it in practice. As described in Chapter 1, there are many barriers to continuous collaboration in traditional research and operational settings, including funding requirements, time constraints, and security concerns. These limitations have long stymied true collaboration within the development process (Doswell et al. 1981; Deal and Hoffman 2010a) and are in part responsible for the inadequacies of modern R2O pipelines. As such, a change in the current development paradigm may be required to overcome these challenges and successfully promote collaboration at all phases of the R2O process. This dissertation introduces the concept of collaborative co-production as a possible solution to these challenges, with the hypothesis that systemic collaboration throughout a product's development will improve the success of that product in the R2O transition.

Chapter 4

Collaborative Co-Production

Co-production is a collaborative process that provides a service or product via an equal, reciprocal relationship between professionals and end users (Boyle and Harris 2009). This relationship might be between software developers and members of the general public, academics and public decision makers, or researchers and operational forecasters, to name but a few examples. Collaborative co-production requires end users to be recognized as experts not only in their domain knowledge and experience but in their technological systems, procedures, and requirements as well (Realpe and Wallace 2010). Conversely, this necessitates researchers and developers to take on the roles of facilitator and practical implementer, working closely with their end users to find solutions to a problem or task. The core principle of co-production is that the end user should be considered a valuable resource and ally of the development process, and no development that ignores their contributions can be efficient (Boyle and Harris 2009). To fully realize these principles ultimately requires a shift of power, resources, and responsibility from researchers and developers to end users through deliberate, user-led, collaborative processes (Boyle and Harris 2009; Realpe and Wallace 2010).

Various degrees of co-production have long been integrated into industrial and commercial applications, but these principles have only recently seen explicit application within the atmospheric sciences. Indeed, the exploration of

co-production has seen a rapid growth particularly within the climate sciences over the past decade (e.g., Meadow et al. 2015; Kruk et al. 2017; Wall et al. 2017; Bremer et al. 2018; Kolstad et al. 2019; Ziaja 2019; Blair et al. 2022). These studies primarily apply and evaluate collaborative co-production principles as a means to bridge the often contentious divide between climate scientists and public decision makers to produce actionable climate research and knowledge. Within the R2O space, past publications such as Doswell (1986) and Auciello and Lavoie (1993) have indirectly advocated for the application of co-production between the research, academic, and operational communities; however the concept of collaborative co-production in the modern R2O process is perhaps best formalized by Hoffman et al. (2010) and their Practitioner's Cycles model.

Hoffman et al. (2010)'s Practitioner's Cycles is a conceptual collaborative co-production model generically designed to optimize the procurement and implementation of new technologies within commercial and government institutions. As described in Chapter 1, the Practitioner's Cycles model is based on the principle that successful R2O transitions are the result of active engagement with end users through extensive collaboration to learn their operational needs, desires, and procedures. The Practitioner's Cycles, then, place great emphasis on the application of cognitive work analysis (Scott et al. 2005; Roth et al. 2006), or research that reveals and exploits the knowledge and strategies employed by end users during the course of their regular duties. Such analysis might include field studies, job shadowing, and structured interviews to discover leverage points where changes and improvements to a product could have significant impacts to operational efficiency and effectiveness. Where more traditional R2O cycles might focus on recursively performing research activities

and statistical evaluations, the Practitioner's Cycles strive to continuously produce and evaluate iterative prototypes in the real work environment irrespective of project milestones and schedules.

Etgar (2008) presents a descriptive model of the co-production process that can be modified to break the development cycle into five primary phases: the initiating phase, design phase, production phase, distribution phase, and evaluation phase. These phases are not independent of each other and may be performed concurrently during the development cycle. This descriptive model (summarized in Fig. 4.1) can be applied to the Practitioner's Cycles, and each phase is described in detail in the following subsections. To aid in this description, readers are encouraged to refer to Hoffman et al. (2010), their Fig. 1-5 for graphical depictions of the full Practitioner's Cycles.

4.1 Initiating Phase

The first step of collaborative co-production is to determine if co-production is truly the best development model for the task at hand. There are many scenarios in which the degree of structured collaboration required by this process may not be appropriate or desirable to meet a project's goals. For example, highly innovative technology or techniques that fall contrary to operational norms may face stiff resistance and push-back from those hesitant to drastically change their procedures or risk compromising operational systems (Deal and Hoffman 2010b). While structured collaboration would be appropriate and likely required to eventually transition such a product into operations, the user-centric approach necessitated by collaborative co-production risks the elimination of

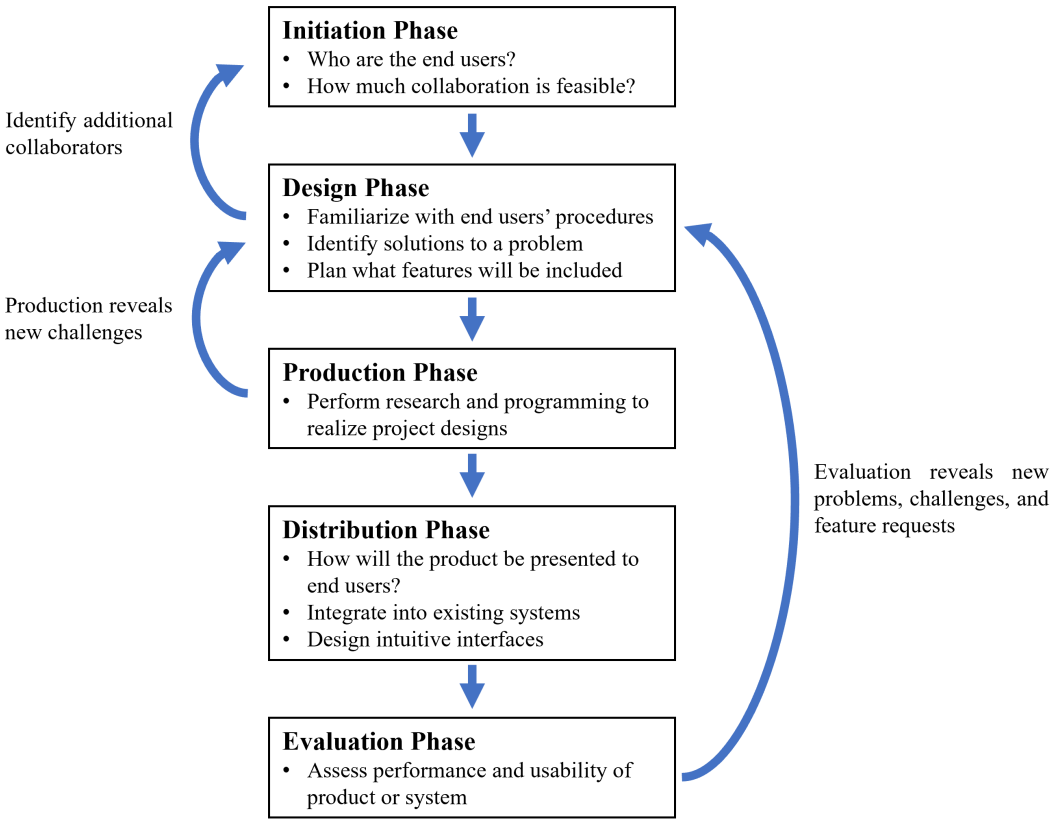


Figure 4.1: Schematic of the collaborative co-production process.

more novel or radical ideas in favor of maintaining the status quo. In some instances, it might be beneficial to first develop a basic prototype or mock-up in a pure research environment as a proof of concept before subjecting the product to the rigors of the co-production cycle. One should also consider the many environmental factors that might influence the effectiveness of collaborative co-production. Turnhout et al. (2020) describes unequal power relations between developers and end users as the primary cause of failure in co-production processes. These power imbalances ultimately limit or extinguish collaboration effectiveness and may derive from a variety of sources such as managerial pressures, regulations, contractual obligations, and workplace culture.

Once collaborative co-production is confirmed as the optimal design strategy, the next step in the initiating phase is to facilitate contact between developers and end users. This phase should identify all parties involved in the proposed collaboration, the level of collaborative participation requested of each party, and where prototype evaluation and testing will occur. The initiation phase can be considered a pre-step to the Practitioner's Cycles and is intended to establish the basic details required for later phases.

4.2 Design Phase

The purpose of the design phase is to collaboratively identify and plan the features and characteristics of the product to be produced (Etgar 2008). During this phase, Deal and Hoffman (2010a) and Hoffman et al. (2010) encourage developers to first become intimately familiar with the tasks, procedures, technologies, and limitations of the end users to better understand the role of the new product and the problems it is intended to address. This might be achieved

through extensive job shadowing, training, structured interviews, or other cognitive work analysis techniques. As an example, Deal and Hoffman (2010a) relay an anecdote of a successful R2O transition in which the developers were told to "Go there, get training, and come back when you have reached the point where the director would say he would let you do the actual job." This degree of familiarity is of course not practical in most collaborations, but active engagement with end users during this early stage is crucial for establishing a comprehensive understanding of the project goals.

The second step of the design phase is for developers and end users to jointly identify problems or needs within the existing system that would improve operational effectiveness. During early stages of development (Hoffman et al. 2010, their Fig. 1), each group might individually or collaboratively suggest potential solutions to these problems before designing research proposals, prototype mock-ups, and initial programs to explore these solutions. The design phase of the co-production model is recursive, and additional problems and needs will likely be identified as product development progresses. In later iterations of a project (e.g., Hoffman et al. 2010, their Fig. 2-4), the design phase may consist of identifying feature requests, interface issues, usability issues, and integration issues of a prototype system and proposing solutions to those problems. Even after a new product has fully transitioned into operations, developers may continue to collaborate with their end users to identify additional feature requests and bug fixes that can be implemented in minor and major updates (Hoffman et al. 2010, their Fig. 5).

4.3 Production Phase

The production phase is primarily concerned with the processes by which the product designs are realized into mock-ups, prototypes, and deployable systems (Etgar 2008). This stage of development may include extensive exploratory research, code creation, and graphic interface design to achieve a product that can be presented for evaluation to the end user. The production phase is perhaps the point in the co-production cycle when the relevant skills and expertise of the collaborative parties are the most distinct, as this step places great emphasis on the technical aspects of development. Nevertheless, there are many opportunities for continued collaboration during product invention that should be explored. For example, end users might participate in, contribute to, or even lead research activities performed over the course of production, particularly when that research falls within their expert domain knowledge. A co-production with the goal to produce a system that predicts tornadoes, for instance, might defer to the domain knowledge of operational forecasters when choosing which variables to include in the predictive model. Developers could also utilize this stage of development to familiarize end users with the methods and technologies being applied to produce the mock-up, prototype, or system. For example, a project that employs ML modeling might apply and demonstrate explainable AI techniques to help end users interpret how the model reaches an output. By including these collaborations during the production phase, developers can impart a sense of ownership to the end users and increase their familiarity with the resulting product prior to evaluation (Deal and Hoffman 2010b).

4.4 Distribution Phase

The next step in the co-production cycle is the distribution phase. This phase focuses on how products are provided to end users for evaluation, testing, and implementation. As described by Hoffman et al. (2010) and Deal and Hoffman (2010a,b), end users should ideally evaluate mock-ups, prototypes, and deployable systems in their native workplace to ensure valid results and reduce logistic challenges. As such, developers are encouraged to collaborate with their end users to identify and create processes by which these products can be evaluated in operationally representative environments. This might mean integrating a product into the end users' primary software or creating interfaces compatible with existing operational technologies. In cases where it is not feasible to deploy a prototype directly to an operational environment, testbeds or other experimental platforms may be utilized to simulate how a product might perform in operational conditions. For example, the NWS has long utilized the HWT (Kain et al. 2003) to host stakeholders as they explore, test, and evaluate experimental technologies in structured weeks-long sessions. The HWT was established on the principles that collaboration increases the speed and success of R2O transitions, benefits both the research and operational communities, and generates outcomes that are more reliable and useful for society (Calhoun et al. 2021). Activities performed within the testbed are designed to mimic operations within a NWS WFO or national center, and experiments are provided the opportunity to leverage live and archived weather data across the United States to assess performance in varied circumstances. One caveat to organized testbeds is that

they tend to meet infrequently, and it might be difficult to maintain the degree of continuous collaboration championed by the collaborative co-production conceptual model.

4.5 Evaluation Phase

The evaluation phase is the point at which developers and end users directly assess the performance and usability of a product or system. In the atmospheric sciences, a product's performance often refers to the goodness of a forecast or analysis as described in section 2.2. This aspect of the evaluation phase frequently relies on the application of verification metrics to identify a product's quality, value, and limitations. Other measures of performance might evaluate technical considerations such as required computing resources, runtime, or cost effectiveness. These metrics are updated and compared with each iteration of development and are useful for measuring progress toward project goals.

The second aspect of evaluation is perhaps less familiar to traditional researchers and developers but equally important to a product's assessment. Stated simply, usability refers to how well end users are able to interact with and apply a product to achieve an objective. Hoffman et al. (2010) and Deal and Hoffman (2010a,b) place great emphasis on continuous, coordinated collaboration between developers and end users to assess a product's usefulness, usability, and understandability throughout the development cycle. As during the design phase, these evaluations might require the application of cognitive work analysis techniques to qualitatively and quantitatively assess product usability. One such technique, a use test, is a process by which developers watch end users interact with a product with the intention of making that product easier to use

(Krug 2009). In this process, developers become facilitators as they encourage constructive feedback from end users and apply that feedback to create a better product. As described by Hoffman et al. (2010), product usability is a core tenet of co-production and should be a major consideration at each stage of the development cycle.

Finally, it is important to understand that although evaluation is the last phase of this discussion, it is not the final phase of development. Instead, the evaluation phase might be considered the beginning of the next iteration of development as the collaborative results obtained from product assessment fold back into the next design phase. End users might reveal a prototype's brittleness in varying circumstances or discover interface, usability, or integration issues that ultimately reduce the effectiveness of the product. These issues are then passed to the next design phase, resolved during production of the next prototype, and then evaluated again until the product finally achieves operational status.

Chapter 5

Lessons from an R2O Success

In the previous chapter, collaborative co-production was presented as a user-centered production cycle that has the potential to improve the success rate of R2O transitions. However, co-production is ultimately an idealized model with many potential limitations to practical real-world application. To assess the feasibility of this proposed paradigm in a high-impact operational environment, I collaborated with expert forecasters and management at the National Weather Service's Storm Prediction Center to apply co-production principles toward the development and operational implementation of a new suite of calibrated thunderstorm forecast guidance using traditional (non-ML) methods. This chapter (modified from Harrison et al. 2022) chronicles that development and implementation process.

5.1 Background

The first step of applied co-production was to identify an operational need that could be filled through collaborative development. The stated mission of the SPC is to deliver timely and accurate forecast information about tornadoes, severe thunderstorms, lightning, wildfires, and winter weather across the contiguous United States (CONUS) to protect lives and property (SPC 2021a). As

part of this mission, the SPC is responsible for issuing forecast products that indicate where and when cloud-to-ground (CG) lightning is anticipated. One such product is the Thunderstorm Outlook, which depicts the probability of thunderstorms across the CONUS in 4- or 8-hour periods (Bright and Grams 2009; SPC 2021b) for the upcoming or current convective day. Specifically, these forecasts represent the probability of at least one CG lightning flash within 20 km (12 miles) of a point location during the valid forecast period. The increased temporal resolution of the Thunderstorm Outlook aids NWS forecasters and partners in time-sensitive decisions related to thunderstorms and lightning hazards (Stough et al. 2012; SPC 2021b).

Accurately predicting the timing and location of thunderstorms across the CONUS can often be a time consuming and mentally taxing challenge for forecasters. Many studies have been published over the past five decades showcasing a variety of automated, gridded thunderstorm probability guidance intended to aid in the prediction of lightning hazards. One of the earliest of these studies dates to the 1970s, when Reap and Foster 1979 created a multiple screening regression to generate medium-range thunderstorm probability forecasts from Model Output Statistics (MOS; Glahn and Lowry 1972). More recent approaches to probabilistic lightning prediction have been incorporated within the NWS's National Blend of Models (NBM), a project intended to generate calibrated, high-resolution forecast guidance from statistically postprocessed multi-model ensembles (Tew et al. 2016; Hamill et al. 2017; Craven et al. 2018). Probabilistic lightning forecast products currently contained within the operational NBM include MOS guidance from the Global Forecast System (GFS; Kanamitsu et al. 1991, Hughes 2001), the North American Mesoscale Forecast

System (NAM; Rogers et al. 2005, Maloney et al. 2009), and the European Centre for Medium-Range Weather Forecasts (ECMWF; Shafer and Rudack 2015) model. MOS thunderstorm forecasts in the NBM are primarily derived from a regression of deterministic large-scale numerical weather prediction (NWP) precipitation forecasts and lightning climatology in 3-hour intervals. Although the GFS, NAM, and ECMWF MOS thunderstorm forecasts have demonstrated skill, their dependence on large-scale NWP forecasts ultimately limits the spatial and temporal detail of the guidance. Additionally, the reliance of MOS schemes on climatology tends to reduce the forecast skill of spatiotemporally infrequent lightning events (Shafer and Fuelberg 2008).

The NBM also includes probabilistic lightning forecasts from the Localized Aviation MOS Program (LAMP; Charba et al. 2019), which combines the aforementioned deterministic large-scale MOS products with fine-scale model output from the deterministic High Resolution Rapid Refresh (HRRR; Benjamin et al. 2016) model, total lightning observations from the National Lightning Detection Network (NLDN; Cummins et al. 1998), and radar reflectivity from the Multi-Radar Multi-Sensor (MRMS; Smith et al. 2016, Zhang et al. 2016) system. Objective verification of the LAMP by Charba et al. (2019) has shown the guidance performs very well in the first forecast hour via extrapolation of MRMS and NLDN observations. However, the influence of observations on the guidance was found to sharply decrease within the first four forecast hours. As such, much of the LAMP's forecast skill comes from shorter lead-time forecasts, with notably decreasing skill at longer lead times.

A fifth lightning forecast product contained within the NBM is derived by SPC through post-processing the 26-member National Centers for Environmental Prediction (NCEP) Short-Range Ensemble Forecast (SREF; Du et al. 2014).

For nearly two decades, SPC forecasters have largely utilized the SREF Calibrated Thunder guidance (SREFCT) as their "first guess" when generating Thunderstorm Outlooks and other thunderstorm forecast products. The SREFCT aids in delineating areas favorable for CG lightning by identifying regions of appropriate instability and thermodynamic factors coinciding with precipitation within the SREF forecast grid (Bright et al. 2005). Specifically, the SREFCT highlights points within the SREF's 40-km NCEP 212 grid where the forecast LCL $\geq$ -10º C, CAPE $>$ 100 Jkg-1 in the 0º to -20º C layer, and the equilibrium level temperature is $\leq$ -20º C. As described by Bright et al. (2005), these parameters are believed to approximately delineate regions where mixed-phase hydrometeors are present above the charge-reversal temperature and coincide with updrafts sufficiently strong enough to replenish supercooled liquid above the charge-reversal zone (Saunders 1993). The SREFCT probability of at least one CG lightning flash within 20 km (12 miles) of a point location is derived by first combining the above environmental parameters into a single probabilistic ensemble composite known as the Cloud Physics Thunder Parameter (CPTP; Bright et al. 2005):

$$CPTP = \frac{(-19^{\circ}\ C - T_{EL})(CAPE_{-20} - K)}{K} \tag{5.1}$$

where T_{EL} is the equilibrium level temperature, $CAPE_{-20}$ is the CAPE between the 0º C to -20º C levels, and K is a constant set to 100 Jkg^{-1}. The calibrated lightning probability is then obtained by calculating the relative frequency of CG flashes observed by the NLDN given the predicted CPTP probability and the SREF ensemble probability of accumulated precipitation $\geq$ 0.01 in. (0.254 mm) during the valid forecast period.

Verification studies of the SREFCT (e.g., Bright and Grams 2009) have shown that the algorithm provides reliable and skillful guidance, particularly as a "first-guess" product for forecasters when producing Thunderstorm Outlooks. A first-guess guidance typically provides an initial automated approximation of a forecast or other official product which human forecasters can then modify to better align with their expert assessment and criteria. Such guidance is generally intended to help forecasters quickly analyze and process complex data, highlight areas of interest within a forecast domain, and reduce the amount of time spent on the technical aspects of drawing and submitting operational products. However, the SREFCT guidance was found to exhibit a notable bias toward under-forecasting CG lightning probabilities in the Plains and Gulf Stream, especially during the warm season when nocturnal convection is more prevalent. Forecasters have also noted a tendency to over-forecast lightning potential along the California coast and in the Pacific Northwest. Bright and Grams (2009) speculate that these biases may be in part due to the SREF's inability to explicitly resolve convection. For example, over the ocean it is common for the SREFCT's thermodynamic parameters to be met despite a dearth of convective precipitation. In these cases, the guidance may undesirably generate probabilities for grid-scale precipitation originating from low clouds in the model's marine boundary layer (Bright and Grams 2009).

Given the apparent limitations of the SREFCT and other probabilistic lightning guidance, I hypothesized that the addition of simulated radar reflectivity and other storm-attribute fields from an ensemble of convection-allowing models (CAMs) may lead to improved probabilistic thunderstorm predictions. This idea was proposed to SPC forecasters with great interest, and I was invited to shadow in SPC operations to observe and learn the processes by which the

Thunderstorm Outlooks are created. By spending dozens of hours shadowing experts and even creating my own experimental forecasts, I was able to witness the aforementioned shortcomings of the existing SREFCT guidance in an operational environment. With this crucial step of the co-production cycle complete, I began collaboration with SPC forecasters and management to design and test a new suite of probabilistic thunderstorm guidance products derived from the NCEP High-Resolution Ensemble Forecast (HREF; Roberts et al. 2019) system.

5.2 Data and Methods

The HREF Calibrated Thunder (HREFCT) forecast products were derived using prognostic fields from the operational HREF version 2 (HREFv2) and an experimental version of the HREF (HREFv2.1) that was tested internally at SPC (Roberts et al. 2020). The HREFv2 is composed of eight ensemble members with four deterministic CAM configurations represented by the High-Resolution Window Advanced Research version of the Weather Research and Forecast Model (HRW ARW; Skamarock et al. 2008), the National Severe Storms Laboratory version of the ARW (HRW NSSL; Kain et al. 2010), the Nonhydrostatic Multiscale Model on the B Grid (HRW NMMB; Janjić and Gall 2012), and the 3-km NAM Nest (Rogers et al. 2017). Each configuration is represented twice within the HREFv2 ensemble by including 12-hour time-lagged initializations of each member. A full list of the model cores, boundary conditions, microphysics schemes, and PBL schemes of each of the eight members is provided by Roberts et al. (2019), their Table 1.

The HREFv2.1 utilized in this study adds the operational HRRR and its 6-hour time-lagged run, giving the ensemble a total of 10 members for the

first 30 forecast hours, 9 members through forecast hour 36, and 5 members through forecast hour 48. The inclusion of the HRRR has been shown to increase member spread and improve the overall skill of the ensemble (Gallo et al. 2018). Note that as of 11 May 2021, the operational version of the HREF version 3 (HREFv3) replaces the NMMB member of the ensemble with a Finite Volume Cubed Sphere (FV3) model and extends the temporal range of the HRRR to 48 hours (EMC 2021; NWS 2021b). However, these changes were not available at the time of this study and thus were not included during the initial design and testing. This is discussed further in section 5.3.

Both 00z and 12z cycles of the HREFv2 were obtained for 1 July 2017 - 1 January 2019, and the HREFv2.1 cycles were collected for 1 January 2019 - 11 May 2021 (the full period available). The HREF is natively produced on a 3-km grid; however, SPC Thunderstorm Outlooks are verified on the 40-km NCEP 212 grid. To ensure the HREFCT forecast probabilities remain consistent with those being issued by the SPC forecasters, all prognostic fields within the HREF ensemble members were remapped to the 212 grid using a nearest neighbor maximum, minimum, or average, depending on the variable (Mesinger et al. 1990; Mesinger 1996; Accadia et al. 2003). For example, a nearest neighbor minimum was used to remap lifted index forecasts because lower values indicate increased instability, while a nearest neighbor maximum was used to remap 1-hour accumulated precipitation. Additionally, observed hourly CG lightning flashes were obtained from the NLDN for the 1 July 2017 - 11 May 2021 period and spatially mapped to the same 40-km grid.

5.2.1 The HREFCT Algorithm

Initial development of the HREFCT attempted to build upon the success of the SREFCT by first focusing on the same environmental parameters used in the existing guidance. However, recreating the SREFCT's algorithm within the HREF framework quickly proved unsuccessful as some HREF members lack the fields necessary to compute the equilibrium level temperature or CAPE in the 0^{o} C to -20^{o} C layer. Instead, the HREFCT guidance was derived from scratch, with the first step to identify which HREF prognostic fields best correlate to the occurrence of at least one CG lightning flash. To accomplish this, the first 24 forecast hours of all 00z and 12z HREF cycles from 1 July 2017 - 1 July 2019 were compared to the corresponding NLDN gridded CG lightning observations. A Pearson correlation coefficient was then computed between the CG flash observations and all prognostic fields common across the HREF members. The resulting correlations from each member were averaged to provide an ensemble mean correlation for each field. Total accumulated QPF was found to have the highest mean correlation to at least one CG lightning flash, with a value of 0.14. Most unstable CAPE (MU CAPE) had the second highest mean correlation, followed by the derived radar reflectivity at -10^{o} C, derived radar reflectivity at 4 km above ground level, maximum 1-hour composite reflectivity, precipitable water, specific humidity, and the most unstable 4-layer lifted index (MU LI). Correlations for each field are provided in Table 5.1.

Once the best correlated prognostic fields were identified, the next step was to develop an algorithm to convert the data into a probabilistic thunderstorm forecast. Several regression analyses utilizing various combinations of the aforementioned fields were tested and subjectively evaluated by SPC forecasters during this stage. Hourly probabilistic thunder forecasts were created from each

Variable	Ensemble mean correlation
Total accumulated QPF	0.14
MU-CAPE	0.13
Derived radar reflectivity at -10$^{\text{o}}$ C	0.12
Derived radar reflectivity at 4 km AGL	0.12
Maximum composite reflectivity	0.11
Precipitable water	0.10
Specific humidity	0.07
MU LI	0.06

Table 5.1: HREF prognostic fields with the greatest ensemble mean Pearson correlation to 1-hour NLDN CG lightning flashes computed between 1 July 2017 and 1 July 2019.

algorithm/input combination for the first 24 forecast hours from the 00z and 12z HREF cycles between 1 July 2017 and 1 July 2019. The mean critical success index (CSI) was computed for each 1-hour forecast, and this process was iteratively repeated until the best combination (i.e., the combination with the greatest mean CSI and subjective forecaster approval) of algorithm and prognostic fields was determined. Ultimately, the derived radar reflectivity at -10 $^{\text{o}}$C ($Z_{\text{-10C}}$), total accumulated QPF (QPF_{accum}), and MU LI were found to be the most successful combination of prognostic fields when paired with a linear regression model of the form:

$$w_1 P(X \geq t_1) + w_2 P(Y \geq t_2) + w_3 P(Z \geq t_3) \tag{5.2}$$

Here, w_1, w_2, and w_3 represent weights summing to 1; X, Y, and Z are HREF prognostic fields; and t_1, t_2, and t_3 are threshold values corresponding to the respective HREF fields. The probability function $P()$ is defined as the fraction of

HREF ensemble members where the inequality is true. As an example, consider a single grid point where five of the ten HREF members predict Z_{-10C}will be greater than a threshold of 40 dBZ over a given 1-hour period. Then $P(Z_{-10C} \geq 40\ dBZ) = 5/10 = 50\%$. The probability of lightning predicted by the algorithm for a given grid point is then the weighted average of the probabilities that each prognostic field meets or exceeds its respective threshold value.

The final step in the initial derivation was to determine which combination of weights and thresholds provide the optimal forecast. This was accomplished by performing a randomized grid search (Bergstra and Bengio 2012), where thunder forecasts were again computed for 1 July 2017 - 1 July 2019 using a random subset of every possible combination of weights and thresholds. The combination of hyperparameters that resulted in the greatest forecast CSI was selected as the optimal configuration. This optimization was performed independently for rolling forecast windows of 1- and 4-hour intervals, where the 4-hour prediction at a given forecast hour represents the cumulative probability of at least one CG flash within 20 km (12 miles) of a point location over the previous 4 hours. For example, the 4-hour forecast at forecast hour f04 represents the f00 - f04 period. Additionally, 24-hour forecasts were generated and optimized for each convective day (12z - 12z) contained within each HREF cycle (00z HREF f12 - f36; 12z HREF f00 - f24 and f24 - f48.) HREFCT forecasts for intervals greater than one hour were computed using the maximum or minimum values of each input variable over the specified period. The best weights and thresholds for each forecast interval are provided in Table 5.2.

	$Z_{\text{-10C}}$; $Z_{\text{4 km AGL}}$	QPF_{accum}	MU LI
1-hour forecast	$t_1 \geq 40\ dBZ$	$t_2 \geq 1\ mm$	$t_3 \leq -3$
	$w_1 = 0.6$	$w_2 = 0.3$	$w_3 = 0.1$
4-hour forecast	$t_1 \geq 40\ dBZ$	$t_2 \geq 2\ mm$	$t_3 \leq -1$
	$w_1 = 0.6$	$w_2 = 0.3$	$w_3 = 0.1$
24-hour forecast	$t_1 \geq 40\ dBZ$	$t_2 \geq 2\ mm$	
	$w_1 = 0.6$	$w_2 = 0.4$	

Table 5.2: The best thresholds (t) and weights (w) for each HREF prognostic field and forecast time interval. MU LI was excluded from the 24-hour forecast due to strong diurnal variations in the parameter.

Though the HREF parameters are not identical to those used in the SREFCT algorithm, the fields and thresholds chosen for the HREFCT formula capture many of the same environmental conditions and physical processes highlighted by the original guidance. For example, the thresholds chosen for MU LI broadly indicate where lapse rates may be steep enough to support sustained updrafts necessary to replenish supercooled liquid above the charge-reversal zone, and $Z_{\text{-10C}}$ might be considered an approximation of mixed-phase hydrometeors present near or above the charge-reversal temperature. Note that some members of the HREFv2 did not initially provide $Z_{\text{-10C}}$, so the derived radar reflectivity at 4 km above ground level ($Z_{\text{4 km AGL}}$) was used as a proxy with the same weights and thresholds prior to the implementation of the HREFv2.1. The improved performance of MU LI over MUCAPE was an unexpected result during the algorithm derivation, as MUCAPE exhibited much higher correlation to CG flashes (Table 1). Anecdotally, forecasts that utilized MUCAPE instead of MU LI tended to over-forecast the spatial coverage of lightning, particularly

in marginally unstable or capped environments. Perhaps the ability of MU LI to represent both stable and unstable environments as a single parameter gives it an advantage over MUCAPE in the HREFCT algorithm, as MUCAPE would need to be paired with another variable such as MUCIN to provide information about stable layers within the thermodynamic profile.

5.2.2 Calibration

Calibration of the HREF probabilistic thunder guidance to be statistically reliable was performed by first generating thunder forecasts from 13 June 2019 through 13 June 2020. These dates were chosen for calibration to avoid inconsistencies in the HREF members that were present during the initial transition from HREFv2 to HREFv2.1. Noise in the raw probability fields was removed by applying a 2D Gaussian filter (σ=80 km) to spatially smooth the forecasts. The smoothed probabilities from the 1-year period were then stratified into 10% bins centered on every 10% (5 - 15%, 15 - 25%, etc.), and the reliability of each bin at each grid point was computed (Fig. 5.1a). For example, at a given grid point, the true probability of the 40% bin was defined as the fraction of 35 - 45% forecasts that verified with at least one observed CG lightning flash. If a given grid point received 40% probability forecasts 100 times throughout the year and lightning occurred at that grid point in 30 of those forecasts, then the true probability was 30% and the 40% bin had a reliability error of +10% (an over-forecast). The reliability error was then recorded for each grid point. This process was performed independently for the 00z and 12z HREF cycles, the 1-hour, 4-hour, and 24-hour forecast products, and for each of the HREF's 48 forecast hours. Thus, the calibration step resulted in a 5-dimensional lookup

table containing the mean reliability error at every grid point for every forecast hour, HREF cycle, and binned forecast probability.

This calibration process revealed a systematic bias in the uncalibrated HREFCT probabilities that generally led to an over-forecast of CG lightning flashes in the 1- and 4-hour forecast products and an under-forecast in the 24-hour product (Fig. 5.1b). The 1-hour probabilities exhibited a consistent mean reliability error of about +5% through at least the first 24 forecast hours, while the 4-hour uncalibrated guidance had an error of +5 to +10%. The 24-hour uncalibrated guidance averaged an under-forecast of -5 to -10%. The reliability errors of the 1- and 4-hour forecasts varied considerably in the last 18 forecast hours, likely due in part to predictability error in the spatial placement of convection at longer lead times. Forecast probabilities between 25% and 55% exhibited the greatest mean reliability error for all three products (not shown). Both the 1- and 4-hour uncalibrated guidance averaged an over-forecast of +10 to +15% at these probabilities, while the 24-hour guidance under-forecast by up to -5% on average. Notably, all three products were found to slightly under-forecast at probabilities $< 5\%$ and over-forecast at probabilities $> 95\%$ on average. As such, the resulting calibration tends to move the final forecast probabilities away from these extremes. Because of this, the calibration rarely changes the areal coverage of HREFCT probabilities or zeros out probabilities in the uncalibrated guidance.

Calibration is applied to new thunder forecasts by first matching the grid point, forecast hour, HREF cycle, and initial forecast probability to the corresponding reliability error in the lookup table. Once this is determined, the guidance is calibrated by simply subtracting that reliability error from the original forecast probability. For example, if at a given grid point the pre-calibrated

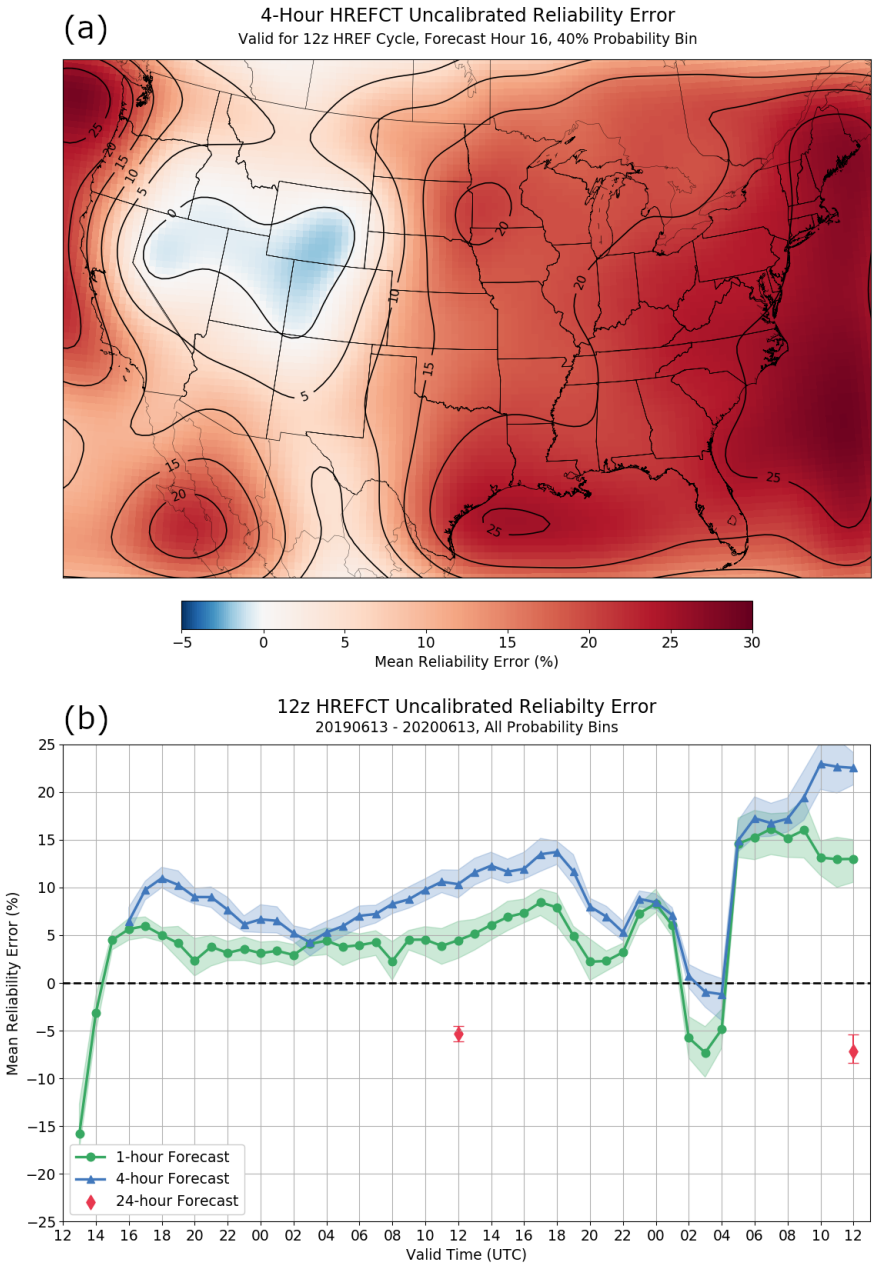


Figure 5.1: (a) Mean uncalibrated reliability error of the 12z HREFCT 4-hour lightning probabilities at forecast hour 16 for the 40% probability bin. Positive values represent an over-forecast compared to NLDN observations from 13 June 2019 - 13 June 2020. (b) Mean uncalibrated reliability error of the 12z HREFCT as a function of lead time.

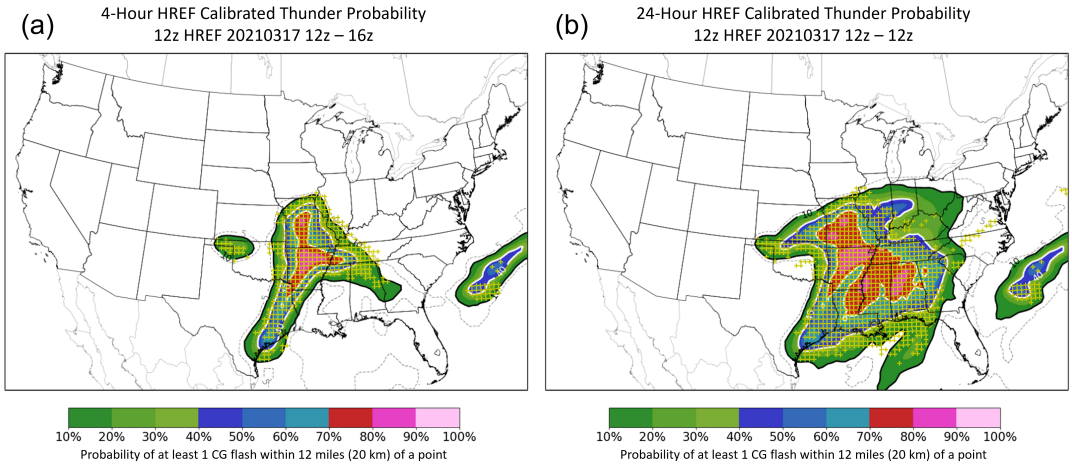


Figure 5.2: (a) HREFCT (a) 4-hour and (b) 24-hour forecasts from the 12z HREF cycle on 17 March 2021. Yellow "+" symbols indicate grid points where there was at least one CG lightning flash detected during the valid forecast period.

guidance had a probability of 44% and the mean reliability error for the 40% bin at that grid point at that forecast hour was +10% (an over-forecast by 10%), then the final, calibrated probability for that grid point would be 44% - 10% = 34%. Note that the calibration is only applied to points where probabilities already exist in the grid. As such, the calibration cannot introduce probabilities where there were none before the calibration step. An example 4-hour and 24-hour calibrated thunderstorm forecast from 17 March 2021 is shown in Fig. 5.2.

5.2.3 Instability and Reflectivity Mask

During iterative testing of the HREFCT guidance, SPC forecasters identified a bias in the algorithm that would result in the prediction of thunder probabilities

for locations that were subjectively analyzed to be unsupportive of deep convection or lightning. This most commonly occurred when several HREF members predicted moderate stratiform precipitation which would activate the QPF_{accum} term of the HREFCT equation. Although such relatively stable environments might fail to meet the thresholds for Z_{-10C} and MU LI, a sufficiently large fraction of HREF members predicting moderate accumulated precipitation values could still generate lightning probabilities.

To correct this bias, a filter was imposed on each member of the HREF to create a simple instability and reflectivity mask. The contribution from any HREF member that forecasts MU LI > 0 and $Z_{-10C} < 35$ dBZ over the valid forecast period is set to zero when creating the probabilities for a given grid point. As an example, consider a grid point where eight of the ten HREF members predict stratiform precipitation with a maximum Z_{-10C} of 30 dBZ and a 4-hour QPF_{accum} of 0.25 in. (6.35 mm). Only one of the ten members predicts MU LI < 0, while the others are all > 0. Without the instability mask, this grid point would be given a 25% probability of thunder, largely driven by the accumulated precipitation term. With the mask applied, however, all but one member would be set to zero in the calculation because the predicted reflectivity is < 35 dBZ and the MU LI is > 0. This would then produce a thunder probability of 4% for the grid point prior to calibration.

Calibrated 1-hour, 4-hour, and 24-hour thunder forecasts were regenerated for 13 June 2019 - 13 June 2020 with the new mask applied, and preliminary verification revealed a slight improvement in the bulk performance of the guidance (not shown). Furthermore, anecdotal case studies and real-time application by SPC forecasters found that the mask was successful at removing most non-meteorological regions of low thunder probabilities, particularly in

the Pacific Northwest and in stratiform precipitation regions of extratropical cyclones. This addition to the HREFCT algorithm highlights a key benefit of the collaborative co-production process, as the need for an instability and reflectivity mask was only discovered through frequent iterative evaluation in an operational environment. Without direct input from SPC forecasters, these edge cases of the guidance may not have been discovered until much later in the development cycle, if at all. Additionally, informal conversation with the forecasters revealed that including the mask at their recommendation anecdotally increased operational buy-in and trust in the final product. All discussion of the HREFCT hereafter refers to the HREFCT with the instability and reflectivity mask applied.

5.3 Results and Discussion

Verification of the 00z and 12z HREFCT forecast products was performed on the 11-month independent dataset of 13 June 2020 - 11 May 2021. Calibrated 1-hour, 4-hour, and 24-hour thunder forecasts were generated for the full verification period, and the probabilities from each forecast were stratified into 10% bins as during calibration. The forecasts were then compared to the observed NLDN CG lightning flashes for each forecast hour, and the POD, FAR, CSI, and statistical reliability were computed for each probability bin. Additionally, 95% confidence intervals for the metrics were computed from 10,000 bootstrapped samples. The following discussion will focus on the 12z HREFCT guidance, but similar results were noted for the 00z guidance as well.

Verification of the HREFCT generally improved as the valid forecast window increased (Fig. 5.3a). The 24-hour forecast product was found to have the

greatest performance over the verification period with a maximum CSI of 0.43 (0.41 - 0.45) at the 40% probability bin. The 4-hour forecast product exhibited the next best performance with a maximum CSI of 0.28 (0.26 - 0.30) at 30%, and the 1-hour HREFCT guidance had the lowest average performance with a maximum CSI of 0.19 (0.17 - 0.20) at 20%. All three forecast products were found to be statistically reliable over the verification period, but the 24-hour forecast tended to under-forecast the observed CG lightning flashes by about 5 - 10% at forecast probabilities $\geq$ 40% (Fig. 5.3b). In contrast, the 1-hour and 4-hour forecasts were, on average, reliable within 5% of observations at all probability levels. Note that the 1-hour HREFCT guidance rarely predicted probabilities $\geq$ 75% during the verification period (Fig. 5.3c), and so the higher bins were excluded when calculating the performance and reliability of the product.

The observed tendency of the 1-hour guidance to produce lower probabilities than the 4-hour or 24-hour guidance is largely a result of forecast uncertainty manifested by the spread of the HREF members. As described in section 2, the HREFCT algorithm is a linear combination of probability functions. These probability functions describe not only how favorable the predicted environment is for lightning, but also the ensemble uncertainty that those conditions will be met. Generating high probabilities in a 1-hour forecast requires an equally large number of HREF members to meet the forecast thresholds at the same grid point and forecast hour (i.e., large ensemble agreement.) Conversely, the 4-hour guidance only requires the HREF membership to meet those forecast thresholds at any time during the valid 4-hour interval. As such, it is generally easier to achieve higher probabilities in the 4-hour (and 24-hour) guidance than in the 1-hour guidance. This holds true at shorter lead times too, as many

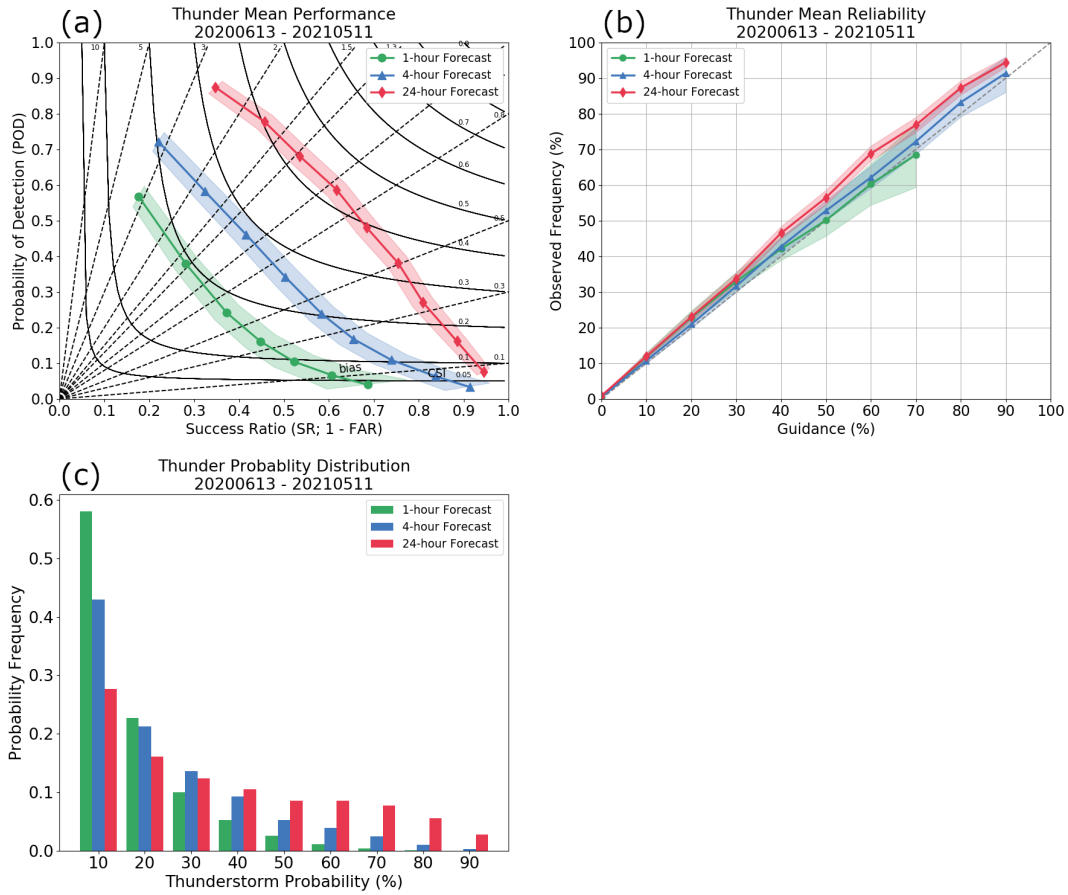


Figure 5.3: 12z HREFCT 1-hour, 4-hour, and 24-hour (a) mean performance, (b) mean reliability, and (c) forecast probability frequency for 20200613 - 20210511. The shaded regions represent the 95% confidence intervals from 10,000 bootstrapped samples.

HREF members utilize different dynamical cores, initial and boundary conditions, microphysics schemes, and PBL schemes that lead to differing solutions early in the forecast cycle.

Additional nuance in the HREFCT's performance was revealed by analyzing a spatial and temporal breakdown of the 4-hour guidance. The average reliability error of the product over the verification dataset was produced at each point within the NCEP 212 grid and across six 4-hour intervals as shown in Fig. 5.4. Despite calibration, the HREFCT continued to exhibit a tendency to over-forecast CG lightning probabilities on average across most of the CONUS and at most hours of the day. Even so, this error was typically within 5 - 10% of observations and much improved from the 20 - 25% error noted in the uncalibrated guidance (Fig. 5.1). These results serve as an example of how an underdispersive or overconfident ensemble may be corrected by applying calibration to reduce probabilities at most locations and times (Raftery et al. 2005; Berrocal et al. 2007; Kann et al. 2009). More notable over-forecasting was observed along and east of the Appalachians between 04z – 16z, with reliability errors of 10 - 15% common across that region. Other forecast biases include a broad area of +10% reliability error along the Gulf Coast from 00z - 04z, a small area of +10% error in the central Plains from approximately 08z - 20z, and a slight under-forecast of up to 5% across the Southwest from 16z - 04z.

These regional biases may be at least partially attributable to systematic error in the underlying HREF forecast. The over-forecast region in the central Plains, for example, anecdotally correlates to the approximate time and location of a number of MCS events that occurred during the 2020 warm season. Although the ability of CAMs to predict MCS events has improved over recent years, some HREF members such as the HRRR have been shown to commonly

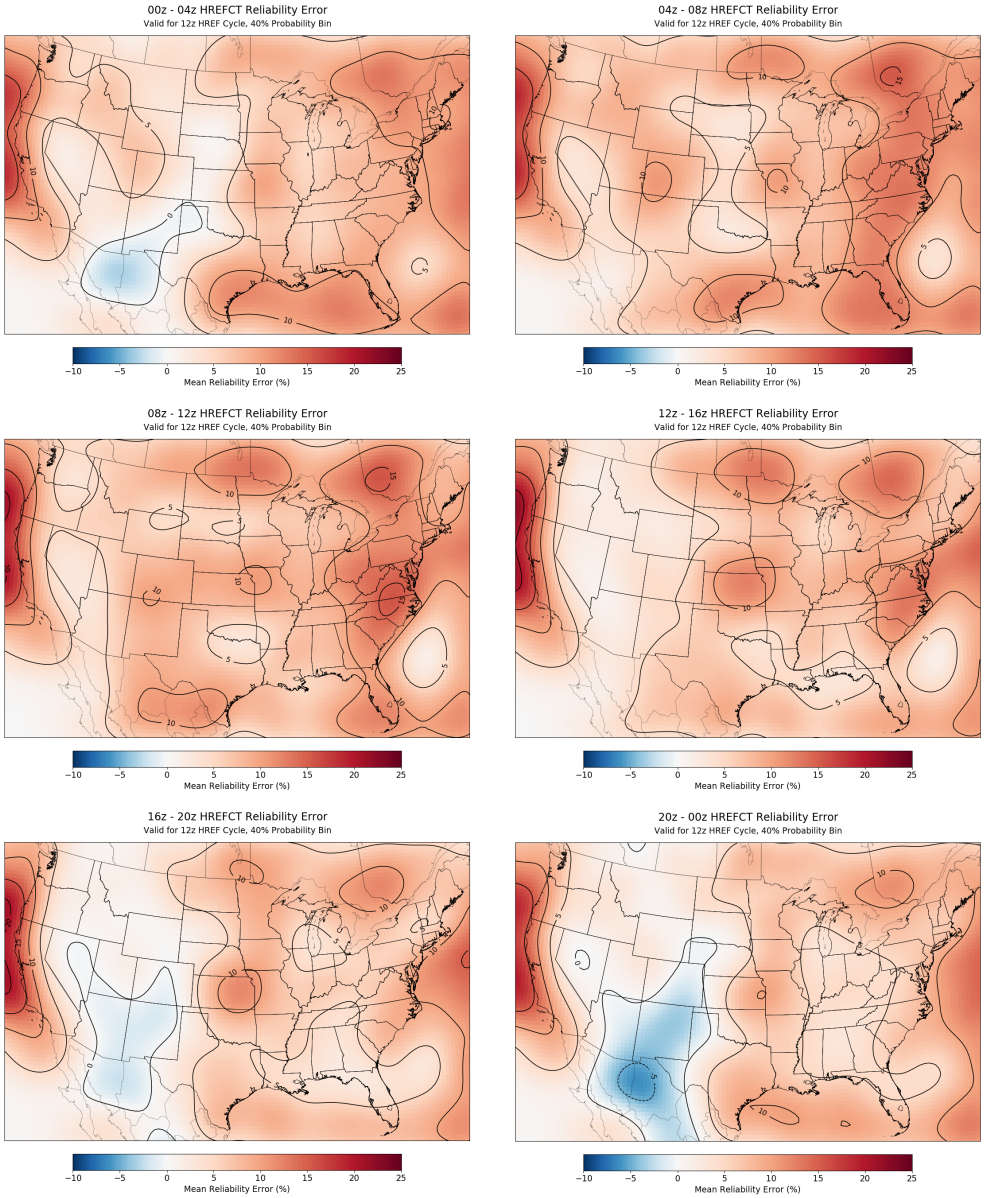


Figure 5.4: 12z HREFCT mean spatial reliability error across six 4-hour periods. Positive values (warmer colors) represent an over-forecast and negative values (cooler colors) represent an under-forecast. The reliability error was calculated for the 40% probability bin from 20200613 - 20210511.

over-forecast MCS convection in the Plains during the overnight hours (Clark et al. 2007; Pinto et al. 2015). As such, the HREFCT may have over-forecast the probability of CG lightning because some members of the HREF consistently predicted too many MCS events in that region. One important caveat to this analysis is that the relatively short 11-month verification period potentially makes the reliability error at any given location sensitive to a small number of events. For instance, there is a consistent area along the Pacific coast where the guidance over-forecast the lightning potential by up to 25% on average. However, the sample size of forecasts in that region is extremely limited, and so these results may not be fully representative of the longer-term performance of the HREFCT in that area. These sample-size limitations are also applicable to the underlying HREFCT calibration, which was necessarily performed on just one year of data. Recalibration and verification on multiple years of forecasts is planned for future updates to the operational guidance as a longer period of record becomes available.

A primary goal when developing the HREFCT was to improve upon the existing SREFCT guidance. As such, it was necessary demonstrate to SPC forecasters how the new HREFCT performance compared to that of the original SREFCT. One-hour, 4-hour, and 24-hour forecasts from the SREFCT were regenerated for the 20200613 - 20210511 verification period, and the POD, FAR, CSI, and reliability were computed as before (Fig. 5.5). There is no 12z SREF cycle to directly compare against the 12z HREF, so the 09z and 15z SREF cycles were used instead. The 09z 4-hour SREFCT exhibited a maximum CSI of 0.20 (0.19 - 0.21) at the 20% probability bin, while the 15z SREFCT had a maximum CSI of about 0.21 (0.19 - 0.21) also at 20%. This performance was notably less than the 4-hour HREFCT's maximum CSI of 0.28 (0.26 - 0.30) at

the 30% bin. Unfortunately, several months of missing data in the local SREFCT archive prevented a direct comparison of the 1-hour and 24-hour products. However, indirect comparisons using the incomplete dataset reveal a similar improvement in the HREFCT guidance over that of the SREFCT (not shown). Despite the notable improvements in CSI, the SREFCT and HREFCT products exhibited similar statistical reliability (Fig. 5.5). The 4-hour forecasts for the 09z and 15z SREFCT tended to slightly under-forecast probabilities $< 25\%$ while slightly over-forecasting probabilities $\geq 35\%$. In contrast, the HREFCT tended to slightly under-forecast CG lightning potential at all probabilities. Additionally, the forecast probability distribution of the HREFCT was notably shifted toward higher probabilities compared to that of the SREFCT (Fig. 5.5c). Approximately 10% of HREFCT 4-hour forecast values exceeded 55% during the verification period compared to only 2% of the 09z and 15z SREFCT forecasts. In fact, the SREFCT 4-hour forecast rarely exceeded 65% during the verification period. This demonstrated ability of the HREFCT to produce higher, statistically reliable forecast probabilities is a noteworthy improvement over the SREFCT.

Next, the mean CSI and reliability error of the SREFCT and HREFCT 4-hour forecasts were computed at each forecast hour to reveal diurnal and lead time variations in the guidance (Fig. 5.6). Both the SREFCT and HREFCT 4-hour forecasts achieved their greatest mean CSI at approximately 23z on day 1 (12z HREF f11; 09z SREF f14; 15z SREF f08). As before, the HREFCT showed marked improvement over the SREFCT with a mean CSI of about 0.31 (0.30 - 0.32) compared to 0.24 (0.22 - 0.25) for the 09z and 15z SREFCT forecasts. After a steady decline in mean performance between approximately 01z and 18z, all versions of the guidance then experienced a secondary peak at about

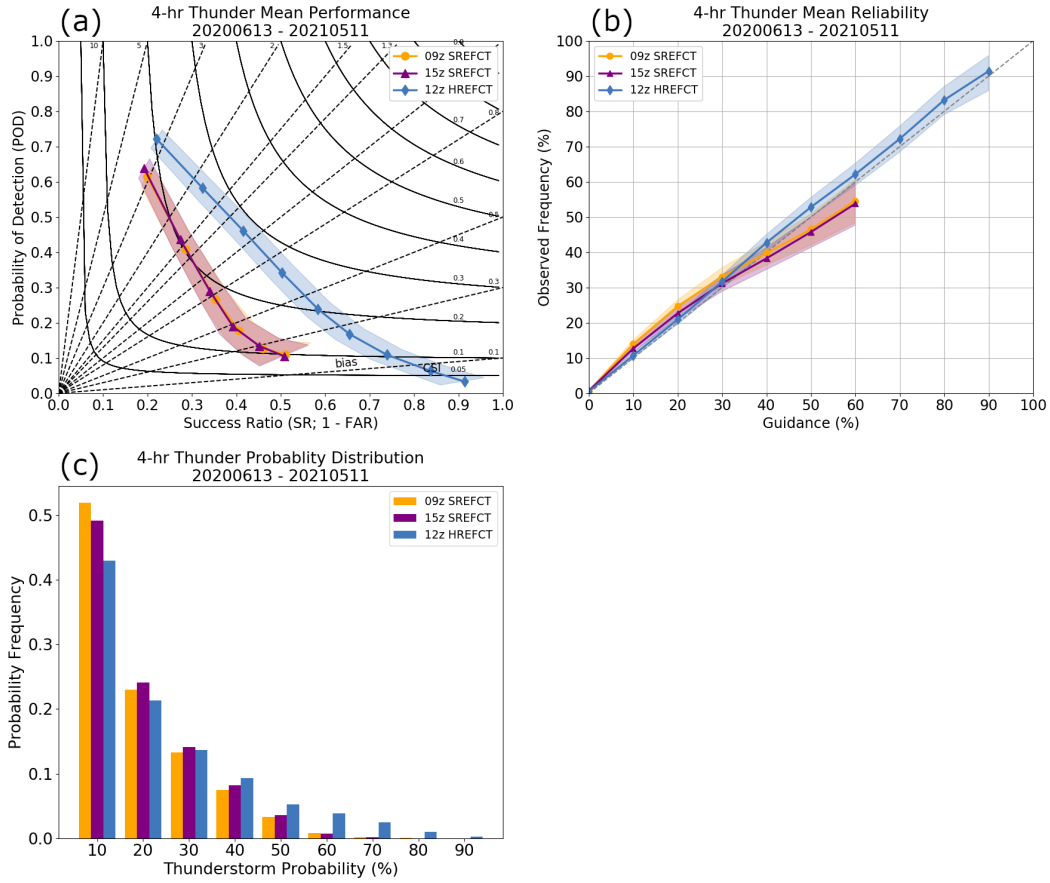


Figure 5.5: A comparison of 09z and 15z SREFCT and 12z HREFCT 4-hour (a) mean performance, (b) mean reliability, and (c) forecast probability frequency for 20200613 - 20210511. The shaded regions represent the 95% confidence intervals from 10,000 bootstrapped samples.

23z - 00z on day 2 (12z HREF f36; 09z SREF f38; 15z SREF f32). This time, the HREFCT achieved a mean CSI of about 0.26 (0.25 - 0.27) while the 09z and 15z SREFCT forecasts had mean CSIs of about 0.21 (0.20 - 0.22) and 0.22 (0.21 - 0.22) respectively. Notably, the day 2 performance peak exhibited by the HREFCT is equal to or greater than the day 1 peak of the SREFCT. The HREFCT demonstrated a minimum mean performance of about 0.18 (0.17 - 0.19) at 10z on day 2 (f46).

More notable differences are evident when comparing the diurnal and lead time variations of the SREFCT and HREFCT mean reliability error (Fig.5.6b). After an initial spin-up period in the first few forecast hours, the HREFCT maintained a reliability error within 5% of observations through 03z of day 2 (f39). Beyond that time, the guidance began to over-forecast CG lightning probabilities by up to 10 - 15%. This reduction in reliability is likely due in part to predictability error in the spatial placement of convection at longer lead times, as well as reduced spread from the ensemble as only five members contribute to the probabilities after forecast hour 36. In contrast, the SREFCT generally remained within 5% of observations up to about 10z (09z SREF f25; 15z SREF f19) before under-forecasting by up to 15% at 16z (09z SREF f31; 15z SREF f25). This strong diurnal signal in the SREFCT reliability error aligns with forecaster observations and past verification studies as discussed in section 5.1. Of note, the HREFCT forecast reliability error did not exhibit a strong diurnal signal and was found to be statistically reliable on average through at least forecast hour 39.

Finally, the HREFCT and SREFCT forecast products were reviewed and compared for several case studies and in real-time SPC forecast operations. Anecdotally, the HREFCT generally produced spatially larger areas of thunder

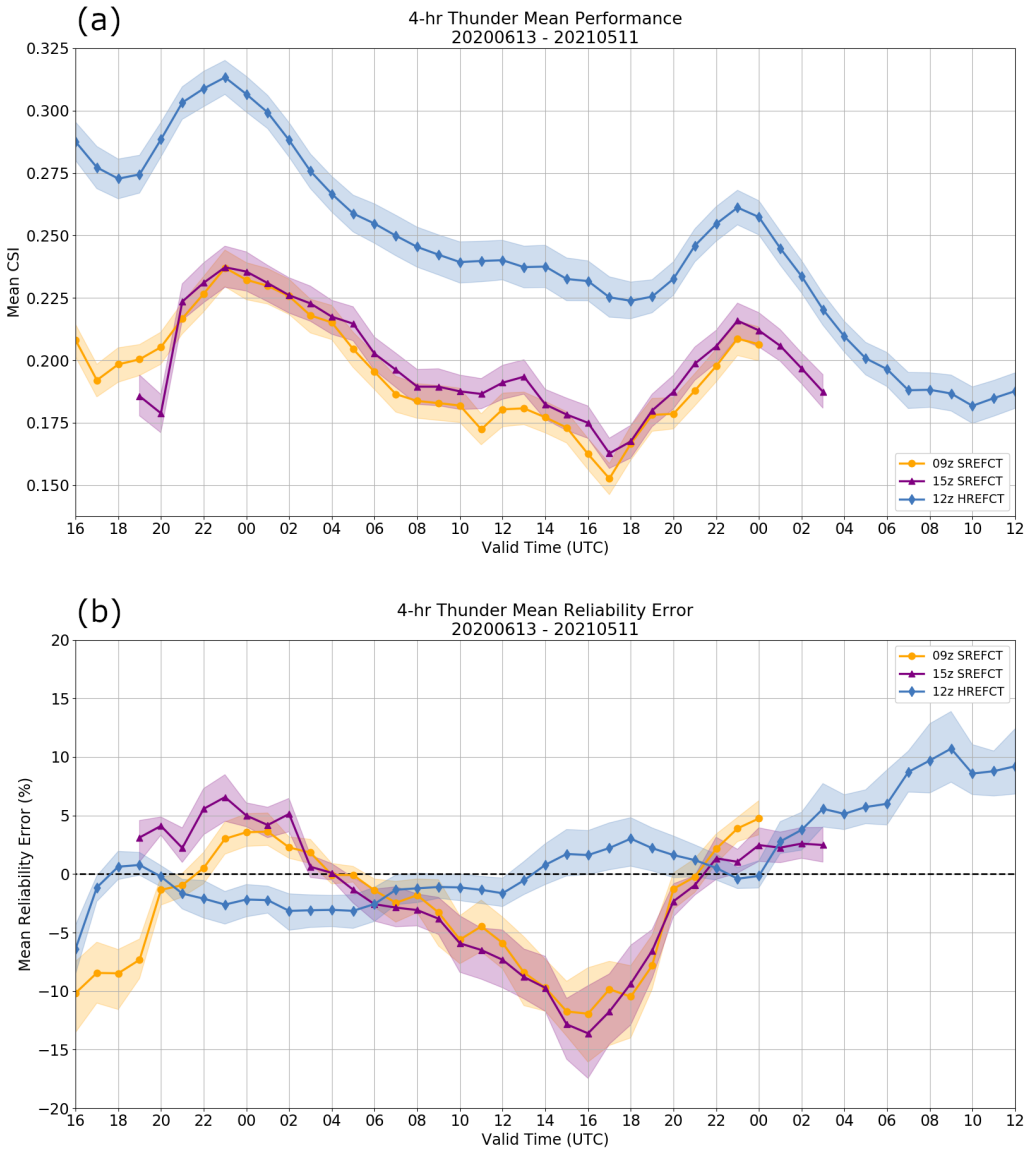


Figure 5.6: A comparison of 09z and 15z SREFCT and 12z HREFCT 4-hour (a) mean performance and (b) mean reliability error as a function of lead time between 20200613 - 20210511. The shaded regions represent the 95% confidence intervals from 1000 bootstrapped samples. Forecast lead time increases to the right for both plots.

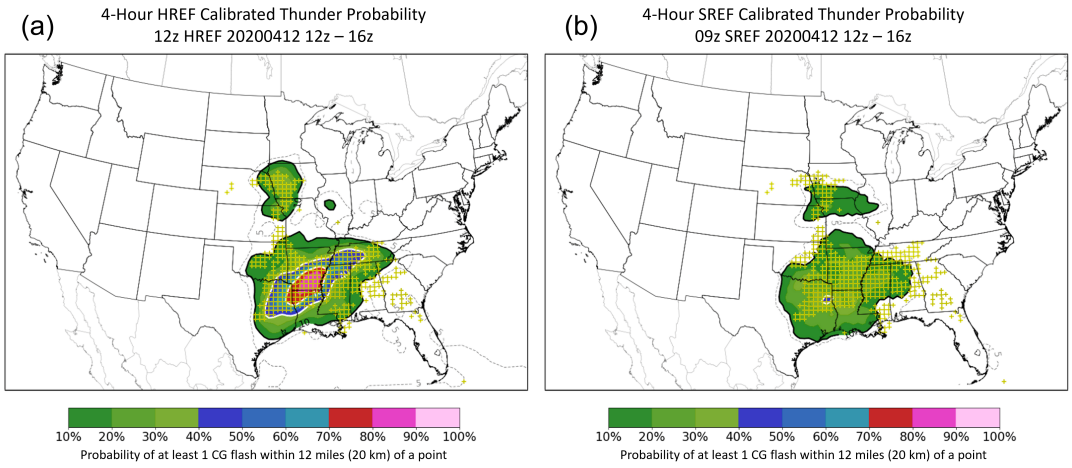


Figure 5.7: (a) 12z HREFCT and (b) 09z SREFCT 4-hour calibrated thunder forecasts for 20200412 12z - 16z. Yellow "+" symbols indicate grid points where there was at least one CG lightning flash detected during the valid forecast period.

probabilities than the SREFCT, and these probabilities were frequently greater in value. One example of this can be seen in Fig. 5.7, which shows a comparison of the HREFCT and SREFCT 4-hour forecast for 12z - 16z 12 April 2020. Both the HREFCT and SREFCT 4-hour guidance correctly predicted the potential for lightning across the southern and central plains and the lower Mississippi Valley. However, the SREFCT probabilities peaked with a small area of 40% in northeast Texas, while the HREFCT painted a broad area of 70 - 80% across parts of Arkansas and Louisiana. The SPC forecaster-created Thunderstorm Outlook for this time period included two regions of 70% across Arkansas, Louisiana, and Texas which more closely aligned with the HREFCT forecast. The HREFCT, SREFCT, and SPC Thunderstorm Outlook all failed to capture the observed lightning over parts of Georgia and Florida, although

the HREFCT did have a few pockets of 5% probability in the vicinity. Scattered convection developed across much of Georgia and north Florida around 15z as a warm front lifted north across the region. Several HREF members accurately depicted the placement and timing of this convection, but the forecast 4-hour Z_{-10C} values were generally < 40 dBZ and 4-hour QPF_{accum} values were < 2 mm. Therefore, the reflectivity and precipitation thresholds of the HREFCT algorithm were not met and the resulting probabilities were $< 5\%$ after calibration. This case highlights a potential limitation of the HREFCT to predict lightning in "dry" thunderstorm scenarios. In particular, the partial reliance of the HREFCT algorithm on QPF_{accum} may greatly limit probabilities when convection is forecast to produce < 1 mm of precipitation in an hour or < 2 mm over a 4-hour period. Additional HREFCT products are currently in development to better account for dry thunderstorm scenarios which play an important role in fire weather forecasting.

5.4 Operational Implementation

Initial prototypes of the HREFCT guidance were iteratively provided to SPC forecasters for evaluation beginning October 2018. To aid in the subjective assessment of the new products, an internal webpage was created that enabled easy comparison of prototype versions, provided objective performance metrics in near-real time, and directly compared prototype forecasts to the operational SREFCT (Fig. 5.8a). Forecasters were also provided access to an online form through which they could send comments about the products' performance during a particular day or event. This virtual form, combined with direct email and

face-to-face conversations, served as the primary conduit for formal communication between the forecasters and myself during the collaborative development process. These evaluation tools were designed in accordance with the principles of collaborative co-production to be easily accessible, reliable, and user-friendly.

Approximately 16 months after the first prototype was introduced to forecasters, a more formal version of the HREFCT guidance was experimentally implemented within SPC operations. This version of the guidance was directly integrated into the SPC's primary operational software, the NCEP-Advanced Weather Interactive Processing System (NAWIPS; Schotz et al. 2008), thus enabling forecasters to directly incorporate the new products as part of their daily forecasting activities (Fig. 5.8b). Daily operational use by an initial few collaborating end users rapidly increased exposure to other forecasters, and formal training sessions were provided as interest increased. Finally, just 2.5 years after the first prototype was created, the SPC formally sponsored the HREFCT guidance for operational implementation. The HREFCT suite of products was officially implemented operationally on NCEP's Weather and Climate Operational Supercomputing System (WCOSS) on 11 May 2021 and is now being distributed by the NWS. Internally, SPC forecasters have largely begun to utilize the HREFCT in combination with or in replacement of the SREFCT when generating thunderstorm forecast products. As of this writing, the 1-hour HREFCT is planned to be added to the NBMv4.1 in combination with the SREFCT, LAMP, and various MOS guidance to inform a blended national thunderstorm probability product.

This demonstrated success of the HREFCT is a testament to the potential of collaborative co-production for rapid R2O transitions. Forecaster feedback was invaluable throughout the development process and both objectively and

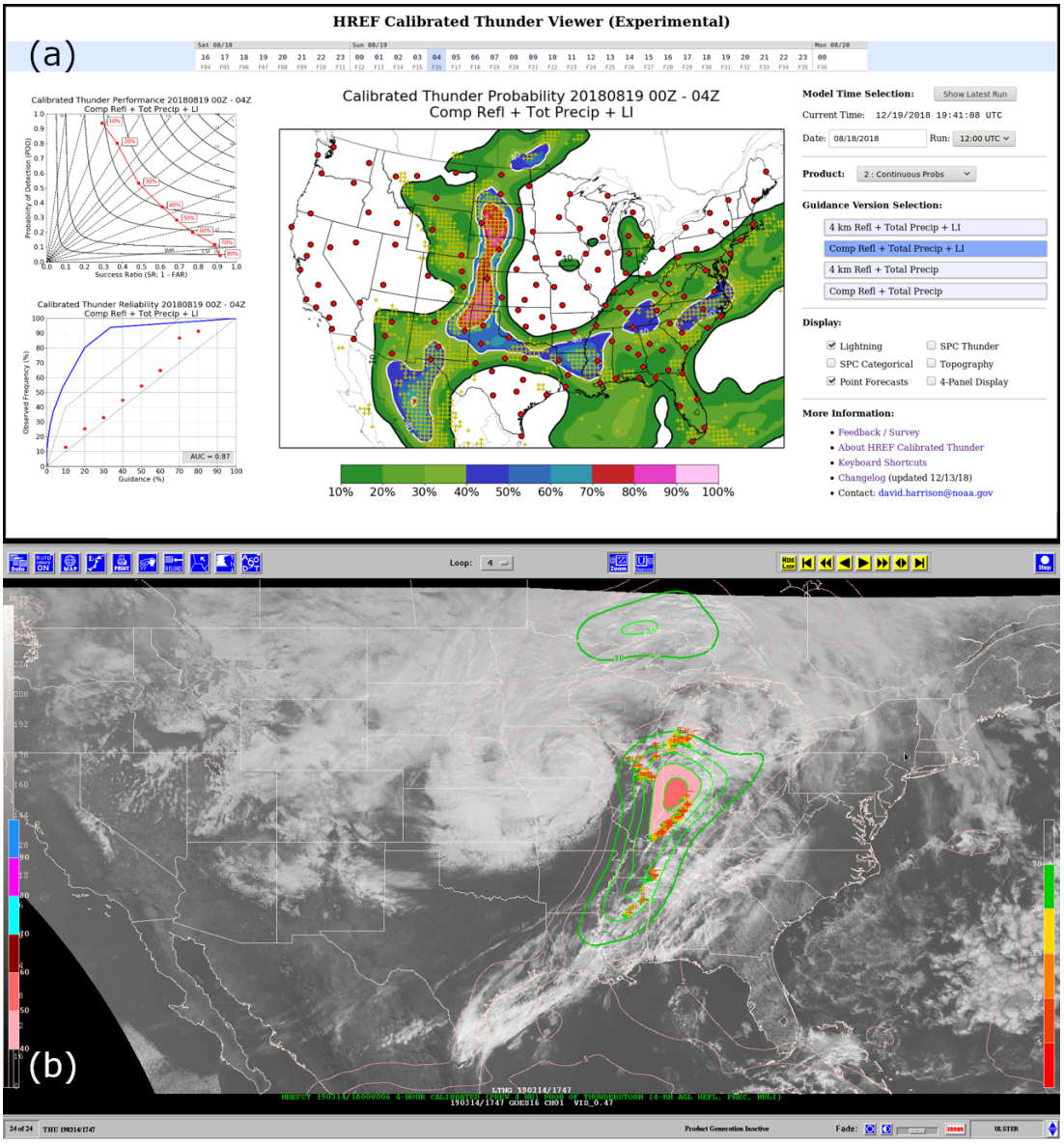


Figure 5.8: (a) An interactive web interface designed to easily compare HREFCT prototypes and operational SREFCT forecasts while also displaying near-real time verification. (b) HREFCT products integrated into the operational NAWIPS software.

subjectively improved the performance of the final product. However, it should be noted that the HREFCT was produced using primarily traditional statistical methods and techniques already familiar to forecasters and researchers at SPC. It is therefore unclear how effective collaborative co-production would have been had the new guidance been developed through less familiar means. To address this question, the remainder of this dissertation will chronicle the application of collaborative co-production to a more complex forecasting challenge that ultimately exposes SPC forecasters and management to the complications and benefits of ML techniques.

Chapter 6

Co-Production of a First-Guess Convective Watch Product

The Storm Prediction Center is responsible for issuing Severe Thunderstorm and Tornado Watch products when conditions become favorable for organized severe thunderstorm development (SPC 2021b). In particular, a Tornado Watch is issued when satellite, radar, and environmental trends appear conducive for multiple tornadoes over a focused geographic area, or when a single intense tornado is forecast. Similarly, Severe Thunderstorm Watches are used when organized convection is expected to result in at least six severe weather events over a confined geographic region, including severe wind gusts ($\geq$ 58 mph), large hail ($\geq$ 1 in. diameter), and brief or weak tornadoes. As described by SPC (2021b), watches are intended to encourage the general public to stay alert to changing weather conditions while providing emergency managers, storm spotters, and broadcast media lead time to prepare for severe weather operations. Additionally, the issuance of watch products has been shown to positively correlate with the quality of NWS warnings (Hales Jr 1990; Krocak and Brooks 2021) and may considerably influence weather awareness among the general public (Gutter et al. 2018).

Forecasters at the SPC issue Severe Thunderstorm Watches with the goal to provide at least 45 minutes of lead time prior to the first severe weather

event (SPC 2021b). Conversely, Tornado Watches are issued with an intended lead time of 2 hours before the first tornado occurrence and at least 1 hour before non-tornado severe weather hazards (i.e., wind or hail). Watch products are initially produced by SPC forecasters as parallelograms that define the approximate area of the predicted severe weather threat (Fig. 6.1b). Those parallelograms are then converted into preliminary county-based watch products prior to direct collaboration with affected NWS WFOs (NWS 2021a) during which WFO forecasters advise SPC of any local considerations that might impact the spatial scope of the watch. As such, some counties included within a Severe Thunderstorm or Tornado Watch may not necessarily fall within the initial parallelogram, and some counties within the parallelogram may not be included in the final watch. Severe Thunderstorm and Tornado Watches typically range in size from 20,000 to 40,000 square miles and have a duration of 6 to 8 hours (SPC 2021b). However, a watch may be canceled early or extended in space and time by local WFOs as conditions require.

SPC convective watches are typically preceded by a mesoscale convective discussion (MCD; SPC 2021b) - a combined graphic and text product that conveys a forecaster's thoughts about how convection will evolve over a mesoscale domain during the next 1 to 6 hours (Fig. 6.1a). Severe weather MCDs are often used to highlight areas of meteorological interest and typically indicate the likelihood that a watch will be issued during the next few hours. Per SPC (2021b), these products are intended to provide extra lead time ahead of potential severe weather development and serve as advance notice to NWS partners that a watch may be issued in the near future. It is the goal of SPC to publish an MCD at least 1 to 2 hours prior to a watch issuance when workload and predictability allow.

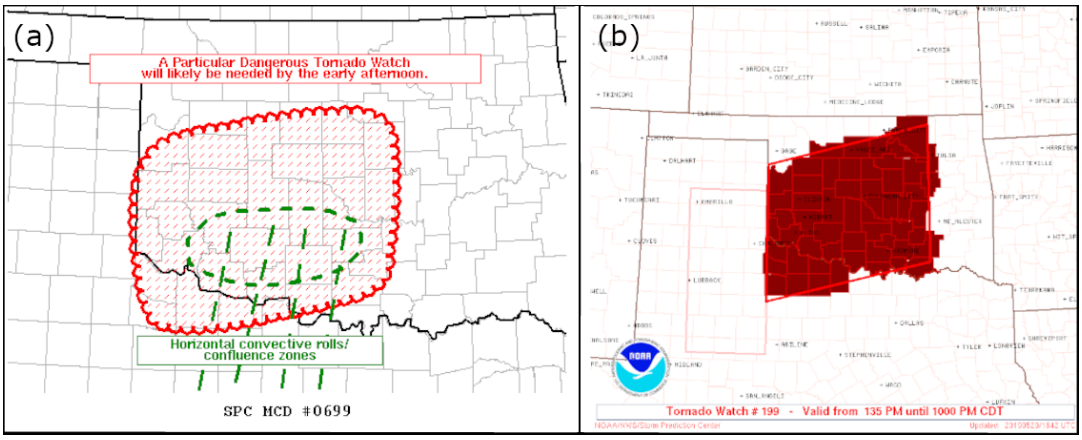


Figure 6.1: Example of an (a) MCD and (b) Tornado Watch as issued by SPC on 20 May 2019. The MCD was issued at 1617z and the Tornado Watch went into effect at 1835z.

Given the stated lead time goals of MCDs and convective watches, SPC forecasters must begin to plan when and where a watch will be issued several hours before the impacts of severe weather hazards are observed. This process is further complicated by NWS procedures which transfer ownership of a watch product from SPC to the affected WFOs immediately after issuance (NWS 2021a). SPC maintains the responsibility to issue 30-minute watch status messages in which SPC forecasters highlight counties within any ongoing watches that are no longer expected to experience severe weather and are recommended to be removed from the watch product. However, any changes to a watch after it is initially transmitted by SPC must be coordinated with and enacted by local WFO forecasters who may be preoccupied issuing warnings, communicating with partners, and performing other high-priority severe-weather operations. As such, it is important that SPC forecasters correctly estimate the location and time of severe weather development to ensure each watch optimally covers the severe weather threat. A watch that does not adequately define the spatial

or temporal domain of the convective weather hazards might require a local extension by WFOs or the issuance of another watch by SPC. These actions ultimately increase workload on both SPC and WFO forecasters and may result in delayed product dissemination, inconsistent messaging, and public confusion.

As with most products in the modern NWS watches, warings, and advisories (WWA) paradigm, Severe Thunderstorm and Tornado Watches are static products that are difficult to modify once issued. Although local WFOs have the authority to cancel or add individual counties to an existing watch, large changes typically require the issuance of a new watch by SPC. These limitations of the current paradigm potentially result in nonuniform lead times across the spatial domains of WWA products (Stumpf and Gerard 2021), where locations on the upstream side of a static product often see impacts from hazardous weather sooner than those farther downstream. For example, consider Tornado Watch 123[1] issued for parts of Louisiana and Mississippi on 13 April 2022 in anticipation of long-track supercells capable of producing tornadoes and damaging wind gusts (Fig. 6.2). Approximately 2 hours after the watch was issued, the first severe storms formed near the western edge of the watch and began producing severe wind gusts and tornadoes eastward across the watch domain (Fig. 6.2b). Local WFOs cleared counties from the upstream side of the watch after the storms passed those locations, but the downstream portion of the watch remained unchanged. Finally, the still-severe storms approached the eastern edge of the valid watch domain six hours after the watch was issued, and SPC decided to issue a new Severe Thunderstorm Watch downstream to capture the continuing severe weather threat (Fig. 6.2d). In this scenario, locations near the western edge of the initial watch domain received about two hours of lead time

[1]https://www.spc.noaa.gov/products/watch/2022/ww0123.html

before observing impacts of hazardous weather, while those on the eastern side saw up to six hours of lead time from the product. This disparity of lead time not only has the potential to be inequitable to populations within the watch (Stumpf and Gerard 2021), but it can also lead to undesirable discrepancies in public response. Krocak et al. (2019) found that members of the public tend to express more uncertainty about how to respond to a severe weather threat as lead time increases, and studies by Doswell (1999), Ewald and Guyer (2002), and Hoekstra et al. (2011) argue that there may be a threshold of lead time at which the public no longer deems the severe weather threat imminent or actionable. As such, the increased lead time provided by downstream portions of a watch may come at a detriment to public response.

To address these challenges of the current WWA system, NOAA's Forecasting a Continuum of Environmental Threats (FACETS; Rothfusz et al. 2018) initiative is tasked with exploring innovative methods that can shift NWS services from static, deterministic products to a dynamic, probabilistic paradigm. Much of FACETs-related research thus far has focused on storm-scale warning generation through dynamically evolving probabilistic hazard information (PHI). Under the PHI concept, WFO forecasters would ideally provide a continuously updating flow of storm-scale hazardous weather information that can be specifically tailored to an end user's location and threat tolerance (Kuhlman et al. 2008; Ling et al. 2015; Karstens et al. 2015, 2018; Harrison 2018; Krocak et al. 2019). However, initial evaluation of PHI in an experimental setting revealed that key NWS partners often rely on deterministic products to trigger critical decisions (Cross et al. 2019) and were generally unfamiliar with how those deterministic thresholds might transfer to a probabilistic paradigm (LaDue et al. 2017; Shivers et al. 2017; Klockow-McClain et al. 2020). As such, more recent

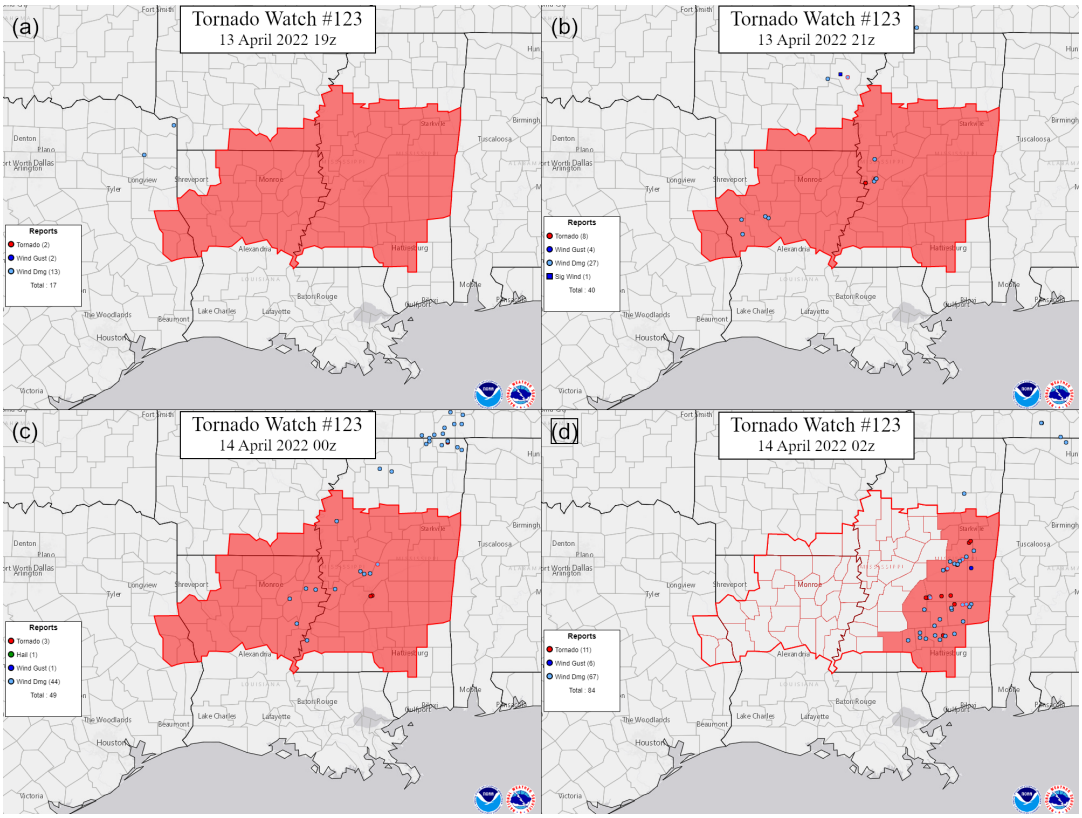


Figure 6.2: Verification of Tornado Watch 123 on (a) 13 April 2022 19z, (b) 13 April 2022 22z, (c) 14 April 2022 00z, and (d) 14 April 2022 02z. Counties with red outlines but no fill represent counties that were cleared from the original watch. Adjacent watches are not shown for clarity.

iterations of PHI have strived to combine the idealized continuous flow of information with the existing WWA structure by deriving static products from the underlying probabilistic information. One such product has been realized in the form of dynamically-updating deterministic warning polygons which automatically track along a severe storm's path in real time. These non-static tornado and severe thunderstorm warnings have been termed "Threats-in-Motion" (TIM; Stumpf et al. 2011; Stumpf and Gerard 2021) and are currently being proposed as a bridge between the deterministic WWA system and the probabilistic PHI paradigm. Although TIM is primarily concerned with storm-scale warnings, I hypothesized that the concept of deriving dynamically-evolving deterministic products from an underlying probabilistic paradigm could be further developed and applied to address the operational challenges and limitations of SPC Severe Thunderstorm and Tornado Watches. To this end, I collaborated with SPC forecasters and management to design and test a prototype system which applies ML techniques to produce a dynamic, first-guess watch product.

6.1 HREF-based ML guidance

In accordance with the collaborative co-production principles described in Chapter 4, the first step towards developing a new first-guess watch product was to become familiar with the current watch issuance process. Ideally, this would have been achieved through in-person observations of the SPC lead forecast desk during severe weather operations; however, restrictions related to the COVID-19 pandemic precluded any in-person shadowing during this collaboration. Instead, I relied upon electronic communication with SPC forecasters, online training

documentation, and personal experience from pre-pandemic shadowing opportunities to compile a working knowledge of the intricacies of convective watch operations. Initial design-phase collaborations identified key objectives for the first-guess watch guidance, including the operational need for both probabilistic and deterministic outputs in a format comparable to existing watch products (i.e., county-based forecasts). Based on the research conclusions of Gutter et al. (2018) and Krocak et al. (2019), informal feedback from WFO forecasters, and formal operational requirements specified by NWS directives, it was determined that the first-guess watch products should optimally provide 2-3 hours of lead time prior to the issuance of storm-based warnings or local storm reports (LSRs). Finally, SPC forecasters and management suggested that these goals should be achieved using both ML and non-ML techniques so that the two approaches might be compared during later evaluations. Development and verification of both versions of the guidance will be featured in this chapter.

Development of the ML-based watch guidance underwent many iterations of the collaborative co-production design and production phases before a viable prototype was considered ready for deeper evaluation. For example, initial research assessed the skill of an ML model trained solely on the individual hazard probabilities contained within SPC's convective outlooks - human-issued forecasts that indicate the likelihood of tornadoes, severe wind, or severe hail within 25 miles of a point location during a convective day (SPC 2021b). This ML model attempted to not only predict where a watch should be issued, but also whether that watch should be a Tornado Watch or a Severe Thunderstorm Watch. Although this initial guidance demonstrated some skill at identifying the type of watch that should be issued, SPC forecasters determined that the location and areal coverage of the first-guess watch products produced

considerable false alarm and did not provide the spatial specificity desired for operational applications. Later iterations added prognostic synoptic-scale environmental parameters and derived storm-scale attributes from the HREF to the ML training data to better localize the spatial placement of the first-guess watches. However, these models continued to produce first-guess watches that forecasters evaluated to be too nonspecific to be useful in SPC operations, and the watch-type predictions also showed decreased skill from earlier attempts.

These failed designs are not discussed in detail in this dissertation for the sake of brevity; however, it is important not to overlook the contributions of this initial research and the collaborative discussion it generated. For example, diagnosing model performance in these early iterations was found to be particularly challenging, as predicting both the type and location of a watch resulted in increased model complexity and reduced interpretability. Therefore, the decision was collaboratively made to only focus on predicting watch placement in the next phase of model development. Additionally, feature importance analysis performed on each iteration of the ML model was crucial for narrowing down the list of input data until the best-performing features were subjectively identified. These lessons learned from failure ultimately led to the creation of a viable ML-based first-guess watch product as detailed in the following subsections.

6.1.1 Feature Engineering

The most viable ML-based first-guess watch guidance was primarily trained using prognostic storm-scale attributes derived from the HREFv2.1 and HREFv3 ensembles. For this study, full 48-hour 00z and 12z HREFv2.1 forecasts were

obtained for 10 March 2018 - 10 May 2021[2], and HREFv3 forecasts were collected for 11 May 2021 - 31 May 2022 (the full period available). As described in Chapter 5 and detailed by Roberts et al. (2019), the HREF is an ensemble of opportunity composed of five deterministic CAM configurations and their 12-hour (6-hour for the HRRR) time-lagged cycles. Each CAM configuration is compiled using different combinations of dynamical cores, initial and boundary conditions, microphysics schemes, and PBL schemes which provide greater forecast spread and more effectively samples forecast uncertainty than unified convection-allowing ensembles (Roberts et al. 2020). This forecast spread is perhaps most apparent in the derived storm-scale attributes produced by each HREF member, as deterministic simulated convection often varies considerably in spatial and temporal placement among the different CAM configurations. However, storm-scale attributes have also been shown to be skillful predictors of severe hazards when smoothed and upscaled to produce probabilistic "surrogate severe" forecasts (Sobash et al. 2011, 2016; Roberts et al. 2019, 2020; Gallo et al. 2021). Such surrogate severe fields were hypothesized to be strong candidates for training an ML-based first-guess watch product as they not only serve as proxies for explicit hazard prediction but also represent localized spatial scales similar to that of operational SPC watches (as opposed to synoptic-scale environmental parameters like MUCAPE). Additionally, these post-processed fields inherently contain information about ensemble spread and forecast uncertainty, thus removing the need to explicitly include individual output from each ensemble member in the training data. This ultimately reduces the complexity

[2]The HRRR member of the HREFv2.1 was retroactively added to the HREFv2 archive after development of the HREFCT had completed, hence the discrepancy in the available dates shown here and in the previous chapter.

of the ML model and makes the product robust to future changes in HREF membership.

Probabilistic surrogate forecasts for tornadoes, severe hail, and damaging wind were derived from the HREF 1-hour maximum updraft helicity (UH), 1-hour maximum updraft vertical velocity (UVV), and 1-hour maximum 10-m wind speed using a technique similar to that described by Roberts et al. (2019). First, each storm-scale field was aggregated using a rolling 3-hour window to produce a 3-hour maximum composite. This was performed in part to satisfy the key project goal of providing up to 3 hours of lead time prior to severe weather occurrence. By training a model on 3-hour composite data, the model forecast at a given forecast hour is inherently valid for the 3-hour window associated with that forecast hour. Recall that HREF forecast hours represent the end of the valid forecast period, so a 12z HREF 1-hour forecast at f09 is valid for the 20z - 21z period. Thus a first-guess watch product derived from the 12z HREF at f09 (20z) would be based on data aggregated from f09 - f11 (20z - 23z) and provide up to 3 hours of lead time for any surrogate severe hazards predicted within that 3-hour window. Each aggregated field from a given ensemble member was transformed into a binary representation of the forecast. Grid points within each ensemble member's 1-hour maximum UH forecast were set to a value of 1 where the forecast UH values exceeded a specified threshold (tuned during model training) within a 40 km x 40 km neighborhood centered on that point. All other points were set to 0. The binary UH fields were then averaged across all ensemble members to produce a neighborhood maximum ensemble probability (NMEP; Schwartz and Sobash 2017), or the ensemble probability that the UH threshold will be exceeded anywhere within the 40-km neighborhood. Finally, these probabilities were spatially smoothed via a Gaussian kernel with $\sigma = 20$

Field	Threshold	Mask
NMEP 1-h Max UH	99.85%	
NMEP 1-h Max UVV	20 ms^{-1}	
NMEP 1-h Max 10-m wind speed	35 kt	$Z_{Comp} > 30$ dBZ
Mean 1-h Max 10 m - 500 mb Shear		$Z_{Comp} > 30$ dBZ

Table 6.1: Derived storm-scale and environmental fields, their optimal exceedance thresholds, and spatial masks that make up the training dataset.

km (tuned during model training). An example of the smoothed UH NMEP can be seen in Fig. 6.3b, and a detailed depiction of the data transformation process is provided by Roberts et al. (2019), their Fig. 1. This process was repeated for the 1-hour maximum UVV and 10-m wind speed fields, and the best exceedance thresholds for each attribute were identified during model tuning as described in the next section.

While UVV and 10-m wind speed exceedance thresholds were calculated from fixed physical values, it was necessary to base the UH threshold on climatological percentiles. Research by Potvin et al. (2019) and Gallo et al. (2021) noted that some members of the HREF produce higher UH values on average than others, and particularly those with an FV3 dynamical core. As such, the predictive skill of a UH value of 75 m^2s^{-2} in the HRRR member, for example, may not be equivalent to that of the same value in the WRF ARW member. To facilitate these differences in model climatology, the UH exceedance threshold was tuned based on percentile values specific to each HREF member. It was also necessary to mask the 1-hour maximum 10-m wind field such that only grid points where the 1-hour maximum composite reflectivity (Z_{Comp}) exceeded 30 dBZ were included in the NMEP calculations. This step was required to

exclude any non-convective wind speeds forecast by the HREF, particularly in mountainous regions and in the stratiform region of extratropical cyclones during the cool season. Finally, the ensemble-mean 10 m - 500 mb bulk wind shear was identified as another promising indicator of severe weather potential, particularly in cases when the forecast UH signal is limited. A full list of training variables and their tuned exceedance thresholds are provided in Table 6.1.

Both SPC parallelogram and county–based Tornado and Severe Thunderstorm Watches were collected for 10 March 2018 - 31 May 2022 and mapped to the HREF's native 3-km grid. Each watch was aligned temporally with the most recent valid HREF cycle and forecast hour, such that a watch valid from 20z to 04z would be paired with the 12z HREF forecast hours f09 - f16. Examples of the positive or target class (WATCH) were compiled by sampling every 50th grid point (2% of all points) within each watch parallelogram at each valid hour. The resulting examples then contained the processed forecast values of the four input fields at the associated grid points and forecast hours. Similarly, examples of the negative or null class (NO WATCH) were compiled by sampling every 50th grid point within at least an SPC Day 1 (D1) convective outlook marginal (MRGL) risk category but not within a watch parallelogram for all hours when a watch was in effect. Conditioning the null class examples to locations within a MRGL ensured that non-severe and pre-severe convective environments were well represented within the dataset and avoided placing too much emphasis on trivial non-convective environments. Additionally, watch parallelograms were used to specify the positive class instead of county-based watches as the simpler geometries considerably reduced the otherwise prohibitive time required to process the data. Note that the number of sampled grid points on a given convective day was determined by the spatial size of the masking watch or MRGL

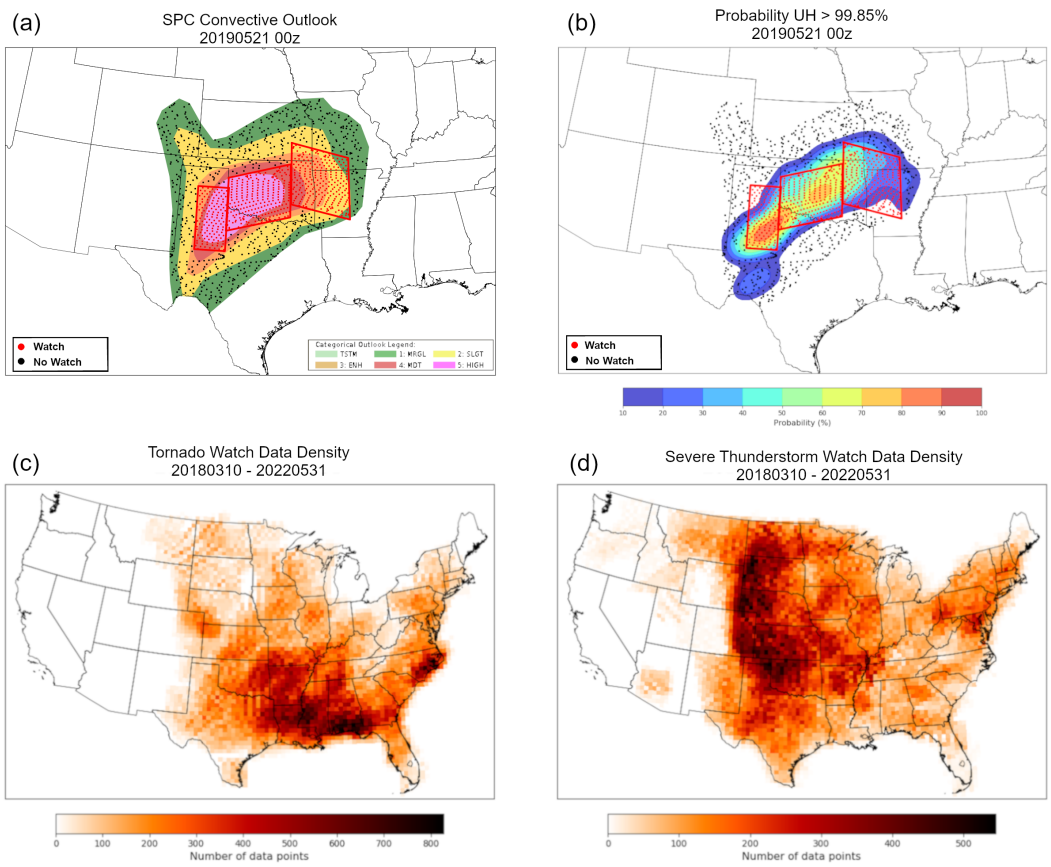


Figure 6.3: (a) Example of grid point sampling within the SPC D1 MRGL risk on 20 May 2019. Red dots indicate positive class (watch) samples while black dots represent negative (no watch) samples. (b) Grid point sampling of NMEP UH values $>$ 99.85% of climatology on 20 May 2019. (c) Spatial distribution of sampled Tornado Watch grid points from 10 March 2018 - 31 May 2022. (d) As in (c) but for Severe Thunderstorm Watches.

risk area such that a smaller mask region produced fewer examples than a larger area. A depiction of the sampling process is shown in Fig. 6.3a,b.

Watches sampled for this research frequently occupied a large percentage of the MRGL risk areas used to mask the dataset, leaving fewer grid points available to sample for the negative examples. As such, the sampling process resulted in a positive skew, with a total of 851,723 positive examples and 495,500 negative examples for the four-year period of study. Most watches were sampled east of the Continental Divide, with the majority of Tornado Watches located in the southern plains, Southeast, and along the Gulf Coast (Fig. 6.3c). Severe Thunderstorm Watches were primarily sampled from the Great Plains, though a secondary maximum of data was obtained along the Northeast corridor (Fig. 6.3d). Finally, null data points represented a largely uniform sampling east of the Rocky Mountains and west of the Appalachians, with slightly reduced density along the East Coast and Florida (not shown).

6.1.2 Model Design

Prior to model development, the dataset was separated into independent training, calibration, and testing sets. Examples from 10 March 2018 - 1 March 2020 were selected for the combined training and validation set, 10 March 2020 - 10 March 2021 was used for model calibration, and the test set contained examples from 20 March 2021 - 31 May 2022. Ten days were withheld between datasets to avoid cross-contamination from temporal autocorrelation within the features. As noted in the previous section, the method used to sample data from the 3-km HREF grid resulted in a greater number of positive class examples than negative examples by a factor of about 1.72. To better balance the two classes

NMEP UH > 99.85% (%)	NMEP UVV > 20 m/s (%)	NMEP 10-m Wind > 35 kt (%)	Bulk Shear (kt)	Class
4.67	10.7	0.00	29.0	NO WATCH
7.22	39.6	13.6	40.7	WATCH
57.9	79.5	79.2	47.9	WATCH
9.68	15.6	0.01	24.9	NO WATCH
15.5	36.1	47.8	22.0	WATCH
0.00	2.36	1.39	51.7	NO WATCH
11.4	44.7	0.00	18.5	NO WATCH
10.8	21.3	25.0	34.5	WATCH

Table 6.2: Example of input values within the training dataset. Note that the "Class" field was removed prior to training and is only provided here for reference.

for training, an oversampling approach was applied so that negative class examples in the training dataset were randomly selected with replacement until the number of negative classes equaled the number of positive classes. This process was applied independently for each convective day within the training dataset to ensure equal representation of the two classes when subsetting by date in later operations. The class ratios remained unchanged in the calibration and testing datasets. An example of the training data is provided in Table 6.2.

This study initially trained and compared the ability of random forest (RFC) and gradient boosted classifiers (GBC) to predict whether each example in the dataset belonged to the WATCH or NO WATCH class. These ML models were chosen for their ability to learn from multiple weak predictors as described in Chapter 2, and for the ease with which probabilistic forecast confidence can be extracted from the model predictions. A randomized grid search (Bergstra and Bengio 2012) with K-fold cross validation (Kohavi et al. 1995) was used to train and tune hyperparameters for each model while also identifying the optimal exceedance thresholds for each NMEP feature. The training dataset was equally partitioned into five folds, with at least one day of examples withheld between each fold to preserve sample independence. One thousand randomized combinations of hyperparameters and NMEP exceedance thresholds were then trained on data from four of the folds and validated on the remaining fold. This process was repeated five times for each model configuration (5-fold cross validation) such that each fold was used for validation exactly once per combination. Finally, the model accuracy (Eq. 2.3) was calculated for each validation fold, and the average accuracy of all five validation folds was reported as the mean validation performance for that selection of model, hyperparameters, and

Number of Estimators:	500	Learning Rate:	0.1
Max Depth:	19	Loss Function:	Deviance
Min Samples/Split:	5	Max Leaf Nodes:	None
Min Samples/Leaf:	5	Max Features:	Auto

Table 6.3: Optimally-tuned hyperparameters used to train a scikit-learn GBC (https://scikit-learn.org).

NMEP exceedance thresholds. Ninety-five percent confidence intervals were also computed for each mean validation score from 10,000 bootstrapped samples.

Models trained using a GBC were found to consistently outperform RFCs throughout the random grid search and cross validation process. Ultimately, the best RFC configuration exhibited a mean validation accuracy of 0.862 (0.859 - 0.865), while the GBC achieved a maximum score of 0.914 (0.910 - 0.917). From these results, the GBC was selected as the best model architecture and the RFC was excluded from future analysis. The model tuning process also found that the surrogate severe storm-scale attributes demonstrated the greatest predictive skill when using exceedance thresholds of 35 kts for 10-m wind speed, 20 m/s for UVV, and 99.85% of model climatology for UH. A list of the tuned model hyperparameters is provided in Table 6.3. Initial analysis of the GBC performance revealed that the model tended to produce overconfident class predictions, with both positive and negative class probabilities heavily skewed towards 0 or 1. This behavior resulted in probabilistic forecasts that were statistically unreliable with the observed class frequency as indicated by a reliability diagram systemically offset from the one-to-one line (not shown). To account for this overconfidence, an isotonic regression model was applied by first running the GBC on the calibration dataset and then training the isotonic regression

on those predictions as described in Section 2.1.4 and (Burke et al. 2020). As before, 5-fold cross validation was used to assess the isotonic regression performance and the 95% confidence interval was calculated via 10,000 bootstrapped samples. The resulting calibrated model did not exhibit any notable change in accuracy scores, and all differences fell well within the 95% confidence interval. However, the class probabilities produced by the GBC and isotonic regression were found to be less skewed and more statistically reliable with observations. All discussion of the GBC model herein refers to the GBC with the isotonic regression applied.

6.1.3 Deriving First-Guess County-Based Watches

During the initial design phase of this research, SPC forecasters and management indicated the operational desire for model output that is directly comparable to the current watch paradigm. As such, it was not sufficient to provide forecasters with a model that only predicts the probability of a watch at a given location and time; rather, the model should provide derived first-guess county-based watch predictions as well. To achieve this goal, it was first necessary to identify the optimal forecast probability threshold to use when stratifying WATCH and NO WATCH predictions. Full-CONUS probabilistic forecasts were generated by applying the calibrated GBC model to each point within the HREF 3-km grid for each forecast hour in the calibration dataset. The resulting 3-km watch probabilities were then interpolated to a 40-km grid and smoothed with a Gaussian kernel (σ = 40 km) to reduce noise and better represent the spatial scales at which SPC watches are typically issued. SPC watches were also mapped to the 40-km grid and temporally aligned with the forecasts as before, and the GBC probabilistic forecasts were compared to the SPC watches

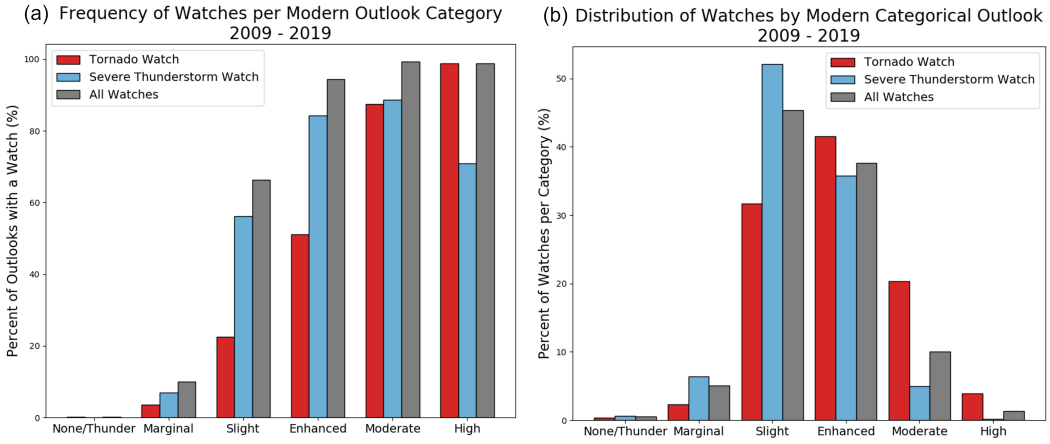


Figure 6.4: (a) Distribution of SPC Convective Outlooks with a watch from 2009 - 2019. (b) Distribution of Tornado and Severe Thunderstorm Watches per modern SPC Convective Outlook category from 2009 - 2019.

on a gridpoint by gridpoint basis. This evaluation employed a similar strategy to that used in Chapter 5, where the forecast probabilities were first stratified into 10% bins centered on every 10% (5-15%, 15 - 25%, etc) prior to calculating mean contingency table metrics for the full 1-year period.

A forecast probability of 70% was identified as the optimal threshold for stratifying WATCH and NO WATCH forecasts, with a mean CSI of 0.24 and a bias of 1.4 (Fig. 6.5a). However, anecdotal observations and bulk statistics revealed considerable false alarm particularly in the lower forecast probabilities. This was corroborated during initial evaluation by SPC forecasters who noted that much of the false alarm was located in environments supportive of precipitation but generally unfavorable for severe convective storms. To help reduce the model tendency to overforecast the spatial extent of watch probabilities, the SPC convective outlook was investigated as a potential way to further

mask non-severe environments. A precursory climatology of SPC watch products revealed that about 65% of all D1 convective outlooks with a maximum Slight (SLGT) risk category received a Severe Thunderstorm or Tornado Watch at some point during the convective day, and this number increased to 93% for Enhanced (ENH) risk days (Fig. 6.4a). Similarly, about 94% of all watches were issued within at least a D1 SLGT risk category, and 99% of watches fell within at least a MRGL (Fig 6.4b). These statistics strongly supported the application of the SPC convective outlook as a mask to remove watch probabilities from environments unsupportive of severe hazards; however, collaborative discussion with SPC forecasters and management resulted in some uncertainty regarding which outlook category to use as the threshold. To objectively address this question, full-CONUS forecasts were once again generated for the 1-year calibration dataset and compared to the SPC watches. This time, however, probabilities were excluded from the verification process if they fell outside of a MRGL or SLGT risk category. The 2000z D1 outlook was used to mask forecasts from the 00z HREF up to forecast hour f12, and the 1300z D1 outlook was applied for 12z HREF cycles up to f12. Similarly, the 1730z D2 outlook was used for 00z HREF cycles from f12 - f36 and the 0600z D2 outlook was selected for 12z cycles through f48. Finally, the 0730z D3 outlook was applied to 00z HREF forecasts f36 - f48. These outlook times were chosen because they represent the most recent SPC forecasts at the time each HREF cycle becomes available for post-processing.

Forecast performance and reliability were calculated for each threshold and compared as shown in Fig. 6.5. Model performance notably increased when only considering points within at least a MRGL, with a maximum mean CSI of

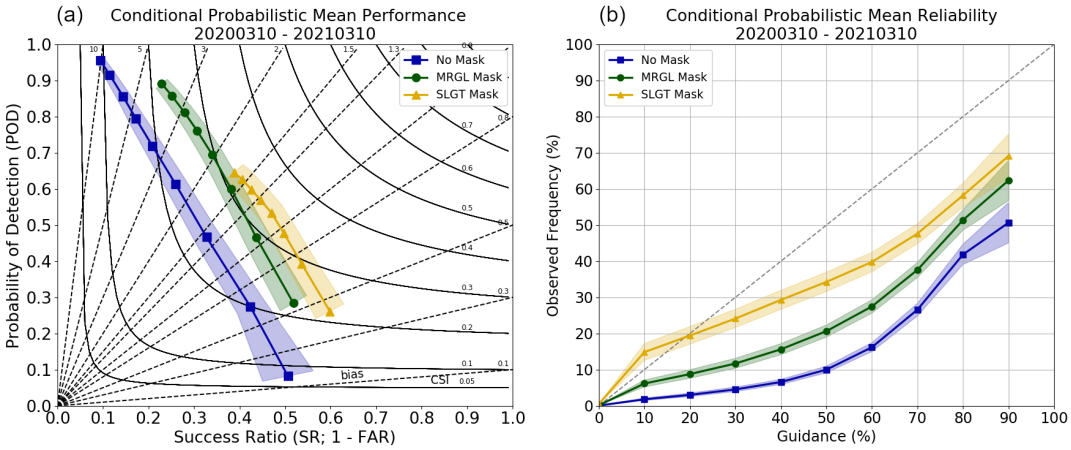


Figure 6.5: GBC (a) mean performance and (b) mean reliability for 20200310 - 20210310 when not masked, masked by an SPC MRGL risk, or masked by an SPC SLGT risk area. Shaded regions represent the 95% confidence intervals from 10,000 bootstrapped samples.

0.29 and a bias of 1.05. Additional improvement was observed with the application of a SLGT risk mask, though diminishing returns were noted compared to the MRGL risk mask. Forecasts within at least a SLGT risk area exhibited a maximum average CSI of 0.32 and a bias of 0.95. GBC forecast probabilities also demonstrated improved statistical reliability as the categorical threshold increased, but all models were still found to overforecast on average at most probability bins. From these results, a set of criteria was proposed to derive deterministic county-based watches from the hourly probabilistic GBC forecasts. A county was included within a first-guess watch product at a given forecast hour if (1) the mean watch probability of all grid points within the county $\geq$ 70% *and* (2) any part of the county falls within at least a SLGT risk area. Counties are also removed from the first-guess watch when these criteria are no longer met. This results in an hourly forecast watch product that ideally

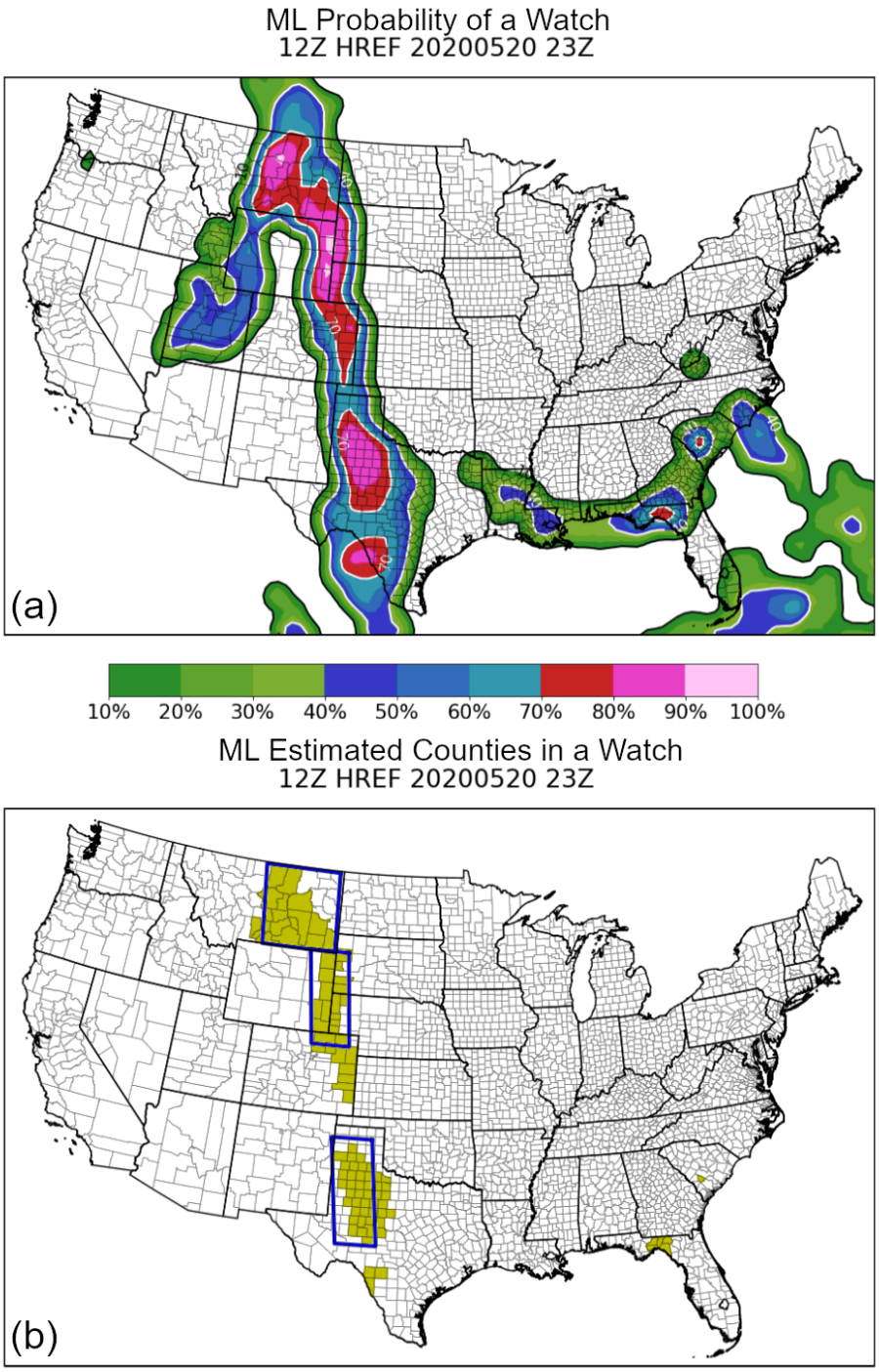


Figure 6.6: (a) Forecast watch probabilities and (b) derived first-guess county-based watches for 20200520 23z. The blue polygons represent operational SPC Severe Thunderstorm Watch parallelograms valid for this hour.

extends about 3-hours downstream of a predicted severe weather hazard and automatically removes counties for locations where the severe weather threat has passed. Recall that the GBC forecast probabilities were previously interpolated to a 40-km grid and smoothed via a Gaussian kernel to remove noise and better represent the scale of SPC watches. To ensure that there was at least one grid point associated with each county in the CONUS, it was necessary to remap these upscaled GBC probabilities back to a 3-km grid prior to calculating which counties should be included in the deterministic first-guess watch product. Note that the remapped probabilities still effectively provide the 40-km scaling but subdivided into the higher-resolution grid. An example of a probabilistic and deterministic watch forecast for 20 May 2020 23z is provided in Fig. 6.6.

6.2 SPC Severe Timing Guidance

A secondary goal of this research was to derive a first-guess county-based watch product using non-ML techniques. SPC forecasters and management suggested this alternative product as a less complex and more transparent baseline to compare against the ML output for the purpose of becoming more familiar with how the ML process differs from more traditional approaches. From the principles of collaborative co-production, it was also hypothesized that the inclusion of this less complex algorithm could give forecasters an increased sense of ownership of the final product and ultimately enhance operational buy-in if the ML-derived guidance proved successful.

This non-ML watch guidance was produced by leveraging the output from an experimental product known as Severe Timing Guidance (Jirak et al. 2020). The SPC Severe Timing Guidance is a prototype system that combines explicit

convective timing and evolution details from the SREF and HREF with human-issued SPC convective outlooks to provide probabilistic information about how and when the severe weather threat will evolve during a convective day. As described by Jirak et al. (2020), the Severe Timing Guidance is produced by first geometrically closing each discrete probability contour of the tornado, wind, and hail D1 convective outlook such that each contour becomes a geometrically-valid polygon. This step primarily fixes instances where the original outlook contours intersect with the outer bounds of the National Digital Forecast Database (NDFD; Glahn and Ruth 2003) domain and thus do not form a closed geometric shape. The closed outlook probabilities are then mapped to the HREF 3-km grid and the discrete probability contours are converted to gridded continuous probabilities via linear interpolation (Loken et al. 2020). This method ensures that the locations of the probability contours issued by SPC forecasters are preserved and that only points within the bounds of the original contours are modified. An example of a continuous-probability SPC outlook is provided by Jirak et al. (2020), their Fig. 1.

The next step of the Severe Timing Guidance algorithm is to derive convective timing information for each severe weather hazard from HREF storm-scale attributes and SREF environmental parameters. This method pairs the aggregated 4-hour HREF NMEP UH $\geq 75\ m^2 s^{-2}$ with the 40-km neighborhood ensemble probability that the SREF significant tornado parameter (STP), MU-CAPE, and effective-layer wind shear fields will exceed hazard-specific thresholds during the same 4-hour period. The environmental fields and exceedance thresholds are different for each hazard (tornado, wind, and hail), and those values are provided in Table 6.4. The HREF/SREF calibrated 4-hour tornado,

Hazard	**HREF**	**SREF**
Tornado	UH $\geq 75\ m^2s^{-2}$	STP ≥ 1
Hail	UH $\geq 75\ m^2s^{-2}$	MUCAPE $\geq 1000\ J/kg$ Eff. Shear $\geq 20\ kts$
Wind	UH $\geq 75\ m^2s^{-2}$	MUCAPE $\geq 250\ J/kg$ Eff. Shear $\geq 20\ kts$

Table 6.4: From Jirak et al. (2020): "Probabilistic inputs from the HREF and SREF to the calibrated hazard guidance."

wind, and hail probabilities are then obtained by calculating the historical frequency of each hazard within 25 miles of a point location given the predicted combination of UH and environmental NMEPs during the valid forecast period. Finally, the gridded HREF/SREF calibrated probabilities are scaled relative to the SPC D1 Convective Outlook such that a scaling factor < 1 indicates that the full-period, grid point-dependent SPC probabilities are less than the 4-hour HREF/SREF guidance. This scaling factor is applied to each 4-hour HREF/SREF calibrated hazard probability to ensure that the Severe Timing Guidance does not exceed probability values issued by SPC forecasters and that lower guidance probabilities are increased to be consistent with SPC messaging. These scaled probabilities are then smoothed via a gaussian kernel to produce the final hourly overlapping 4-hour tornado, wind, and hail probabilities for the convective day (Jirak et al. 2020).

The inputs used to derive the SPC Severe Timing Guidance algorithm exhibit many parallels to those used to train the ML-based first-guess watch product as described previously. For example, both techniques utilize the HREF

NMEP UH as a proxy for severe hazards, and both methods temporally aggregate these storm-scale and environmental attributes to produce rolling windows of lead time. Because of these similarities, the experimental SPC Severe Timing Guidance was selected as a starting point to derive a non-ML first-guess watch product. First, the gridded Severe Timing Guidance individual hazard probabilities for a given forecast hour were mapped to their equivalent SPC Convective Outlook categories. For example, a grid point with a Severe Timing Guidance tornado probability between 5% - 10%, wind probability between 15% - 30%, or hail probability between 15% - 30% was considered equivalent to a SLGT risk. TThe maximum category between the three hazards was then identified for each grid point and assigned a numerical value corresponding to that risk category (1 = MRGL, 2 = SLGT, etc). Finally, first-guess county-based watches were derived using similar criteria as that defined for ML-based watches. Specifically, a county was included within a first-guess watch at a given forecast hour if (1) the maximum Severe Timing Guidance equivalent risk category within the county was at least a SLGT (≥ 2), and (2) any part of that county was included within an SPC convective outlook SLGT risk contour. This second condition was primarily included to be consistent with the ML criteria, as the Severe Timing Guidance by definition should never produce probabilities greater than an equivalent SLGT outside of a SLGT risk contour in the official convective outlook. An example of the Severe Timing Guidance first-guess watch product for 23z 13 April 2022 is provided in Fig. 6.7.

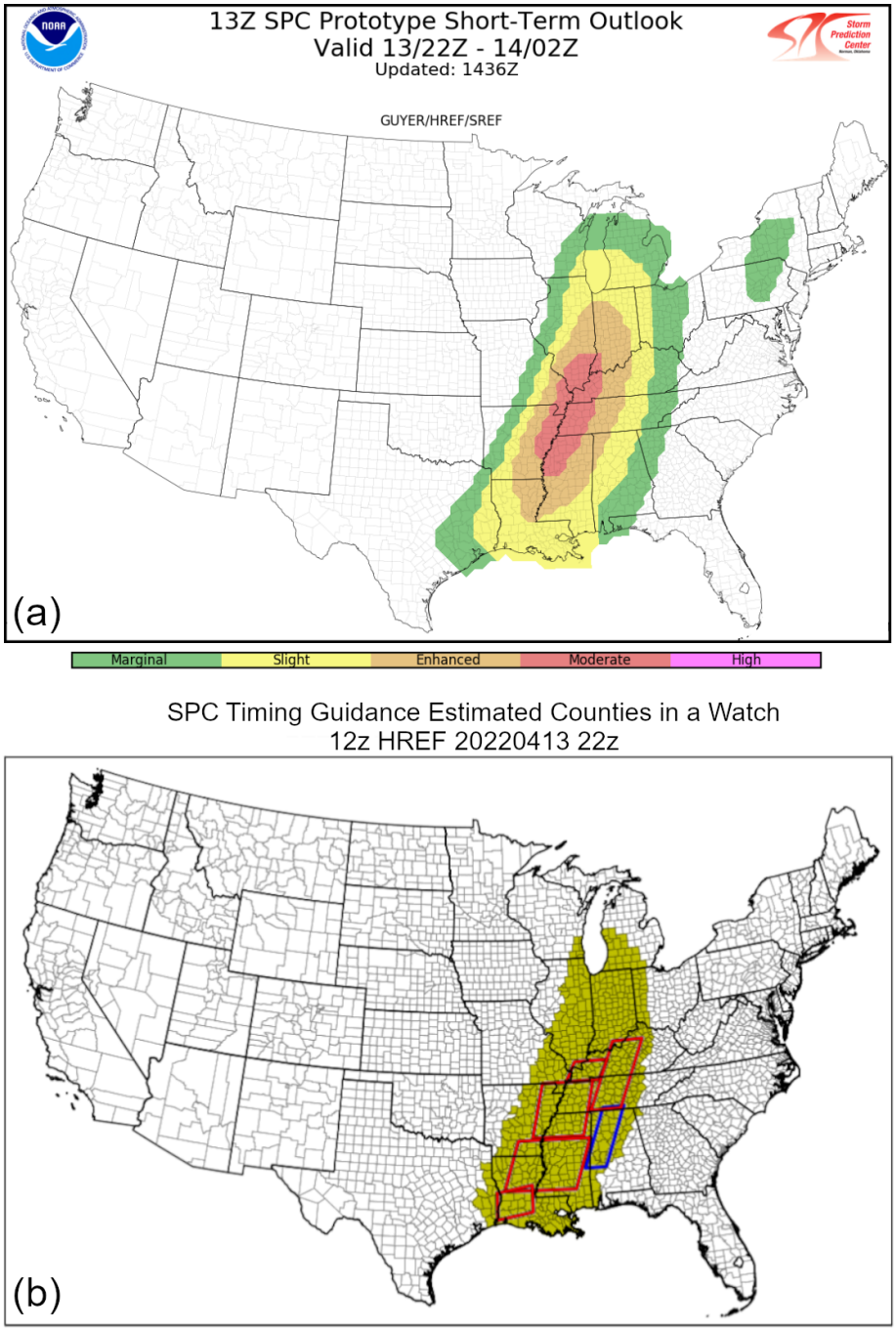


Figure 6.7: (a) SPC Severe Timing Guidance categorical forecast and (b) derived first-guess county-based watches for 20220413 22z. The red and blue polygons represent valid operational SPC Tornado and Severe Thunderstorm Watch parallelograms respectively.

6.3 Results and Discussion

The ML and Severe Timing Guidance first-guess watch products were objectively evaluated on the 14-month independent test set of 20 March 2021 - 31 May 2022. Hourly, full-CONUS, county-based forecasts were generated for each convective day within the evaluation period where the SPC 13z D1 convective outlook contained at least a SLGT risk, and this resulted in a total of 244 days available for verification. During initial collaborations with SPC forecasters and management, two primary goals of the evaluation phase were identified. First, any verification should identify how well the ML and Severe Timing Guidance first-guess watches align spatially and temporally with the official SPC Tornado and Severe Thunderstorm Watches valid during the same period. These metrics are intended to assess whether the forecast products are able to emulate the timing and spatial specificity of a human-issued watch. The second goal of the objective evaluation is to determine how well the forecast guidance is able to correctly predict the true severe weather hazard regardless of when or where SPC issued a watch. This question removes the assumption that SPC watches are always optimal and avoids penalizing the guidance for predicting a watch where severe weather occurred but an operational watch was not issued. By collaborating with SPC forecasters to identify the criteria with which to evaluate these first-guess watch products, this research adheres to the principles of collaborative co-production and ensures the experimental products are assessed in the ways that matter most for operational application.

Prior to this evaluation, it is important to note several caveats in this analysis. First, there are differences in the inherent lead time of the ML and Severe Timing Guidance products. Inputs to the GBC model are aggregated over a

3-hour rolling window to approximate a 3-hour lead time prior to severe weather occurrence; however, the SPC Severe Timing Guidance is derived from rolling 4-hour periods. As such, the hourly Severe Timing Guidance first-guess watch product is produced in such a way that a predicted watch at a given forecast hour should ideally provide up to 4 hours of lead time prior to the observation of severe weather. This discrepancy complicates comparisons of the ML and Severe Timing Guidance products, as the Severe Timing Guidance should in theory produce first-guess watches earlier than the ML approach. Future iterations of this research are planned to better align the temporal component of the two methods; however those results are not available for this dissertation. Additionally, SPC watches are designed to produce about 1 hour of lead time for non-tornado hazards and 2 hours of lead time for tornadoes. These differences in intended lead time could result in situations where the first-guess guidance recommends a watch earlier than the operational watches were issued, and this may be desirable behavior. Another factor to consider is that the dynamic paradigm represented by the hourly ML and Severe Timing Guidance first-guess watch products is not entirely compatible with the static nature of operational Tornado and Severe Thunderstorm Watches. An operational watch, for example, may initially cover a large spatial area even though storms may not move into the downstream portion of the watch until several hours after issuance. Conversely, the forecast guidance products are designed to dynamically evolve with the severe weather threat, such that locations downstream should ideally only be added once they are within 3-4 hours of the predicted hazard. This may result in instances where the forecast guidance technically underforecasts the spatial extent of an SPC watch at a given forecast hour, but

the reduced coverage may be a desirable behavior. These limitations will be discussed in greater detail in the following subsections and in Chapter 7.

6.3.1 Comparison to SPC Watches

The first stated goal of this objective evaluation is to assess how well the forecast guidance emulates the timing and placement of SPC Severe Thunderstorm and Tornado Watches. Deterministic, county-based, first-guess watches produced by the ML and Severe Timing Guidance algorithms were first mapped to a 40-km grid for each forecast hour in the evaluation dataset. SPC watches were also mapped to the same 40-km grid, and the products were compared on a grid point by grid point basis. Contingency table metrics were calculated for the ML and Severe Timing Guidance using the operational SPC watches as "true" observations. As such, a predicted first-guess watch at a given grid point verified as a TP only if there was an SPC-issued Severe Thunderstorm or Tornado Watch valid at that grid point and forecast hour. The resulting performance metrics serve as an indication of how well the guidance predicted *exactly* when and where an operational watch would be valid on a given convective day (within 40 km and 1 hour). Because both the ML and Severe Timing Guidance forecasts are designed to provide greater lead time than operational SPC watches, only forecast hours with at least one valid SPC watch were included in this analysis. This limitation was applied to avoid penalizing the guidance for recommending first-guess watches earlier than those issued by SPC; however it also precludes evaluation of the true false alarm produced by the models. As such, the results presented in this subsection are conditional, subject to the existence of at least one operational Severe Thunderstorm or Tornado Watch at a given hour. A

more detailed discussion of the products' true false alarm is included in the next subsection.

The conditional POD, FAR, CSI, and bias were calculated for each predictive model for each convective day, and these values were averaged across the full 14-month evaluation period (Fig. 6.8). Note that the SPC Severe Timing Guidance is only run through forecast hour f24, whereas the HREF–based ML guidance is available through f48. To ensure a fair comparison between the two products, only forecasts up to f24 were considered for these calculations. The verification presented herein is for ML forecasts generated from the 12z HREF and Severe Timing Guidance based on the 12z HREF and 13z D1 convective outlook. Similar results were noted for the 00z HREF forecast cycle. Both the ML and Severe Timing Guidance first-guess watches exhibited a mean CSI of 0.32, though the Severe Timing Guidance was found to have a slightly larger 95% confidence interval (0.27 - 0.38) compared to that of the ML guidance (0.28 - 0.36). At first glance, the ML guidance appeared to generally exhibit a higher POD and FAR compared to the Severe Timing Guidance, suggesting that the ML approach produced a greater frequency of positive class predictions (both TP and FP) than the non-ML approach on average. However, these differences were not found to be statistically significant at the 95% confidence level. Indeed, both products demonstrated little forecast bias, with mean scores of 1.09 (0.99 - 1.23) for the ML-based approach and 0.89 (0.78 - 1.03) for the Severe Timing Guidance. Furthermore, the optimal forecast bias score of 1 was observed to fall within the 95% confidence intervals for each model, indicating that neither product strongly overforecast or underforecast the areal coverage of SPC watches at times when one was in effect.

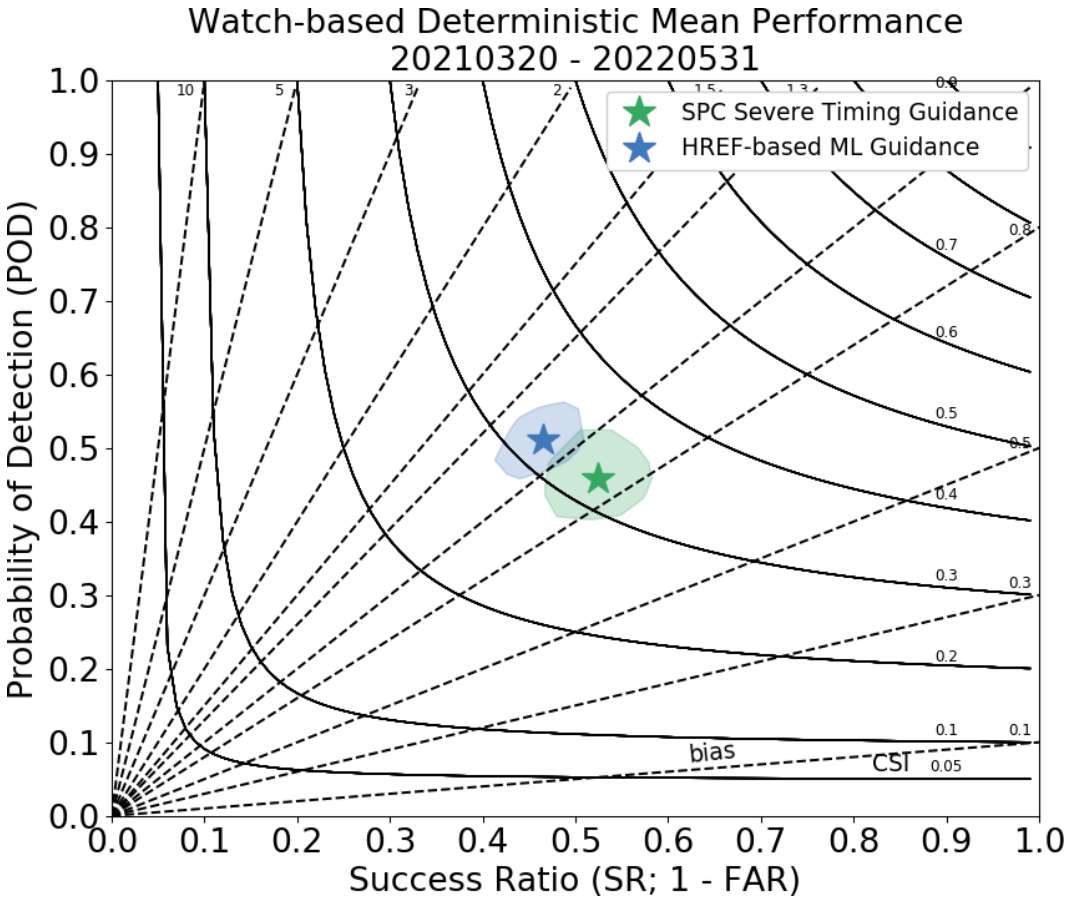


Figure 6.8: Mean performance of the 12z HREF-based ML and 13z SPC Severe Timing Guidance deterministic, first-guess, county-based watch predictions for 20 March 2021 - 31 May 2022. Shaded regions denote 95% confidence intervals from 10,000 bootstrapped samples.

Greater nuance in forecast performance was observed when calculating the mean conditional CSI on an hourly basis as shown in Fig. 6.9. The HREF-based ML guidance exhibited greater performance than the Severe Timing Guidance on average from the beginning of the forecast period through D1 19z (f07). The two products demonstrated similar CSI scores from D1 20z - 03z (f08 - f15), then the Severe Timing Guidance outscored the ML guidance from about D1 04z - 06z (f16 - f23). These results suggest that the HREF-based ML watches may better align with SPC watches during the beginning of a severe weather event, while the Severe Timing Guidance may better capture the end of the event. Intuitively, these discrepancies were first thought to be a result of the increased inherent lead time built into the design of the Severe Timing Guidance product as described previously. However, anecdotal case studies revealed that the Severe Timing Guidance actually tends to produce first-guess watches later in an event than the ML guidance, and this resulted in the product missing the initial timing of SPC watches by up to several hours in some events. These observations will be discussed further in Chapter 7. The ML-based guidance demonstrated a maximum hourly CSI of 0.37 (0.35 - 0.38) at D1 21z (f09), while the Severe Timing Guidance maximum CSI of 0.38 (0.31 - 0.39) occurred one hour later at D1 23z (f10). Both products exhibited a strong diurnal signal in the mean verification scores, with the worst mean performance occurring around 12 - 16z during both the D1 and D2 periods. Note that there was also a strong diurnal signal in the number of forecasts available for evaluation, as SPC watches were less common during the overnight hours. As such, these reduced performance metrics may be influenced in part by a much smaller sample size compared to that of the afternoon and evening hours. Alternatively, the reduced forecast performance may be due in part to dirunal variation in the HREF

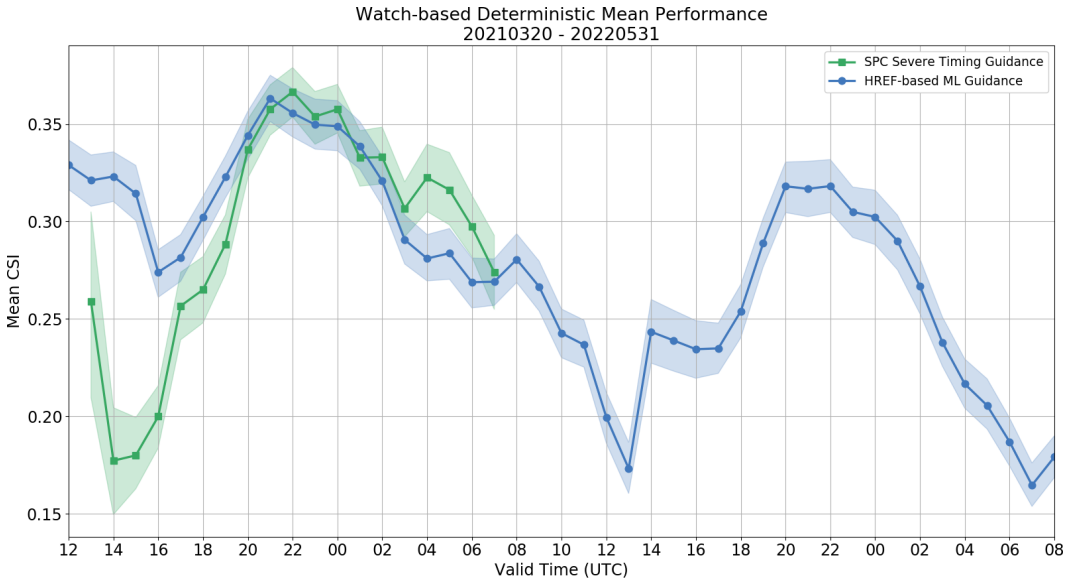


Figure 6.9: A comparison of 12z HREF-based ML and 13z Severe Timing Guidance mean conditional performance as a function of lead time between 20 March 2021 - 31 May 2022. The shaded regions represent the 95% confidence intervals from 10,000 bootstrapped samples. Forecast lead time increases to the right.

ensemble climatology, resulting in reduced nocturnal signal in the storm-scale attributes that may not meet the exceedance thresholds described in sections 6.1.1 and 6.2. This will need to be studied further in future iterations of this research.

A secondary performance maximum was noted for the HREF-based ML first-guess watch product, with a score of 0.32 (0.31 - 0.34) at D2 22z (f39). This is an encouraging result as it indicates that the ML guidance is skillful in the D2 period with performance only slightly degraded from the D1 timeframe. It is therefore conceivable that these first-guess watch products could have value at forecasting the placement of SPC watches at least 36 hours in advance of a severe weather event. Such potential lead time for a watch product may allow SPC forecasters to better plan their watch and MCD strategies, staffing, and shift

activities long before workload increases due to severe weather operations. This was a key motivating factor of this research, and these results are encouraging for future operational impact.

As described in Chapter 2, one limitation of grid-point-dependent contingency table metrics is that it requires the forecast to exactly match observations within a spatial and temporal window specified by the data. In this case, a first-guess watch only verified if a Severe Thunderstorm or Tornado Watch was valid within a 40-km and 1-hour neighborhood of the forecast. However, this does not address the value of a forecast that is close to but not quite aligned with observations. For example, is a first-guess watch invalid because it did not include one county that the SPC watch contained? Should a forecast be penalized because it recommended a watch one hour earlier than the SPC? To address these challenges, a mean FSS (Eq. 2.10) was calculated for both the HREF-based ML and the SPC Severe Timing Guidance first-guess county-based watch forecasts. The FSS was computed at increasing horizontal scales from 40 km to 480 km in 40-km increments, and at time windows of 0 hours, $\pm$ 1 hour, and $\pm$ 2 hours to identify how the guidance verifies with increasing spatial and temporal buffers (Fig. 6.9). The ML guidance exhibited greater FSS on average than the Severe Timing Guidance at all horizontal and temporal scales, with a FSS of 0.43 (0.42 - 0.44) compared to 0.37 (0.36 - 0.39) at native 40 km horizontal spacing and no temporal buffer. FSS generally increased with increasing horizontal scales for both products, but diminishing returns were observed beyond a horizontal scale of about 300 km. Significant increases (at 95% confidence) were noted with a temporal window of $\pm$ 1 hour for both the ML and Severe Timing Guidance forecasts, with FSS increasing to 0.46 (0.45 - 0.47) and 0.42 (0.40 - 0.43) at 40 km spacing, respectively. Additional improvements were noted with a $\pm$ 2

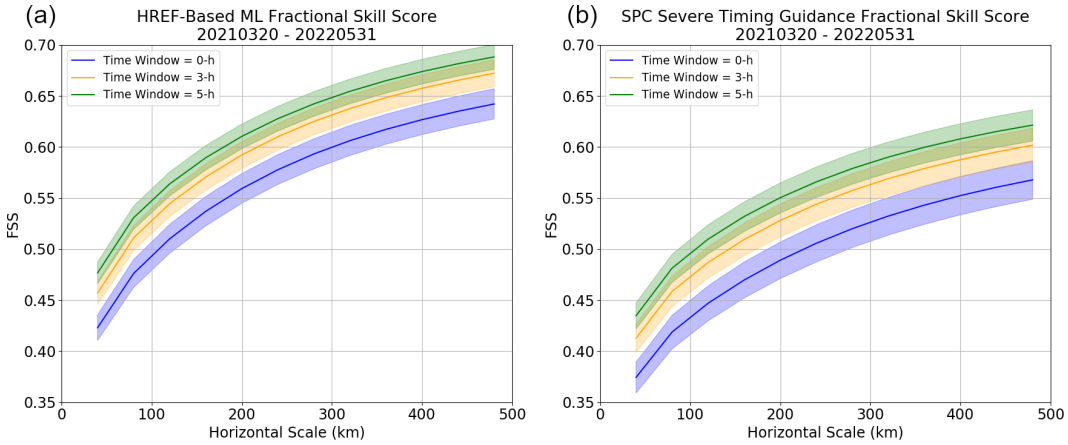


Figure 6.10: (a) 12z HREF-based ML and (b) 13z SPC Severe Timing Guidance FSS as a function of horizontal and temporal scales for 20 March 2021 - 31 May 2022. The shaded regions represent the 95% confidence intervals from 10,000 bootstrapped samples.

hour temporal buffer, though these differences were not found to be statistically significant compared to the ± 1 hour buffer.

The results from this FSS analysis suggest that the first-guess watch forecasts produced by the ML and Severe Timing Guidance may be offset temporally from SPC watches by 1 to 2 hours. This appears to particularly apply to the Severe Timing Guidance product, which saw more notable improvement with a ± 1 hour buffer than the ML-based approach. These results align with the hourly verification metrics and anecdotal observations discussed previously, which also indicated the Severe Timing Guidance first-guess watches tend to be forecast later in an event than those produced by the ML guidance and SPC. Additionally, this analysis suggests that both products may be spatially offset from SPC watches by up to several hundred kilometers in some instances. However, it is unclear if these spatial displacements are true misforecasts or a side effect of the partially-incompatible dynamic and static paradigms. For

example, the ML or Severe Timing Guidance forecasts may predict a first-guess watch for an area that falls downstream of an ongoing severe weather threat, dynamically expanding the leading edge of the watch and clearing the trailing counties as the storms evolve. If the SPC has a static Severe Thunderstorm or Tornado Watch in effect at that hour but haven't yet issued a downstream watch, then the first-guess watch would be deemed spatially displaced from the "true" observations. Such discrepancies in the two paradigms make direct comparisons of the forecast guidance and SPC watches challenging, and indicate the need for more subjective evaluation. This is addressed in Chapter 7.

6.3.2 Capturing the Severe Weather Threat

The second objective of this evaluation was to assess how well the HREF-based ML and SPC Severe Timing Guidance first-guess watch products capture observed severe weather hazards regardless of when or where SPC issued a watch. To accomplish this, LSRs were obtained for each day in the evaluation database and filtered to exclude any reports that fell outside of a 13z D1 SLGT risk area. Filtering reports by the SPC convective outlook avoids penalizing the watch guidance for missing severe weather events in locations where it was systematically precluded from producing a forecast and keeps the verification consistent with methods described in the previous subsection. Mean contingency table metrics were calculated for the ML and Severe Timing Guidance first-guess county-based watch products using a similar method to that described by Anthony and Leftwich Jr (1992). First, POD was calculated as the percentage of LSRs contained within a first-guess watch at the time of the report. Similarly, the percent verified (PV) was defined as the percentage of counties included in the first-guess watch product that contained an LSR during the watch's valid

duration. Finally, Anthony and Leftwich Jr (1992) proposed a modified calculation of FAR to assess the spatial false alarm of watch products. This modified FAR accounts for the spatial distribution of LSRs within a watch product by first mapping those reports to a 40-km grid. Next, an estimated area of impact is assigned to each LSR by defining a 200 x 200 km (5 x 5 40-km grid blocks) neighborhood centered on the report. Anthony and Leftwich Jr (1992) then define the "good area percentage" (A) as the cumulative area of impact contained within a watch divided by the total area of that watch. The modified FAR of the watch guidance for a given convective day is then $1 - A$, where A is spatially summed for all LSRs and predicted watch counties during the convective day. Note that the modified FAR proposed by Anthony and Leftwich Jr (1992) also includes a temporal component which was excluded for this study, as the metrics were calculated on an hourly basis before being aggregated into daily verification scores.

These metrics were computed for both the HREF-based ML and Severe Timing Guidance watch products and averaged across the 14-month evaluation period. Operational SPC Tornado and Severe Thunderstorm Watches were also evaluated using this method, and the mean CSI was calculated from the POD and modified FAR (Fig. 6.11). As before, the ML and Severe Timing Guidance watch products achieved similar mean CSI scores of 0.39 (0.37 - 0.41) and 0.35 (0.33 - 0.38) respectively. However, more notable differences were observed between the products' POD and FAR. The ML watch guidance was found to have a mean POD of 0.65 (0.61 - 0.71), significantly greater than the POD of 0.48 (0.44 - 0.51) exhibited by the Severe Timing Guidance. Conversely, the Severe Timing Guidance demonstrated markedly improved FAR over the ML, with mean scores of about 0.41 (0.36 - 0.48) and 0.52 (0.48 - 0.57) respectively.

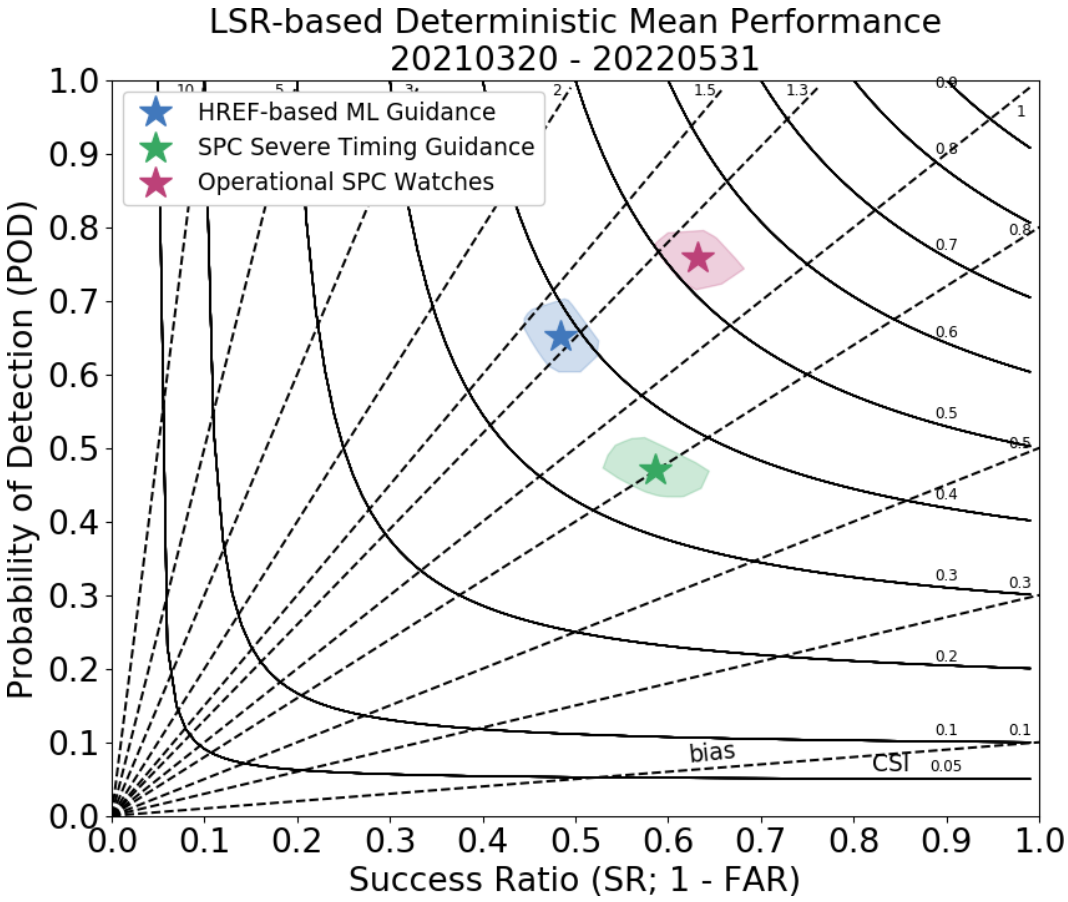


Figure 6.11: Mean performance of the 12z HREF-based ML, 13z Severe Timing Guidance, and operational SPC watch products evaluated against local storm reports for 20 March 2021 - 31 May 2022. The shaded regions represent the 95% confidence intervals from 10,000 bootstrapped samples.

From these scores, the ML guidance was determined to have an overforecast bias of about 1.3 (1.2 - 1.5) while the Severe Timing Guidance underforecast with a bias of 0.8 (0.7 - 0.9). These results are further supported by the mean PV of each model, with an average 37% of ML-predicted watch counties verifying with an LSR on a given convective day. In contrast, 57% of counties predicted by the Severe Timing Guidance verified with an LSR on average.

Both forecast guidance products exhibited notable skill at predicting observed severe weather hazards; however, the human-issued SPC Tornado and Severe Thunderstorm Watches outperformed the two algorithms by a considerable margin. Operational watches had a mean POD of 0.75 (0.71 - 0.79) and a mean FAR of 0.38 (0.32 - 0.42) over the 14-month evaluation period, resulting in an average CSI of about 0.53 (0.50 - 0.55). SPC watches did exhibit a tendency to overforecast with a mean bias of 1.2 (1.1 - 1.3), and about 51% of operational watch counties verified with an LSR on average. This impressive performance serves as a reminder of the skill and expertise of SPC forecasters and again demonstrates the importance of collaboration to ensure that expert knowledge is incorporated into product design. However, it should also be noted that SPC forecasters had access to real-time data including radar and satellite observations which aided in the issuance of the watch products. This gives SPC watches an inherent advantage over the ML and Severe Timing guidance forecast products.

The objective evaluation thus far has revealed the skill of the forecast guidance and operational SPC watches to capture the severe weather threat at time of occurrence; however, watch products are intended to provide some amount of lead time as well. To assess this aspect of the ML and Severe Timing Guidance forecasts, each LSR in the test dataset was mapped to the county it was

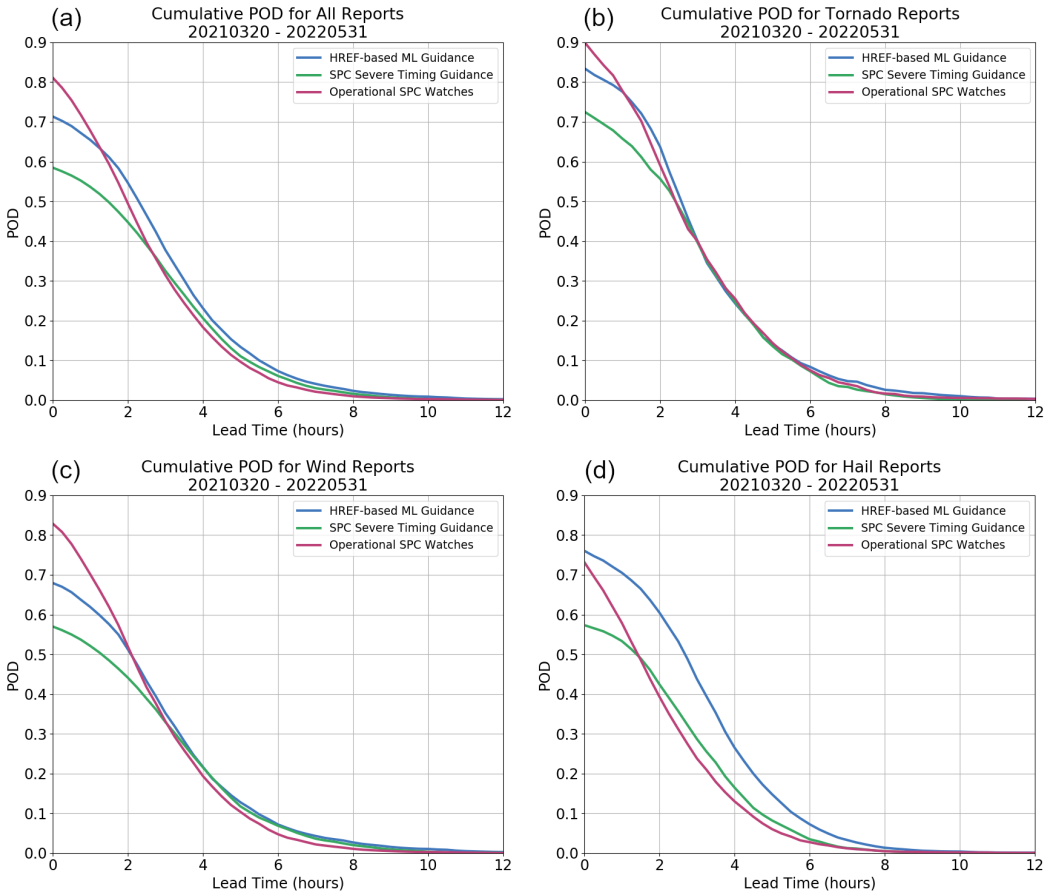


Figure 6.12: 12z HREF-based ML, 13z Severe Timing Guidance, and operational SPC watch product POD as a function of lead time for (a) all reports, (b) tornado reports, (c) wind reports, and (d) hail reports for 20 March 2021 - 31 May 2022.

reported in, and a check was performed to determine if the LSR fell within a first-guess watch at the time severe weather was observed. LSRs not contained within a watch at time of observation were assessed no lead time from the watch product and set aside for later operations. Note that some LSR locations may have been included within a watch product prior to issuance, but the location was removed from the watch by the time severe weather was observed. While it may be argued that those locations still received lead time from the watch product, that lead time was not continuous and thus excluded from this evaluation. LSRs contained within a watch at the time of the report were further assessed by comparing the report to incrementally earlier watch forecasts in 1-hour steps. When a watch forecast that no longer contained the location of a report was identified, the lead time for that LSR was calculated by subtracting the time of the earliest forecast hour with a verifying watch from the time of the LSR. This lead time was then binned into 15-minute intervals and cumulatively plotted for each hazard as shown in Fig. 6.12. Here, POD refers to the fraction of all reports contained within a watch at a given lead time.

The ML-derived first-guess watch product was found to produce very similar lead times to operational SPC watches overall, and particularly for tornado and damaging wind reports. About 60% of tornado reports received a lead time of at least 2 hours from operational watches, while the ML guidance captured about 64% of tornado reports at this lead time (Fig. 6.12b). Both products contained about 52% of wind reports with at least 2 hours of notice (Fig. 6.12c). Conversely, the SPC Severe Timing Guidance first-guess watches tended to capture fewer reports overall, and thus exhibited reduced POD at time frames of 2 hours or less. However, these discrepancies vanished at lead times greater than 3 hours, with all three watch products exhibiting similar performance

at longer time frames. More notable differences in the products' performance were observed for hail reports, with the ML-guidance not only capturing more reports at time of observation than the operational SPC watches (78% and 73% respectively; Fig. 6.12d). The ML first-guess watches captured 60% of hail reports at lead times of at least 2 hours and 28% at lead times of at least 4 hours. Conversely, the operational SPC watches only contained about 40% of hail reports at 2 hours and 13% at 4 hours. The SPC Severe Timing guidance captured 43% of hail reports at 2 hours and 18% at 4 hours. It is unclear why the ML guidance exhibited such improved performance for hail reports specifically, and this will be a topic of future study.

Lead time calculations were also computed for NWS-issued Tornado (TOR) and Severe Thunderstorm (SVR) storm-based warnings to determine how well the guidance predicted the timing and placement of local storms assessed to be potentially severe by expert forecasters (Fig. 6.13). The ML guidance was once again found to perform similarly to the operational SPC watches for TOR warnings, with about 65% of TOR warnings captured by the ML product at 2 hours of lead time and 61% contained by an SPC watch. However, SPC watches exhibited greater POD at shorter lead times, such that 89% of TOR warnings were contained within an operational watch at time of issuance compared to 82% for the ML guidance. The SPC Severe Timing guidance was again found to capture fewer warnings overall, with only 70% of TOR warnings captured by the non-ML first-guess watches.

The ML-derived watch product provided notably greater lead time for SVR warnings than the operational SPC watches, capturing 60% and 50% of warnings with at least 2 hours of lead time, respectively. This discrepancy may be in part a result of SPC watch product definitions, which specify Severe Thunderstorm

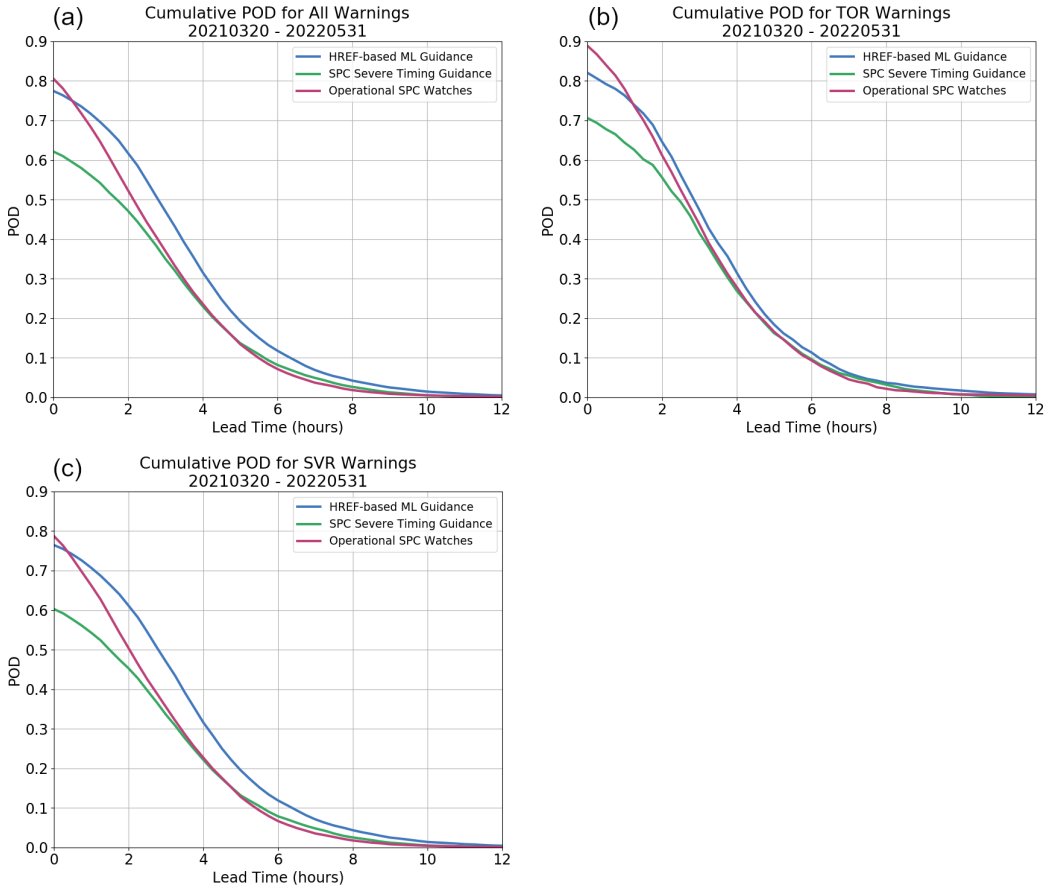


Figure 6.13: 12z HREF-based ML, 13z Severe Timing Guidance, and operational SPC watch product POD as a function of lead time for (a) all warnings, (b) TOR warnings, (c) SVR warnings for 20 March 2021 - 31 May 2022.

Watches should produce at least 45 minutes of lead time prior to the first non-tornado severe weather event. As the ML guidance was designed with the intent to provide 3 hours of lead time for all severe observations, it makes sense that the product would produce greater lead time than the SPC operational watches for non-tornado events. It should be noted that the Severe Timing Guidance first-guess watches provided equivalent or reduced lead time compared to the other watch products, even though it was produced using inputs aggregated over the longest temporal window (4 hours compared to a 3-hour period for the ML guidance). As such, the Severe Timing Guidance should theoretically produce the greatest lead time of the forecast products, but this was not observed. As previously noted, the Severe Timing Guidance first-guess watch algorithm results in fewer watch predictions than the ML guidance, and consequently the guidance has a reduced overall POD. It is possible that the conservative spatial forecasts produced by the non-ML guidance also limit the maximum achievable lead time of the product, as the smaller forecasts may be more sensitive to spatially or temporally displaced HREF/SREF forecasts. Additional work is needed to assess these discrepancies, and future work is planned to compare the ML and non-ML products using the same temporal windows and inherent lead times.

To summarize, both the ML and SPC Severe Timing Guidance first-guess watch products were found to be skillful at emulating human-issued SPC watches and capturing observed severe weather hazards. The ML-based approach tends to overforecast in both instances, with increased POD and FAR over the non-ML algorithm. Conversely, the Severe Timing Guidance demonstrated an underforecast of both SPC watches and observed severe weather hazards, resulting

in decreased POD and FAR. The ML-derived first-guess watch products generally provide equivalent or greater lead time than operational SPC watches on average, with notable improvement observed for hail reports and SVR warnings. These results are very encouraging, particularly for the ML guidance, but also suggest potential room for additional improvement. To further assess how the products perform in real time severe weather scenarios, the ML and Severe Timing Guidance products were presented to a combination of expert operational forecasters and researchers as part of the 2022 HWT SFE for subjective analysis. Their evaluation is discussed in detail in the next chapter.

Chapter 7

Results from the 2022 Hazardous Weather Testbed

As described in Chapter 4, a key step of the collaborative co-production process is to iteratively present product prototypes to intended end users for frequent, subjective evaluation. For example, various experimental versions of the HRE-FCT guidance described in Chapter 5 were evaluated by SPC forecasters for more than a year before the final product was operationally implemented by the NWS. In similar fashion, prototypes of the ML and SPC Severe Timing Guidance first-guess watch products were originally intended to be subjectively evaluated by SPC forecasters over several seasons of severe weather operations. Unfortunately, restrictions related to the COVID-19 pandemic greatly reduced access to SPC operations, and in-person evaluation was not possible during this phase of the research. Additionally, restricted network access and remote work orders prevented deployment of a web-based display or other experimental interface with which forecasters could virtually assess the performance of the guidance on a regular basis. Instead, SPC forecasters provided targeted evaluation feedback through scheduled virtual meetings, during which specific case studies were presented and discussed. While these evaluations provided valuable feedback, the limited sample size and case studies precluded subjective

evaluation of the watch products' long-term performance. To increase the number of product evaluations, the forecast watch products were included as part of the 2022 HWT SFE and subjectively assessed by a combination of operational forecasters, researchers, developers, and students.

7.1 Testbed Design

The 2022 SFE was conducted as part of NOAA's HWT and co-led by NSSL and SPC over a 5-week period from 2 May 2022 - 3 June 2022. Similar to the previous year's SFE described in Chapter 3, this experiment was held virtually via the Google Meet video-communication service, and participants interacted with web-based interfaces to assess and evaluate experimental products. The 2022 SFE hosted a total of 166 participants over the 5-week experiment, including individuals from local NWS WFOs, NOAA research laboratories, universities, cooperative institutes, and international agencies. This was the largest single experiment in the history of the SFE, and participation exceeded the record of 133 SFE attendees set during the previous year. To accommodate such a large group, experimental product evaluations were divided into four groups, and each attendee participated in one group of evaluations per day on Tuesday through Friday. Mondays were reserved for introductions, and no product evaluations occurred on that day. These groups then rotated every other day such that each participant experienced two of the four evaluation groups over the course of the experiment. Previous iterations of the virtual SFE instructed participants to rotate groups each day, allowing every attendee to partake in each evaluation group once during the week. However, feedback from the 2021

SFE indicated that some participants desired more than one opportunity to observe an experimental product to better understand and assess its performance. By rotating participants every other day, the 2022 SFE enabled attendees to observe and evaluate how experimental products performed on two consecutive days of real-weather cases; however, this also meant that each participant was only exposed to half of the experimental products demonstrated in the SFE.

The ML and Severe Timing Guidance first-guess watch products were included as part of the fourth group of evaluations, known as Group D or the "medley" group. Where products evaluated in the other three groups tended to fit a group theme (e.g., derived probabilistic guidance, deterministic CAM configurations, ensemble modeling, etc.), Group D contained a variety of mostly unrelated experimental products including several innovative applications of ML for the prediction and analysis of severe weather hazards. Developers of each experimental product provided a brief introduction to the group each day, and the participants were then given approximately 10 minutes to independently explore the product and fill out an associated evaluation survey. At the end of the evaluation period, the group reconvened and openly discussed their thoughts about the product performance, giving developers the opportunity to ask questions, explain design decisions, and receive direct feedback from their prospective end users. The full 2022 SFE program overview and operational plan is provided in NSSL (2022).

The first-guess county-based watch products were presented to SFE participants via an interactive webpage with three graphic panels as shown in Fig. 7.1. Hourly forecasts from the SPC Severe Timing Guidance were displayed in the left-most panel, the ML guidance forecasts were presented in the middle panel, and the "observed" SPC-issued Severe Thunderstorm and Tornado Watches

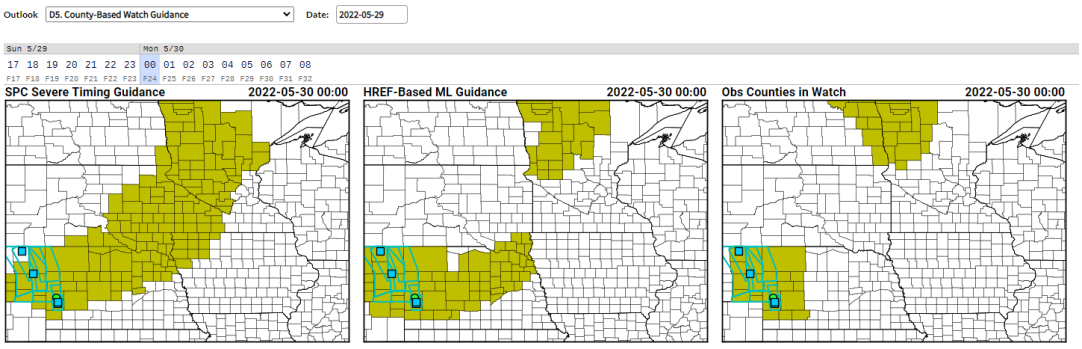


Figure 7.1: Web display presented to 2022 SFE participants while evaluating the performance of the 12z HREF-based ML and 13z SPC Severe Timing Guidance first-guess watch products. Blue polygons represent SVR warnings valid at the displayed hour, blue squares indicate wind LSRs, and green circles represent hail LSRs.

were provided in the right-most panel. An interactive slider bar at the top of the webpage enabled participants to step through each available forecast hour (17z - 08z), and overlays of LSRs, NWS storm-based warnings, and the 13z D1 SPC outlook could be toggled on all three panels. The spatial scope of the evaluation was limited to a rectangular domain of 15° longitude × 8.721° latitude, and this domain was set each day by the SFE facilitators to best contain the severe weather event. All evaluations of the ML and Severe Timing Guidance first-guess watches were performed for the *previous* day's severe weather.

The evaluation survey presented to Group D participants consisted of five questions, including three open response and two matrix-table questions. The first two questions contained metadata, asking respondents to enter the date of the forecast being evaluated and their unique participant number. These questions were included on all evaluation surveys during the 2022 SFE and enabled facilitators to remove results from participants who did not agree to

share their responses for scientific study. Question 3 (Q3) asked respondents to subjectively rate how similar the placement and timing of the ML and Severe Timing Guidance watch products were to the operational Tornado and Severe Thunderstorm Watches issued by the NWS. Each product was assessed independently on a 5-point Likert scale with values ranging from "Not at all similar" to "Extremely similar." Respondents were instructed to consider the full 16-hour forecast period when determining their responses, and an option of "N/A" was provided if there were no operational watches issued for the event. Similarly, Q4 directed participants to subjectively evaluate how well the ML and Severe Timing Guidance watch products captured the location and timing of the severe weather threat during the available 16-hour forecast period. Again, the ML and non-ML products were independently assessed via a 5-point Likert scale ranging from "Terrible" to "Excellent." This evaluation was to be performed using both the LSR and NWS storm-based warning overlays to indicate the observed location and time of severe weather occurrence. Additionally, respondents were instructed to only consider reports and warnings that fell within at least a 13z D1 SLGT to avoid penalizing the forecast products for not capturing severe hazards in locations where the guidance was systematically precluded from issuing forecasts. Finally, Q5 provided an open response field for participants to describe their thoughts about the guidances' performance for the day. This study was approved by the University of Oklahoma Institutional Review Board. A copy of the survey is provided for reference in Appendix B.

7.2 Participant Evaluation

A total of 122 responses were received for the watch guidance evaluation survey over the 5-week SFE, including 43 (35%) from operational NWS forecasters. These survey results were processed and assessed to subjectively identify how well each guidance product emulated SPC watches and captured the true severe weather hazards according to the expertise of the SFE participants. To aid in this evaluation, a KDE was applied to the data in much the same way as described in Chapter 3. As before, the KDE curves represent the relative frequency of responses provided across the 5-point Likert scales and are useful for identifying and interpreting response variance. This time, however, the resulting probability density functions were smoothed using a Gaussian kernel bandwidth of 0.55 to best represent the data. The mean scores and standard deviations were also computed for Q3 and Q4 via 10,000 bootstrapped samples.

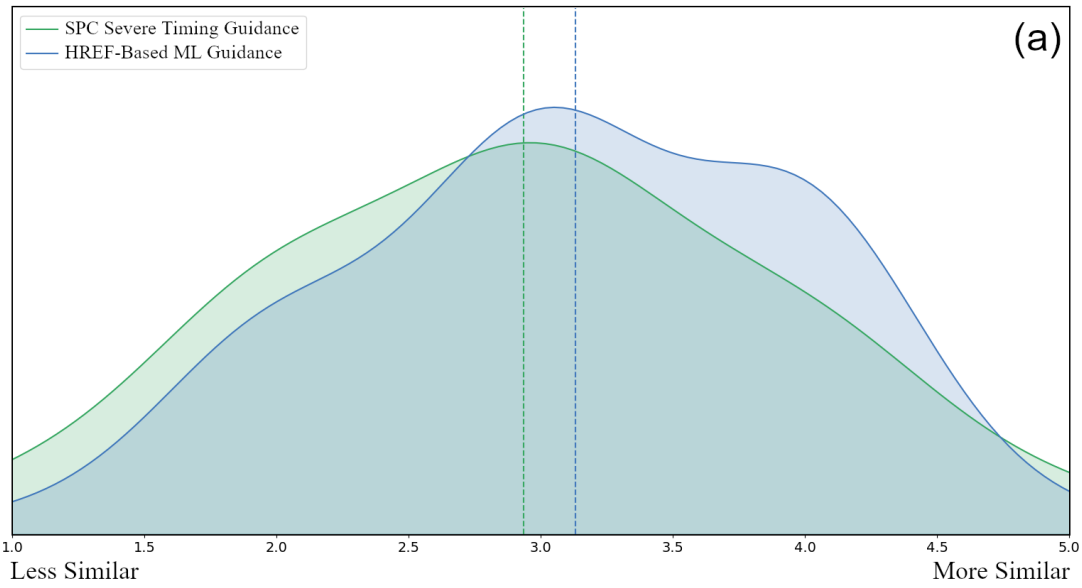
Respondents were neutral on average when rating how similar the ML and Severe Timing Guidance first-guess watch products were to the SPC-issued Tornado and Severe Thunderstorm watches (Fig. 7.2a). The ML guidance received a bootstrapped mean score of 3.13 with a standard deviation of 0.82. Similarly, the non-ML guidance was given a mean rating of 2.93 and had a standard deviation of 0.93. Differences between the two products were small and ultimately not statistically significant at the 95% confidence level; however, the distribution of survey responses does at least indicate a slight trend in favor of the ML-derived first-guess watch products. Approximately 77% of survey responses indicated the ML guidance was at least "moderately" similar to the SPC watches, and 36% of responses found it to be"very" or "extremely" similar.

Conversely, the Severe Timing Guidance was at least "moderately" similar in 67% of responses and "very" or "extremely" similar in only 28% of the results.

Q4 responses generally rated ML and Severe Timing Guidance first-guess watch products more favorably in regard to how well they captured the spatial and temporal domains of the true severe weather hazards, with bootstrapped mean scores of 3.58 and 3.37 respectively (Fig. 7.2b). Additionally, respondent agreement was nearly identical for both products as indicated by a standard deviation of 0.82 for the ML and 0.81 for the non-ML products. As before, these minute differences between the product ratings were not found to be statistically significant at the 95% confidence level, but the response distribution of the ML guidance again trended towards somewhat higher ratings than that of the Severe Timing Guidance. About 86% of responses stated that the ML first-guess watches captured the timing and spatial coverage of the observed NWS warnings and LSRs with at least "average" skill, and 61% said the model performance was "good" or "excellent." In comparison, the Severe Timing Guidance performance was rated as "average" or better in 83% of responses and "good" or "excellent" in 48% of the results.

Anecdotal observations of the guidance products found that the ML model often forecast first-guess watches 1-2 hours earlier than the Severe Timing Guidance algorithm, and the additional lead time was viewed favorably by participants during open discussion. This sentiment was shared in the open response Q5, with one respondent assessing the product performance on 3 May 2022 as, "The ML put up a Watch over southern Ohio a couple of hours before the timing guidance, so it gave better lead time as to where the initial problem area might develop." Another participant commented with regard to 18 May 2022: "I was impressed that both new methods produced watches in advance of

(Q3) How similar was the guidance to operational SPC watches?



(Q4) How well did the guidance capture the severe weather threat?

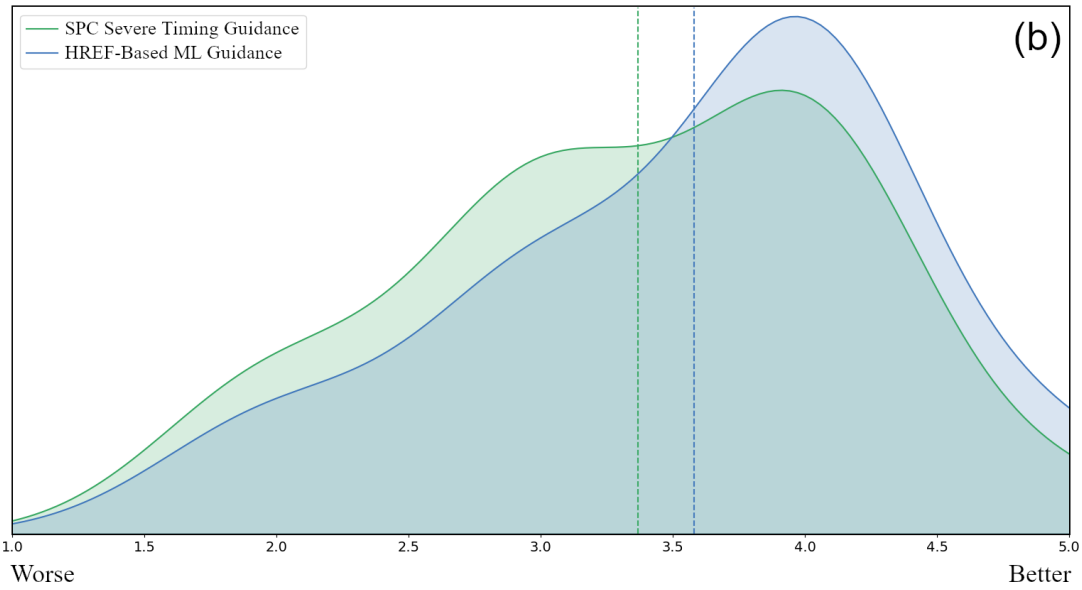


Figure 7.2: (a) Survey Q3 and (b) Q4 responses approximated as KDE curves. Dashed vertical lines represent the mean score for each guidance product.

the official watch issuance." The ML guidance was also observed to frequently produce larger watch areas than the Severe Timing Guidance, and this was evaluated with more mixed reviews. Many participants noted that the smaller Severe Timing Guidance first-guess watches often missed LSRs or storm-based warnings due to the more conservative watch areas, while others argued that the ML-based guidance produced areas that were too large even if they fully captured the threat. Group discussion on this topic often referenced personal preference for better POD or FAR, with opinions varying considerably each day. This was also noted in the survey Q5 responses. One participant indicated a preference for reduced FAR in their evaluation of 24 May 2022, stating, "The HREF-based ML guidance would have given a lot of advanced lead time, but with the caveat that FAR was really high. If I were a user of this product, the Severe Timing Guidance is more accurate. While lead time was not gained as much, it more effectively captured the event while minimizing FAR and canceling watches when appropriate." Another respondent explicitly stated this dichotomy in their evaluation for 31 May 2022, noting that each product's rating, "Goes back to valuing POD (ML guidance better) vs FAR (Severe Timing Guidance better)." This was echoed by several other respondents for the same day, with comments including: "I initially wanted to dismiss the HREF-based ML guidance because of its extensive coverage. And, while it does seem to indicate that the FAR would run high using this guidance alone, it is worth asking about the trade-off between FAR and POD numbers. This then turns into a social science question of how best to serve the public," and "ML HREF guidance was by far the most superior, in some cases better than SPC (especially over TX/OK/KS). Only drawback [in my opinion] was that it was a bit heavy

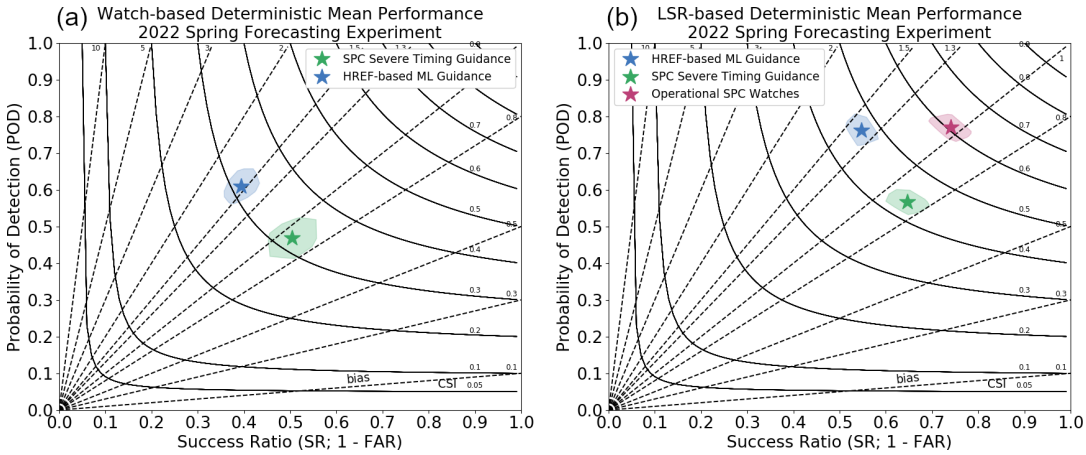


Figure 7.3: Mean performance of the 12z HREF-based ML and 13z Severe Timing Guidance watch products compared to (a) operational SPC watches and (b) LSRs during the 2022 Spring Forecasting Experiment.

handed with watches over MO. But given that this was a frontal boundary with widely spaced convective development, that's a minor consideration."

The comments raised by the SFE participants demonstrate the need for continuous collaboration with end users during development to ensure the final product is tuned to their specific needs. In this case, survey respondents revealed widely varying sensitivities to POD vs FAR based on their own needs and requirements. However, these sensitivities often differ from one end user to another and may change from day to day depending on the circumstances. For example, an operational forecaster may be more concerned with FAR on a day with only low chances for severe weather, but may place greater value on POD during a high-impact severe weather outbreak (Karstens et al. 2018). These variable sensitivities and asymmetric penalties (Doswell 2008) are an ongoing challenge in the study of risk communication and a primary target of probabilistic hazard information research (e.g., LaDue et al. 2017; Shivers et al. 2017; Rothfusz et al. 2018; Klockow-McClain et al. 2020).

The anecdotal observations and subjective analysis presented thus far generally aligned well with an independent, objective verification of the forecast products during the 2022 SFE. Contingency table metrics were calculated for the ML and Severe Timing Guidance first-guess watches with respect to operational SPC-issued watches (Fig. 7.3a) and LSRs (Fig. 7.3b) as described in Chapter 6. These metrics were computed for Monday - Thursday of each week (recall that evaluations were performed for the previous day during the SFE) and from 2 May 2022 - 3 June 2022. The HREF-based ML first-guess watch products exhibited a mean CSI of 0.32 (0.28 - 0.36) when compared to the operational watches, and the SPC Severe Timing Guidance performed similarly with a CSI of 0.32 (0.28 - 0.38). Although both products achieved a nearly identical CSI, the products' biases were more distinct. The ML guidance demonstrated a mean POD of 0.61 (0.58 - 0.68) and a mean FAR of 0.62 (0.58 - 0.64), resulting in an overforecast bias of about 1.5 (1.4 - 1.7). In comparison, the Severe Timing Guidance performed with a mean POD of 0.47 (0.42 - 0.52) and a mean FAR of 0.5 (0.44 - 0.55) for a bias of about 0.9 (0.8 - 1.1).

Similar results were noted when objectively evaluating how well the forecast watches captured the severe weather threat as represented by LSRs. The ML first-guess watches performed with a mean CSI of 0.48 (0.43 - 0.51) and the Severe Timing Guidance exhibited a CSI of 0.43 (0.39 - 0.48). In comparison, the operational watches achieved a mean CSI of 0.61 (0.55 - 0.66) during the same evaluation period. The ML guidance was found to once again overforecast the severe weather potential with a mean POD of 0.77 (0.73 - 0.8), a mean FAR of 0.45 (0.42 - 0.49), and a bias of about 1.4 (1.3 - 1.5). Similarly, the Severe Timing Guidance generally underforecast compared to LSRs, with a mean POD

of 0.57 (0.54 - 0.60), a mean FAR of 0.36 (0.31 - 0.40), and a bias of about 0.9 (0.8 - 1.0).

These objective verification metrics corroborate the subjective evaluation provided by the 2022 SFE participants, demonstrating that the survey respondents were well calibrated on average and unbiased toward a particular forecast product. Additionally, the verification of the ML and Severe Timing Guidance first-guess watch products during the 5-week SFE correspond well to the 14-month verification presented in Chapter 6. Although both products exhibited a higher mean CSI with respect to LSRs during the shorter verification period, that increase was achieved while maintaining approximately constant forecast bias. Similarly, the mean CSI of the ML and non-ML forecast guidance with respect to operational watches remained largely unchanged between the 5-week SFE and the full 14-month verification period, although the ML overforecast bias did increase slightly during the shorter time period. These results suggest that the mean behavior and performance of the ML and Severe Timing Guidance first-guess watches during the severe weather events assessed in the 2022 SFE are largely representative of the products' long-term performance.

Finally, it is worth noting that the 2022 SFE participants were overwhelmingly favorable to the concept of the first-guess watch products and strongly encouraged continued development. This sentiment was captured by one respondent who commented, "From an operational standpoint, this product could be very useful for communicating severe hazards to core partners, especially in the hours leading up to an event." Such a positive response is encouraging and suggests a degree of buy-in from prospective end users.

7.3 Future Work

The results from the 2022 SFE provide valuable subjective feedback from potential end users of the first-guess watch products and offer insight into future improvements. For example, one participant noted that much of the false alarm produced by the ML guidance occurred upstream from the severe weather threat, and particularly later in the severe weather event. This false alarm is hypothesized to be, in part, a result of the Gaussian kernel that is applied to the input fields before generating the forecasts as described in section 6.1.1. This technique, used to spatially smooth the NMEP surrogate severe fields, equally weights all directions radially from a given grid point. As such the spatial extent of the surrogate severe fields is smoothed and expanded in all directions without regard for the relative motion of the predicted storms or any boundaries or sharp environmental gradients in the HREF forecast. In some cases, this may result in counties that fall upstream of a predicted severe weather hazard being included in a first-guess watch as noted during the SFE. To address this undesirable behavior, future iterations of the ML guidance may incorporate a nonuniform smoothing that places greater weight on the Gaussian kernel downstream of the predicted storm motion vector. Alternatively, prognostic environmental fields from the HREF ensemble may be considered to mask counties no longer at risk for severe weather.

Several SFE participants also expressed concern about how the forecast products were masked by the 13z D1 SLGT risk area. In many days of the experiment, the SPC expanded or upgraded the predicted severe weather hazard in later outlooks, and severe events that occurred in those areas were not captured by the masked guidance. As stated in one survey response, “I do think

that it is problematic that the guidance is restricted to the slight risk area. The main corridor of severe weather can occur outside this delineated area and thus this product can be misleading and miss important features." Based on this and similar feedback, the categorical mask applied to both the ML and SPC Severe Timing Guidance first-guess watches will need to be reevaluated, as well as the probabilistic thresholds used to derive deterministic watches from the Severe Timing Guidance product. Future versions of the guidance products are also anticipated to utilize equivalently-sized rolling windows when deriving temporally aggregated storm-scale attributes and environmental parameters used for watch prediction as discussed in Chapter 6. This will enable a more direct comparison between the ML and non-ML products, as the products' inherent lead times will be homogenous.

With the restrictions of COVID-19 gradually rolling back, all future iterations of the first-guess watch guidance are expected to be run in real-time within SPC operations and presented to SPC forecasters via an experimental web interface. Frequent face-to-face collaborative discussions are also planned, and forecasters have already begun offering ideas for improvements and expansions of the current research. Ideally, this increased in-person collaboration will allow future development of the first-guess watch guidance to more closely follow the principles of collaborative co-production, with the goal to reach full operational implementation within the next few years.

Chapter 8

Conclusions

ML, DL, and other AI methods have demonstrated great potential for advancing the state of science within the field of meteorology, but this potential has been somewhat stymied by the slow adoption of AI products by domain experts within operational settings. This hesitancy is often attributed to a perceived inherent distrust that forecasters and domain experts have of systems that are not easily understood or interpretable; however, complexity and opaqueness are not qualities exclusive to ML products within the field of meteorology. Indeed, any nonlinear model or algorithm can be difficult to understand or interpret for those unfamiliar with its design. Therefore, attributing suboptimal R2O performance of new ML products to forecaster distrust of "black boxes" alone implies that operational forecasters must evaluate ML products with different priorities than products derived from non-ML techniques. This notion is directly contradicted by survey results from the 2021 SFE which indicate that operational forecasters on average do not consciously evaluate ML-derived forecast products any differently than they do more traditional or non-ML products. These results were further corroborated by open response comments from the survey participants, in which one respondent explicitly stated, "I would treat a machine learning-produced product pretty similarly to any other probabilistic product." Instead, the survey results revealed that respondents who identified as operational forecasters do evaluate products in general with priorities that differ

somewhat from those who identified researchers and developers. Whereas researchers tended to place more importance on factors that represent a forecast's quality (i.e., verification), operational forecasters emphasized considerations of a forecast's value and consistency (i.e., usability). These differing perspectives are perhaps reflective of the different skills, experience, and responsibilities of the two professions, and can represent a healthy diversity within the meteorological community. However, such varied priorities can also be a source of conflict and confusion during product development if left unmitigated by structured communication and collaboration.

The survey results presented in this dissertation strongly support the dissertation hypothesis that increased communication and structured collaboration between the research and operational communities may improve the success of products in the R2O process. To this end, collaborative co-production is proposed as an approach to product development that enables an equal, reciprocal relationship between developers and their end users. This development cycle extends the principles of Hoffman et al. (2010)'s Practitioner's Cycles, and is based on the ideal that the end user should be considered a valuable resource and ally of the development process. Collaborative co-production ultimately requires a shift of power and responsibility away from researchers and developers to end users through deliberate, user-led collaboration in the initiation, design, production, distribution, and evaluation phases of development. In support of this process, operational forecasters in the 2021 SFE noted that it is important for forecasters to be "involved throughout" the development process, including "early interaction" to provide insight about where additional guidance and support is needed within operations. These comments indicate the desire of operational forecasters to be actively involved with the design and production of

products intended for operational implementation. It should be noted that this sentiment was not unanimous among survey respondents, and some researchers and developers suggested forecasters serve in the role of consultant rather than co-developer.

Collaborative co-production was hypothesized to improve the speed and success rate of R2O transitions by ensuring new products address a real operational need, satisfy operational requirements, and are presented in a way that is accessible by forecasters. Additionally, collaborative co-production principles are expected to expedite management and forecaster buy-in by imparting a sense of ownership of the final product. However, collaborative co-production is an idealized development model with many potential limitations to practical real-world application. To assess the feasibility of this proposed paradigm, collaborative co-production principles were applied to the development of two operations-oriented forecast products using both ML and non-ML techniques. The first project involved frequent collaboration with SPC forecasters to develop a new suite of probabilistic thunderstorm guidance products derived from the HREF. Forecasters were able to view and assess performance of the experimental HREFCT guidance daily in their native operational software, and this enabled rapid, iterative development to tune the products to their operational needs. After just 2.5 years of development, the new products were found to consistently outperform the existing operational guidance by a considerable margin, and the HREFCT guidance is now operational within the NWS.

A similar strategy of collaboration with SPC forecasters was applied to develop ML and non-ML approaches for producing first-guess watch forecasts. These products predict where and when conditions will be favorable for a Severe Thunderstorm or Tornado Watch each hour, and provide a probabilistic,

dynamic framework to help SPC forecasters plan watch issuance strategies, staffing, and shift activities long before severe weather occurs. This dynamic, probabilistic framework is a key innovation for the evolution of operational SPC products toward FACETs goals, and is anticipated to serve as a starting point for the future development of a dynamic "watch-in-motion" product. As before, this watch guidance was intended to be evaluated by SPC forecasters during operational activities over several seasons of severe weather operations. However, restrictions due to the COVID-19 pandemic made in-person communication and evaluation impossible, and greatly restricted access to SPC operational systems for deployment. Instead, the products were tested and evaluated during the 2022 SFE to mostly positive reviews. Feedback from experiment participants has been critical to future development, and this experience highlights one way collaborative co-production can be applied in circumstances where direct access to end users is not possible.

The successful operational implementation of the HREFCT and the promising reviews of the first-guess watch guidance serve as a testament to the potential of collaborative co-production for improving the speed and success of R2O transitions. Additionally, this new production paradigm offers the opportunity to improve our understanding and prediction of atmospheric processes via increased multi-disciplinary knowledge transfer between the operational and research communities. Although the focus of this dissertation is on the co-production of ML in operational meteorology, the results presented here have shown co-production to apply equally to non-ML applications as well. As such, the principles of collaborative co-production are recommended to be integrated into existing development procedures, including the NOAA Readiness Levels described in Chapter 1. Within the current RL system, developers are not

required to directly interact with their intended end users until RL 6, when projects must demonstrate a working prototype within a formal testbed or similar environment. However, this dissertation has shown that it is beneficial to include end users during each stage of the development cycle. For example, RL 2 (Applied Research) may be considered analogous to the first design and production phases of the co-production model shown in Chapter 4. End users (e.g., operational forecasters) could work with researchers and developers at this stage to identify key goals of the research and how the anticipated knowledge gain may improve operational skill or efficiency. Similarly, RL 3 - 7 generally reference various degrees of product evaluation which may be comparable to multiple iterations through the co-production evaluation phase. Frequent, iterative forecaster feedback during such evaluation stages was found to be particularly important to ensure the new product continued to meet end-user needs and requirements. Codifying these researcher-forecaster interactions into the existing RLs would formalize the co-production process within NOAA R2O procedures, increase the scope and frequency of researcher-forecaster collaboration, and potentially improve the success and speed of R2O transitions. Ultimately, this research is presented as a call to the meteorological community as a whole to strive for greater structured collaboration in product development, so that we may better apply our diverse expertise toward our shared scientific goals.

Appendix A

2021 SFE Survey

Consent to Participate in Research at the University of Oklahoma

You are invited to participate in research designed to improve severe weather forecasts.

If you agree to participate, during the time period of the Hazardous Weather Testbed you will make forecasts using experimental products and data. In addition, you may be asked questions while making your forecasts or may be involved in group discussions with other forecasters and researchers. We may also ask you to complete surveys or participate in focus group discussions, which will last no longer than 45 minutes each. These activities may be photo recorded.

There are no risks, benefits or compensation.

Some data will be collected via an online platform not hosted by OU that has its own privacy and security policies for keeping your information confidential. Please note no assurance can be made as to the use of the data you provide for purposes other than this research.

Your participation is voluntary. Even if you choose to participate now, you may stop participating at any time and for any reason. Your photographs or video images may be used in University research reports unless you tell me not to do this. Your data may be used in future research studies, unless you contact me to withdraw your data. After removing all identifiers, we might share your data with other researchers without obtaining additional consent from you.

If you have questions about this research, please contact the researcher of the Hazardous Weather Testbed at this email: [redacted] or [redacted]. Or you can contact [redacted].

You may also contact the University of Oklahoma – Norman Campus Institutional Review Board at 405-325-8110 or irb@ou.edu with questions, concerns or complaints about your rights as a research participant, or if you don't want to talk to the researchers.

Name of Participant:

Date:

← June 2022 →

Su	Mo	Tu	We	Th	Fr	Sa
29	30	31	1	2	3	4
5	6	7	8	9	10	11
12	13	14	15	16	17	18
19	20	21	22	23	24	25
26	27	28	29	30	1	2

I agree to participate in this research.

- ○ Yes
- ○ No

Which of the following best describes your professional background? *Check all that apply.*

- ○ Student
- ○ Academic faculty/staff
- ○ Researcher
- ○ Operational forecaster
- ○ Other (please specify)

For the purposes of this survey, probabilistic forecast products are defined as any probabilistic forecast derived from a numerical weather prediction (NWP) model or ensemble. Probabilistic forecast products represent the inherent uncertainty and spread associated with NWP and can often extract information not explicitly provided by the original NWP model. A few examples of probabilistic forecast products include HREF Calibrated Thunder guidance, HREF/SREF Calibrated Severe guidance, and NOAA/CIMSS ProbSevere.

Please indicate how often you utilize probabilistic forecast products as part of your work-related duties.

	Never	Once a week	2-3 times a week	4-6 times a week	Daily
I use probabilistic forecast products...	○	○	○	○	○

Consider a probabilistic forecast product you have utilized in the past. When evaluating how useful that product might be to your personal forecasting process, how important are each of the following factors?

	Not at all important	Not very important	Somewhat important	Very important	Extremely important
Performance of the product in case studies	○	○	○	○	○
The statistical verification of the product	○	○	○	○	○
Knowledge of the product's limitations and possible failure conditions	○	○	○	○	○
Previous experience with the developers of the product	○	○	○	○	○
Previous experience evaluating experimental versions of the product	○	○	○	○	○
Knowledge of how the probabilistic output is derived	○	○	○	○	○
How closely the probabilistic output aligns with human-generated forecasts	○	○	○	○	○
Timeliness and availability of the product	○	○	○	○	○
How closely the variables used as input to the product align with traditional meteorological knowledge	○	○	○	○	○
Use by other experts in your field	○	○	○	○	○

Please list other factors, if any, that you consider when evaluating how useful a new probabilistic forecast product might be to your personal forecasting process.

Please indicate your knowledge of **machine learning** including random forests, deep learning, and other artificial intelligence techniques.

	Not knowledgeable at all	Not very knowledgeable	Somewhat knowledgeable	Very knowledgeable	Extremely knowledgeable
I am ____ about machine learning	○	○	○	○	○

You are told that a new probabilistic forecast product is the result of a **machine learning** model. How important are each of the following factors when determining how useful the new product might be to your personal forecasting process?

	Not at all important	Not very important	Somewhat important	Very important	Extremely important
Knowledge of the product's limitations and possible failure conditions	○	○	○	○	○
How closely the variables used as input to the product align with traditional meteorological knowledge	○	○	○	○	○
Performance of the product in case studies	○	○	○	○	○
Timeliness and availability of the product	○	○	○	○	○
Previous experience with the developers of the product	○	○	○	○	○
Use by other experts in your field	○	○	○	○	○
The statistical verification of the product	○	○	○	○	○
Previous experience evaluating experimental versions of the product	○	○	○	○	○
Knowledge of how the probabilistic output is derived	○	○	○	○	○
How closely the probabilistic output aligns with human-generated forecasts	○	○	○	○	○

Please list other factors, if any, that you might consider when evaluating how useful a new **machine learning** product might be to your personal forecasting process.

When developing a new probabilistic forecast product intended for operational use, please indicate how important you believe it is for developers to collaborate with operational forecasters during each of the following steps.

	Not at all important	Not very important	Somewhat important	Very important	Extremely important
Exploratory research	○	○	○	○	○
Initial design and planning	○	○	○	○	○
Technical and logistical development	○	○	○	○	○
Product testing	○	○	○	○	○
Publication, training, and outreach	○	○	○	○	○

If you have any additional comments related to this survey, please share them here.

This is the end of the survey. Click the "→" button in the bottom right corner to submit your responses. You will not be able to go back or change your answers once you submit. Thank you!

Powered by Qualtrics

Appendix B

2022 SFE Survey

Please enter the date of the forecasts that you are evaluating (typically, this will be yesterday):

← June 2022 →

Su	Mo	Tu	We	Th	Fr	Sa
29	30	31	1	2	3	4
5	6	7	8	9	10	11
12	13	14	15	16	17	18
19	20	21	22	23	24	25
26	27	28	29	30	1	2
3	4	5	6	7	8	9

Please enter your participant number:

The following questions ask about the performance of automated, dynamic watch products produced by machine learning (ML) and non-machine learning methods. These products are designed to emulate county-based Severe Thunderstorm and Tornado watches while dynamically evolving with the time and location of the severe weather threat. Please navigate to the model comparison webpage here. For this evaluation, please consider the overall performance of each product through the ***entire 16-hour period***.

Subjectively rate how similar the placement and timing of the ML and Timing Guidance watch products were to the operational Tornado and Severe Thunderstorm watches issued by the NWS. If there were no operational watches issued during the period, please select N/A.

	Not at all similar	Slightly similar	Moderately similar	Very similar	Extremely similar	N/A
Severe Timing Guidance	○	○	○	○	○	○
HREF-Based ML Guidance	○	○	○	○	○	○

Subjectively evaluate how well the ML and Timing Guidance watch products captured the location and timing of the severe weather threat during the period. You can overlay NWS warnings and LSRs using the toggles on the right side of the experimental outlook webpage.

	Terrible	Poor	Average	Good	Excellent	N/A
Severe Timing Guidance	○	○	○	○	○	○
HREF-Based ML Guidance	○	○	○	○	○	○

If you have any comments about the ML and Timing Guidance automated watch products, please share them here.

Reference List

Accadia, C., S. Mariani, M. Casaioli, A. Lavagnini, and A. Speranza, 2003: Sensitivity of precipitation forecast skill scores to bilinear interpolation and a simple nearest-neighbor average method on high-resolution verification grids. *Weather and forecasting*, **18**, 918–932.

Alzubi, J., A. Nayyar, and A. Kumar, 2018: Machine learning from theory to algorithms: An overview. *Journal of physics: conference series*, IOP Publishing, volume 1142, 012012.

Anthony, R. W. and P. W. Leftwich Jr, 1992: Trends in severe local storm watch verification at the national severe storms forecast center. *Weather and forecasting*, **7**, 613–622.

Atger, F., 2001: Verification of intense precipitation forecasts from single models and ensemble prediction systems. *Nonlinear processes in Geophysics*, **8**, 401–417.

Auciello, E. P. and R. L. Lavoie, 1993: Collaborative research activities between National Weather Service operational offices and universities. *Bulletin of the American Meteorological Society*, **74**, 625–630.

Barlow, R. E., 1972: Statistical inference under order restrictions; the theory and application of isotonic regression. Technical report, Missouri Univ Columbia Dept of Statistics.

Benjamin, S. G., S. S. Weygandt, J. M. Brown, M. Hu, C. R. Alexander, T. G. Smirnova, J. B. Olson, E. P. James, D. C. Dowell, G. A. Grell, H. Lin, S. E. Peckham, T. L. Smith, W. R. Moninger, J. S. Kenyon, and G. S. Manikin, 2016: A north american hourly assimilation and model forecast cycle: The rapid refresh. *Monthly Weather Review*, **144**, 1669–1694.

Bergstra, J. and Y. Bengio, 2012: Random search for hyper-parameter optimization. *Journal of machine learning research*, **13**, 218–305.

Berrocal, V. J., A. E. Raftery, and T. Gneiting, 2007: Combining spatial statistical and ensemble information in probabilistic weather forecasts. *Monthly Weather Review*, **135**, 1386–1402.

Blair, B., M. Müller, C. Palerme, R. Blair, D. Crookall, M. Knol-Kauffman, and M. Lamers, 2022: Coproducing sea ice predictions with stakeholders using simulation. *Weather, Climate, and Society*, **14**, 399–413.

Boukabara, S.-A., V. Krasnopolsky, J. Q. Stewart, E. S. Maddy, N. Shahroudi, and R. N. Hoffman, 2019: Leveraging modern artificial intelligence for remote sensing and NWP: Benefits and challenges. *Bulletin of the American Meteorological Society*, **100**, ES473–ES491.

Boyle, D. and M. Harris, 2009: The challenge of co-production. *London: new economics foundation*, **56**, 18.

Breiman, L., 1984: *Classification and regression trees*. Routledge.

—, 1997: Arcing the edge. Technical report, Technical Report 486, Statistics Department, University of California at Berkeley.

—, 2001a: Random forests. *Machine learning*, **45**, 5–32.

—, 2001b: Statistical modeling: The two cultures. *Statistical science*, **16**, 199–231.

Bremer, S., M. Stiller-Reeve, A. Blanchard, N. Mamnun, Z. Naznin, and M. Kaiser, 2018: Co-producing "post-normal" climate knowledge with communities in northeast bangladesh. *Weather, Climate, and Society*, **10**, 259–268.

Bright, D. R. and J. S. Grams, 2009: Short range ensemble forecast (SREF) calibrated thunderstorm probability forecasts: 2007-2008 verification and recent enhancements. *Preprints, 3rd Conf. on Meteorological Applications of Lightning Data*, Amer. Meteor, Soc., Phoenix, AZ, 4.3.

Bright, D. R., M. S. Wandishin, R. E. Jewell, and S. J. Weiss, 2005: A physically based parameter for lightning prediction and its calibration in ensemble forecasts. *Preprints, Conf. on Meteor. Appl. of Lightning Data*, Amer. Meteor. Soc., San Diego, CA, 6.3.

Burke, A., N. Snook, D. J. Gagne II, S. McCorkle, and A. McGovern, 2020: Calibration of machine learning–based probabilistic hail predictions for operational forecasting. *Weather and Forecasting*, **35**, 149–168.

Cains, M. G., C. D. Wirz, J. L. Demuth, A. Bostrom, A. McGovern, I. Ebert-Uphoff, D. J. Gagne, A. Burke, and R. A. Sobash, 2022: NWS Forecasters' Perceptions and Potential Uses of Trustworthy AI/ML for Hazardous Weather Risks. *21st Conf. on Artificial Intelligence for Environmental Science*, Amer. Meteor. Soc., Houston, TX, 1.3.

Calhoun, K. M., K. L. Berry, D. M. Kingfield, T. Meyer, M. J. Krocak, T. M. Smith, G. Stumpf, and A. Gerard, 2021: The Experimental Warning Program of NOAA's Hazardous Weather Testbed. *Bulletin of the American Meteorological Society*, **102**, E2229–E2246.

Chapman, W., A. Subramanian, L. Delle Monache, S. Xie, and F. Ralph, 2019: Improving atmospheric river forecasts with machine learning. *Geophysical Research Letters*, **46**, 10627–10635.

Charba, J. P., F. G. Samplatsky, A. J. Kochenash, P. E. Shafer, J. E. Ghirardelli, and C. Huang, 2019: LAMP upgraded convection and total lightning probability and "potential" guidance for the conterminous united states. *Weather and Forecasting*, **34**, 1519–1545.

Chase, R. J., D. R. Harrison, A. Burke, G. M. Lackmann, and A. McGovern, 2022: Machine learning tutorial for operational meteorology, Part I: Traditional machine learning. *Weather and Forecasting*.

Cintineo, J. L., M. J. Pavolonis, J. M. Sieglaff, L. Cronce, and J. Brunner, 2020a: NOAA ProbSevere v2.0—ProbHail, ProbWind, and ProbTor. *Weather and Forecasting*, **35**, 1523–1543.

Cintineo, J. L., M. J. Pavolonis, J. M. Sieglaff, A. Wimmers, J. Brunner, and W. Bellon, 2020b: A deep-learning model for automated detection of intense midlatitude convection using geostationary satellite images. *Weather and Forecasting*, **35**, 2567–2588.

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

Clark, A. J., I. L. Jirak, B. T. Gallo, K. H. Knopfmeier, B. Roberts, M. Krocak, J. Vancil, K. A. Hoogewind, N. A. Dahl, E. D. Loken, et al., 2022: The second real-time, virtual spring forecasting experiment to advance severe weather prediction. *Bulletin of the American Meteorological Society*, **103**, E1114–E1116.

Clark, A. J., I. L. Jirak, B. T. Gallo, B. Roberts, A. R. Dean, K. H. Knopfmeier, L. J. Wicker, M. Krocak, P. S. Skinner, P. L. Heinselman, et al., 2021: A real-time, virtual spring forecasting experiment to advance severe weather prediction. *Bulletin of the American Meteorological Society*, **102**, E814–E816.

Craven, J. P., D. E. Rudack, R. S. James, E. Engle, T. M. Hamill, S. Scallion, P. E. Shafer, J. Wagner, M. N. Baker, J. R. Wiedenfeld, et al., 2018: Overview of national blend of models version 3.1. Part I: Capabilities and an outlook for future upgrades. *25th Conf. on Probability and Statistics*, Amer. Meteor. Soc., Austin, TX, 7.3.

Cross, R. N., D. LaDue, T. Kloss, and S. Ernst, 2019: When uncertainty is certain: The creation and effects of amiable distrust between emergency managers and forecast information in the southeastern united states. *14th Symp. on Societal Applications*, Amer. Meteor. Soc., Phoenix, AZ, TJ3.3.

Cummins, K. L., M. J. Murphy, E. A. Bardo, W. L. Hiscox, R. B. Pyle, and A. E. Pifer, 1998: A combined TOA/MDF technology upgrade of the US National Lightning Detection Network. *Journal of Geophysical Research: Atmospheres*, **103**, 9035–9044.

Damrath, U., 2004: Verification against precipitation observations of a high density network—What did we learn. *Int. Verification Methods Workshop*.

Deal, S. V. and R. R. Hoffman, 2010a: The Practitioner's Cycles, Part 1: Actual world problems. *IEEE Intelligent Systems*, **25**, 4–9.

—, 2010b: The practitioner's cycles, Part 3: Implementation opportunities. *IEEE intelligent systems*, **25**, 77–81.

Ding, L., 2018: Human knowledge in constructing AI systems—neural logic networks approach towards an explainable AI. *Procedia computer science*, **126**, 1561–1570.

Doswell, C., 1999: Are warning lead times the most important issue in tornado events. *Weatherzine*, **17**, 3–3.

—, 2008: The unbearable burden of certainty.
URL `http://www.flame.org/c̆doswell/forecasting/probability/Burden_of_Certainty.html`

Doswell, C. A., 1986: The human element in weather forecasting. *Forecasting–National Weather Digest*, **11**.

—, 2004: Weather forecasting by humans—heuristics and decision making. *Weather and Forecasting*, **19**, 1115–1126.

Doswell, C. A., L. R. Lemon, and R. A. Maddox, 1981: Forecaster training—A review and analysis. *Bulletin of the American Meteorological Society*, **62**, 983–988.

Drucker, H., 1997: Improving regressors using boosting techniques. *ICML*, volume 97, 107–115.

Du, J., G. DiMego, B. Zhou, D. Jovic, B. Yang, B. Ferrier, G. Manikin, M. Pyle, E. Rogers, Y. Zhu, et al., 2014: NCEP regional ensemble update: Current systems and planned storm-scale ensembles. *Proc. 26th Conf. on Weather Analysis and Forecasting/22nd Conf. on Numerical Weather Prediction*, Amer. Meteor. Soc., Atlanta, GA, J1.4.

Dykstra, R. L. and T. Robertson, 1982: An algorithm for isotonic regression for two or more independent variables. *The Annals of Statistics*, 708–716.

EMC, 2021: HREFv3 official evaluation. Accessed 5 August 2021. URL https://www.emc.ncep.noaa.gov/users/meg/hrefv3/.

Etgar, M., 2008: A descriptive model of the consumer co-production process. *Journal of the Academy of Marketing Science*, **36**, 97–108.

Ewald, R. and J. L. Guyer, 2002: The ideal lead time for tornado warnings-a look from the customer's perspective. *Publications, Agencies and Staff of the US Department of Commerce*, 39.

Frese, R. and V. Sauter, 2003: Project success and failure: What is success, what is failure, and how can you improve your odds for success. URL http://cs.franklin.edu/%7Esmithw/ITEC495_Resources/Project%20Success%20and%20Failurepdf.pdf

Freund, Y. and R. E. Schapire, 1997: A decision-theoretic generalization of on-line learning and an application to boosting. *Journal of computer and system sciences*, **55**, 119–139.

Freund, Y., R. E. Schapire, et al., 1996: Experiments with a new boosting algorithm. *Icml*, Bari, Italy, volume 96, 148–156.

Friedman, J. H., 2001: Greedy function approximation: a gradient boosting machine. *Annals of statistics*, 1189–1232.

—, 2002: Stochastic gradient boosting. *Computational Statistics & Data Analysis*, **38**, 367–378.

Gagne, D. J., A. McGovern, and J. Brotzge, 2009: Classification of convective areas using decision trees. *Journal of Atmospheric and Oceanic Technology*, **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. *Weather and forecasting*, **32**, 1819–1840.

Gagne, D. J., A. McGovern, and M. Xue, 2014: Machine learning enhancement of storm-scale ensemble probabilistic quantitative precipitation forecasts. *Weather and Forecasting*, **29**, 1024–1043.

Gallo, B. T., A. J. Clark, I. Jirak, J. S. Kain, S. J. Weiss, M. Coniglio, K. Knopfmeier, J. Correia Jr, C. J. Melick, C. D. Karstens, et al., 2017: Breaking new ground in severe weather prediction: The 2015 NOAA/hazardous weather testbed spring forecasting experiment. *Weather and Forecasting*, **32**, 1541–1568.

Gallo, B. T., B. Roberts, I. L. Jirak, A. J. Clark, C. P. Kalb, and T. Jensen, 2018: Evaluating potential future configurations of the High Resolution Ensemble Forecast System. *29th Conf. on Severe Local Storms*, Amer. Meteor. Soc., Stowe, VT, 76.

Gallo, B. T., J. K. Wolff, A. J. Clark, I. Jirak, L. R. Blank, B. Roberts, Y. Wang, C. Zhang, M. Xue, T. Supinie, et al., 2021: Exploring convection-allowing model evaluation strategies for severe local storms using the finite-volume cubed-sphere (fv3) model core. *Weather and Forecasting*, **36**, 3–19.

Géron, A., 2017: *Hands-on machine learning with Scikit-Learn and TensorFlow: Concepts, tools, and techniques to build intelligent systems*. O'Reilly Media, Inc.

Gilbert, G. K., 1884: Finley's tornado predictions. *American Meteorological Journal. A Monthly Review of Meteorology and Allied Branches of Study (1884-1896)*, **1**, 166.

Glahn, H. R. and D. A. Lowry, 1972: The use of model output statistics (MOS) in objective weather forecasting. *Journal of Applied Meteorology and Climatology*, **11**, 1203–1211.

Glahn, H. R. and D. P. Ruth, 2003: The new digital forecast database of the national weather service. *Bulletin of the American Meteorological Society*, **84**, 195–202.

Gutter, B. F., K. Sherman-Morris, and M. E. Brown, 2018: Severe weather watches and risk perception in a hypothetical decision experiment. *Weather, climate, and society*, **10**, 613–623.

Haberlie, A. M. and W. S. Ashley, 2018a: A method for identifying midlatitude mesoscale convective systems in radar mosaics. Part I: Segmentation and classification. *Journal of Applied Meteorology and Climatology*, **57**, 1575–1598.

—, 2018b: A method for identifying midlatitude mesoscale convective systems in radar mosaics. Part II: Tracking. *Journal of Applied Meteorology and Climatology*, **57**, 1599–1621.

Hales Jr, J. E., 1990: The crucial role of tornado watches in the issuance of warnings for significant tornadoes. *Natl. Wea. Dig*, **15**, 30–36.

Hamill, T. M., E. Engle, D. Myrick, M. Peroutka, C. Finan, and M. Scheuerer, 2017: The US national blend of models for statistical postprocessing of probability of precipitation and deterministic precipitation amount. *Monthly Weather Review*, **145**, 3441–3463.

Harrison, D. R., 2018: *Correcting, Improving, and Verifying Automated Guidance in a New Warning Paradigm*. Master's Thesis. University of Oklahoma.

Harrison, D. R., M. S. Elliott, I. L. Jirak, and P. T. Marsh, 2022: Utilizing the High-Resolution Ensemble Forecast System to produce calibrated probabilistic thunderstorm guidance. *Weather and Forecasting*.

Herman, G. R. and R. S. Schumacher, 2016: Extreme precipitation in models: An evaluation. *Weather and Forecasting*, **31**, 1853–1879.

Hill, A. J., G. R. Herman, and R. S. Schumacher, 2020: Forecasting severe weather with random forests. *Monthly Weather Review*, **148**, 2135–2161.

Hoekstra, S., K. Klockow, R. Riley, J. Brotzge, H. Brooks, and S. Erickson, 2011: A preliminary look at the social perspective of warn-on-forecast: Preferred tornado warning lead time and the general public's perceptions of weather risks. *weather, climate, and society*, **3**, 128–140.

Hoerl, A. E. and R. W. Kennard, 1970: Ridge regression: Biased estimation for nonorthogonal problems. *Technometrics*, **12**, 55–67.

Hoffman, R. R., S. V. Deal, S. Potter, and E. M. Roth, 2010: The Practitioner's Cycles, Part 2: Solving envisioned world problems. *IEEE Intelligent Systems*, **25**, 6–11.

Hoffman, R. R., M. Johnson, J. M. Bradshaw, and A. Underbrink, 2013: Trust in automation. *IEEE Intelligent Systems*, **28**, 84–88.

Hoffman, R. R., D. S. LaDue, H. M. Mogil, P. J. Roebber, and J. G. Trafton, 2017: *Minding the weather: How expert forecasters think*. MIT Press.

Hoffman, R. R., K. Neville, and J. Fowlkes, 2009: Using cognitive task analysis to explore issues in the procurement of intelligent decision support systems. *Cognition, Technology & Work*, **11**, 57–70.

Hughes, K. K., 2001: Development of MOS thunderstorm and severe thunderstorm forecast equations with multiple data sources. *18th Conference on Weather Analysis and Forecasting*, Amer. Meteor. Soc., Fort Lauderdale, FL, 191–195.

Janjić, Z. I. and R. L. Gall, 2012: Scientific documentation of the NCEP nonhydrostatic multiscale model on the b grid (NMMB). Part 1: Dynamics. Technical report, NCAR/TN-489+STR.
URL `https://doi.org/10.5065/D6WH2MZX`.

Jebb, A. T., V. Ng, and L. Tay, 2021: A review of key likert scale development advances: 1995–2019. *Frontiers in psychology*, **12**, 1590.

Jirak, I. L., M. S. Elliott, C. D. Karstens, R. S. Schneider, P. T. Marsh, and W. F. Bunting, 2020: Generating probabilistic severe timing information from SPC outlooks using the HREF. *Severe Local Storms Symposium*, Amer. Meteor. Soc., Boston, MA, 3.1.

Jirak, I. L., C. J. Melick, and S. J. Weiss, 2014: Combining probabilistic ensemble information from the environment with simulated storm attributes to generate calibrated probabilities of severe weather hazards. *27th Conf. on Severe Local Storms*, Amer. Meteor. Soc., Madison, WI, P2.5.

Justin, A. D., C. Willingham, A. McGovern, and J. T. Allen, 2022: Toward operational real-time identification of frontal boundaries using machine learning: A 3D model. *21st Conf. on Artificial Intelligence for Environmental Science*, Amer. Meteor. Soc., Houston, TX, 3.3.

Kain, J. S., S. R. Dembek, S. J. Weiss, J. L. Case, J. J. Levit, and R. A. Sobash, 2010: Extracting unique information from high-resolution forecast models: Monitoring selected fields and phenomena every time step. *Weather and forecasting*, **25**, 1536–1542.

Kain, J. S., P. R. Janish, S. J. Weiss, M. E. Baldwin, R. S. Schneider, and H. E. Brooks, 2003: Collaboration between forecasters and research scientists at the NSSL and SPC: The Spring Program. *Bulletin of the American Meteorological Society*, **84**, 1797–1806.

Kanamitsu, M., J. Alpert, K. Campana, P. Caplan, D. Deaven, M. Iredell, B. Katz, H.-L. Pan, J. Sela, and G. White, 1991: Recent changes implemented into the Global Forecast System at NMC. *Weather and Forecasting*, **6**, 425–435.

Kann, A., C. Wittmann, Y. Wang, and X. Ma, 2009: Calibrating 2-m temperature of limited-area ensemble forecasts using high-resolution analysis. *Monthly Weather Review*, **137**, 3373–3387.

Karstens, C. D., J. Correia Jr, D. S. LaDue, J. Wolfe, T. C. Meyer, D. R. Harrison, J. L. Cintineo, K. M. Calhoun, T. M. Smith, A. E. Gerard, et al., 2018: Development of a human–machine mix for forecasting severe convective events. *Weather and Forecasting*, **33**, 715–737.

Karstens, C. D., G. Stumpf, C. Ling, L. Hua, D. Kingfield, T. M. Smith, J. Correia Jr, K. Calhoun, K. Ortega, C. Melick, et al., 2015: Evaluation of a probabilistic forecasting methodology for severe convective weather in the 2014 hazardous weather testbed. *Weather and Forecasting*, **30**, 1551–1570.

Klockow-McClain, K. E., K. Berry, C. A. Shivers-Williams, M. J. Krocak, K. A. Wilson, J. J. James, G. J. Stumpf, Z. Stanford, A. MacDonald, J. E. Trujillo,

et al., 2020: Putting multiple probabilistic products before end users: The 2019 hwt emergency manager experiments. *15th Symposium on Societal Applications: Policy, Research and Practice*, Amer. Meteor. Soc., Boston, MA, 17.6.

Kohavi, R. et al., 1995: A study of cross-validation and bootstrap for accuracy estimation and model selection. *Ijcai*, Montreal, Canada, volume 14, 1137–1145.

Kolstad, E. W., O. N. Sofienlund, H. Kvamsås, M. A. Stiller-Reeve, S. Neby, Ø. Paasche, M. Pontoppidan, S. P. Sobolowski, H. Haarstad, S. E. Oseland, et al., 2019: Trials, errors, and improvements in coproduction of climate services. *Bulletin of the American Meteorological Society*, **100**, 1419–1428.

Krocak, M. J. and H. E. Brooks, 2021: The influence of weather watch type on the quality of tornado warnings and its implications for future forecasting systems. *Weather and Forecasting*, **36**, 1675–1680.

Krocak, M. J., J. T. Ripberger, H. Jenkins-Smith, and C. Silva, 2019: The impact of hours of advance notice on protective action in response to tornadoes. *Weather, Climate, and Society*, **11**, 881–888.

Krug, S., 2009: *Rocket surgery made easy: The do-it-yourself guide to finding and fixing usability problems*. New Riders.

Kruk, M. C., B. Parker, J. J. Marra, K. Werner, R. Heim, R. Vose, and P. Malsale, 2017: Engaging with users of climate information and the coproduction of knowledge. *Weather, Climate, and Society*, **9**, 839–849.

Kuhlman, K. M., T. M. Smith, G. J. Stumpf, K. L. Ortega, and K. L. Manross, 2008: Experimental probabilistic hazard information in practice: Results from the 2008 ewp spring program. *24th Conf. on Severe Local Storms*, Amer. Meteor. Soc., Savannah, GA, 8A.2.

Kumler-Bonfanti, C., J. Stewart, D. Hall, and M. Govett, 2020: Tropical and extratropical cyclone detection using deep learning. *Journal of Applied Meteorology and Climatology*, **59**, 1971–1985.

LaDue, D., C. D. Karstens, J. Correia Jr, J. E. Hocker, S. J. Sanders, M. A. Dovil, C. A. Sivers, A. Bean, T. Adams, and A. Gerard, 2017: emporal and spatial aspects of emergency manager use of prototype probabilistic hazard information. *5th Symposium on Building a Weather Ready Nation*, Amer. Meteor. Soc., Seattle, WA, 896.

Lagerquist, R., A. McGovern, and D. J. Gagne II, 2019: Deep learning for spatially explicit prediction of synoptic-scale fronts. *Weather and Forecasting*, **34**, 1137–1160.

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

Ling, C., L. Hua, C. D. Karstens, G. J. Stumpf, T. M. Smith, K. M. Kuhlman, and L. Rothfusz, 2015: A comparison between warngen system and probabilistic hazard information system for severe weather forecasting. *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, SAGE Publications Sage CA: Los Angeles, CA, volume 59, 1791–1795.

Lipton, Z. C., 2018: The mythos of model interpretability. *Queue*, **16**, 31–57.

Loken, E. D., A. J. Clark, and C. D. Karstens, 2020: Generating probabilistic next-day severe weather forecasts from convection-allowing ensembles using random forests. *Weather and Forecasting*, **35**, 1605–1631.

Maloney, J., E. Engle, P. Shafer, and G. Wagner, 2009: The NMM MOS replacement for the Eta MOS. *Preprints, 23rd Conf. on Wea. Analysis and Forecasting/19th Conf. on Numerical Wea. Prediction*, Amer. Meteor. Soc., Omaha, NE, 6A.3.

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

Maulud, D. and A. M. Abdulazeez, 2020: A review on linear regression comprehensive in machine learning. *Journal of Applied Science and Technology Trends*, **1**, 140–147.

McGovern, A., A. Bostrom, P. Davis, J. L. Demuth, I. Ebert-Uphof, R. He, J. Hickey, D. J. Gagne II, N. Snook, J. Q. Stewart, et al., 2022: NSF AI institute for research on trustworthy AI in weather, climate, and coastal oceanography (AI2ES). *Bulletin of the American Meteorological Society*.

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. *Bulletin of the American Meteorological Society*, **98**, 2073–2090.

McGovern, A., R. Lagerquist, D. J. Gagne, G. E. Jergensen, K. L. Elmore, C. R. Homeyer, and T. Smith, 2019: Making the black box more transparent: Understanding the physical implications of machine learning. *Bulletin of the American Meteorological Society*, **100**, 2175–2199.

McGovern, A., R. A. Lagerquist, and D. Gagne, 2020: Using machine learning and model interpretation and visualization techniques to gain physical insights in atmospheric science. *AI for Earth Sciences Workshop*.

Meadow, A. M., D. B. Ferguson, Z. Guido, A. Horangic, G. Owen, and T. Wall, 2015: Moving toward the deliberate coproduction of climate science knowledge. *Weather, Climate, and Society*, **7**, 179–191.

Mercer, A. E., A. D. Grimes, and K. M. Wood, 2021: Application of unsupervised learning techniques to identify Atlantic tropical cyclone rapid intensification environments. *Journal of Applied Meteorology and Climatology*, **60**, 119–138.

Mesinger, F., 1996: Improvements in quantitative precipitation forecasts with the Eta regional model at the National Centers for Environmental Prediction: The 48-km upgrade. *Bulletin of the American Meteorological Society*, **77**, 2637–2650.

Mesinger, F., T. L. Black, D. W. Plummer, and J. H. Ward, 1990: Eta model precipitation forecasts for a period including Tropical Storm Allison. *Weather and Forecasting*, **5**, 483–493.

Molnar, C., 2020: *Interpretable machine learning*. Lulu.

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

Murphy, A. H., B. G. Brown, and Y.-S. Chen, 1989: Diagnostic verification of temperature forecasts. *Weather and Forecasting*, **4**, 485–501.

Muszynski, G., K. Kashinath, V. Kurlin, M. Wehner, et al., 2019: Topological data analysis and machine learning for recognizing atmospheric river patterns in large climate datasets. *Geoscientific Model Development*, **12**, 613–628.

Niebler, S., A. Miltenberger, B. Schmidt, and P. Spichtinger, 2022: Automated detection and classification of synoptic-scale fronts from atmospheric data grids. *Weather and Climate Dynamics*, **3**, 113–137.

NOAA, 2020: NOAA artificial intelligence strategy.
URL `https://sciencecouncil.noaa.gov/LinkClick.aspx?fileticket=pJUx_XRePbI%3D.`

—, 2022: Research Transitions: Readiness Levels. Accessed 27 March 2022.
URL `https://wpo.noaa.gov/R2O/Transitions/RLevels.`

Nowotarski, C. J. and A. A. Jensen, 2013: Classifying proximity soundings with self-organizing maps toward improving supercell and tornado forecasting. *Weather and forecasting*, **28**, 783–801.

NSSL, 2022: Spring forecasting experiment 2022: Program overview and operations plan.
URL https://hwt.nssl.noaa.gov/sfe/2022/docs/HWT_SFE2022_operations_plan.pdf

NWS, 2021a: National severe weather products specification. National Weather Service Instruction 10-512.
URL https://www.nws.noaa.gov/directives/sym/pd01005012curr.pdf

—, 2021b: Service change notice 21-38. accessed 5 august 2021.
URL https://www.weather.gov/media/notification/scn21-38hiresw_v8_hrefaaa.pdf.

Olah, C., A. Mordvintsev, and L. Schubert, 2017: Feature visualization. *Distill*, **2**, e7.

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

Potvin, C. K., J. R. Carley, A. J. Clark, L. J. Wicker, P. S. Skinner, A. E. Reinhart, B. T. Gallo, J. S. Kain, G. S. Romine, E. A. Aligo, et al., 2019: Systematic comparison of convection-allowing models during the 2017 noaa hwt spring forecasting experiment. *Weather and Forecasting*, **34**, 1395–1416.

Raftery, A. E., T. Gneiting, F. Balabdaoui, and M. Polakowski, 2005: Using bayesian model averaging to calibrate forecast ensembles. *Monthly weather review*, **133**, 1155–1174.

Realpe, A. and L. M. Wallace, 2010: What is co-production. *London: The Health Foundation*, 1–1.

Reap, R. M. and D. S. Foster, 1979: Automated 12-36 hour probability forecasts of thunderstorms and severe local storms. *Journal of Applied Meteorology and Climatology*, **18**, 1304–1315.

Roberts, B., B. T. Gallo, I. L. Jirak, A. J. Clark, D. C. Dowell, X. Wang, and Y. Wang, 2020: What does a convection-allowing ensemble of opportunity buy us in forecasting thunderstorms? *Weather and Forecasting*, **35**, 2293–2316.

Roberts, B., I. L. Jirak, A. J. Clark, S. J. Weiss, and J. S. Kain, 2019: Postprocessing and visualization techniques for convection-allowing ensembles. *Bulletin of the American Meteorological Society*, **100**, 1245–1258.

Roberts, N. M. and H. W. Lean, 2008: Scale-selective verification of rainfall accumulations from high-resolution forecasts of convective events. *Monthly Weather Review*, **136**, 78–97.

Rogers, E., J. Carley, B. Ferrier, E. Aligo, G. Gayno, Z. Janjic, Y. Lin, S. Liu, G. Lou, M. Pyle, et al., 2017: Upgrades to the NCEP North American Mesoscale (NAM) system. *Working Group on Numerical Experimentation Blue Book*, 2 pp.

Rogers, E., Y. Lin, K. Mitchell, W. Wu, B. Ferrier, G. Gayno, M. Pondeca, M. Pyle, V. Wong, and M. Ek, 2005: The NCEP North American Mesoscale Modeling System: Final Eta model/analysis changes and preliminary experiments using the WRF-NMM. *Preprints, 21st Conf. on Wea. Analysis and Forecasting/17th Conf. on Numerical Wea. Prediction*, Amer. Meteor. Soc., Washington, DC, 4B.5.

Roth, E., R. Scott, S. Deutsch, S. Kuper, V. Schmidt, M. Stilson, and J. Wampler, 2006: Evolvable work-centred support systems for command and control: Creating systems users can adapt to meet changing demands. *Ergonomics*, **49**, 688–705.

Rothfusz, L. P., R. Schneider, D. Novak, K. Klockow-McClain, A. E. Gerard, C. Karstens, G. J. Stumpf, and T. M. Smith, 2018: Facets: A proposed next-generation paradigm for high-impact weather forecasting. *Bulletin of the American Meteorological Society*, **99**, 2025–2043.

Russell, S. and P. Norvig, 2010: *Artificial Intelligence: A modern approach*. Pearson Education Press, 1091 pp. pp.

Samuel, A. L., 1959: Some studies in machine learning using the game of checkers. *IBM Journal of Research and Development*, **3**, 210–229.

Saunders, C., 1993: A review of thunderstorm electrification processes. *Journal of Applied Meteorology and Climatology*, **32**, 642–655.

Schölkopf, B., A. Smola, and K.-R. Müller, 1998: Nonlinear component analysis as a kernel eigenvalue problem. *Neural computation*, **10**, 1299–1319.

Schotz, S., J. Tuell, S. Jacobs, D. Plummer, S. Gilbert, and R. Henry, 2008: Integrating NAWIPS into the new NWS Service Oriented Architecture. *24th Conf. on Interactive Information and Processing Systems for Meteorology, Oceanography, and Hydrology*.

Schumacher, R. S., A. J. Hill, M. Klein, J. A. Nelson, M. J. Erickson, S. M. Trojniak, and G. R. Herman, 2021: From random forests to flood forecasts: A research to operations success story. *Bulletin of the American Meteorological Society*, **102**, E1742–E1755.

Schwartz, C. S. and R. A. Sobash, 2017: Generating probabilistic forecasts from convection-allowing ensembles using neighborhood approaches: A review and recommendations. *Monthly Weather Review*, **145**, 3397–3418.

Scott, R., E. M. Roth, S. E. Deutsch, E. Malchiodi, T. E. Kazmierczak, R. G. Eggleston, S. R. Kuper, and R. D. Whitaker, 2005: Work-centered support systems: A human-centered approach to intelligent system design. *IEEE Intelligent Systems*, **20**, 73–81.

Sejnowski, T. J., 2018: *The deep learning revolution*. MIT press.

Serafin, R. J., A. E. MacDonald, and R. L. Gall, 2002: Transition of weather research to operations: Opportunities and challenges. *Bulletin of the American Meteorological Society*, **83**, 377–392.

Shafer, P. E. and H. E. Fuelberg, 2008: A perfect prognosis scheme for forecasting warm-season lightning over Florida. *Monthly Weather Review*, **136**, 1817–1846.

Shafer, P. E. and D. E. Rudack, 2015: Development of a mos thunderstorm system for the ecmwf model. *Seventh Conf. on the Meteorological Applications of Lightning Data*, Pheonix, AZ, 2.1.

Shield, S. A. and A. L. Houston, 2022: Diagnosing supercell environments: A machine learning approach. *Weather and Forecasting*, **37**, 771–785.

Shivers, C. A., S. J. Sanders, T. Adams, D. LaDue, A. Gerard, and L. P. Rothfusz, 2017: An examination of the impact of probabilistic vs. deterministic information on the decision-making practices of emergency managers: 2016 HWT case study. *5th Symposium on Building a Weather Ready Nation*, Amer. Meteor. Soc., Seattle, WA, 7.1.

Skamarock, W. C., J. B. Klemp, J. Dudhia, D. O. Gill, D. Barker, M. G. Duda, X. Huang, W. Wang, and J. G. Powers, 2008: A description of the Advanced Research WRF version 3. Technical report, NCAR/TN-475+STR.
URL `https://doi.org/10.5065/D68S4MVH.`

Smith, T. M., V. Lakshmanan, G. J. Stumpf, K. L. Ortega, K. Hondl, K. Cooper, K. M. Calhoun, D. M. Kingfield, K. L. Manross, R. Toomey, et al., 2016: Multi-radar multi-sensor (mrms) severe weather and aviation products: Initial operating capabilities. *Bulletin of the American Meteorological Society*, **97**, 1617–1630.

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. *Weather and Forecasting*, **26**, 714–728.

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

SPC, 2021a: About the SPC. Accessed 2 August 2021.
URL `https://www.spc.noaa.gov/misc/aboutus.html`.

—, 2021b: SPC products. Accessed 2 August 2021.
URL `https://www.spc.noaa.gov/misc/about.html`.

Stough, S., E. Leitman, J. Peters, and J. Correia Jr, 2012: The role of Storm Prediction Center products in decision making leading up to severe weather events. *11th Annual AMS Student Conf. And Career Fair*, Amer. Meteor. Soc., New Orleans, LA.

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

Stumpf, G. J. and A. E. Gerard, 2021: National weather service severe weather warnings as threats-in-motion. *Weather and Forecasting*, **36**, 627–643.

Stumpf, G. J., S. Stough, and S. T. M, 2011: Examining potential improvements to severe weather warnings from a geospatial verification perspective. *First Conf. on Weather Warnings and Communication*, Amer. Meteor. Soc., Oklahoma City, OK, P1.4.

Tew, M., A. Horvitz, K. Gilbert, and D. Myrick, 2016: National Blend of Models: Transformational forecast change by the National Weather Service. *Fourth Symp. on Building a Weather-Ready Nation*, Amer. Meteor. Soc., New Orleans, LA, 6.2.

Tibshirani, R., 1996: Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society. Series B (Methodological)*, 267–288.

Toms, B. A., E. A. Barnes, and I. Ebert-Uphoff, 2020: Physically interpretable neural networks for the geosciences: Applications to earth system variability. *Journal of Advances in Modeling Earth Systems*, **12**, e2019MS002002.

Turnhout, E., T. Metze, C. Wyborn, N. Klenk, and E. Louder, 2020: The politics of co-production: Participation, power, and transformation. *Current Opinion in Environmental Sustainability*, **42**, 15–21.

van Straaten, C., K. Whan, D. Coumou, B. van den Hurk, and M. Schmeits, 2022: Using explainable machine learning forecasts to discover subseasonal drivers of high summer temperatures in western and central europe. *Monthly Weather Review*, **150**, 1115–1134.

Wall, T. U., A. M. Meadow, and A. Horganic, 2017: Developing evaluation indicators to improve the process of coproducing usable climate science. *Weather, Climate, and Society*, **9**, 95–107.

Wilks, D. S., 2011: *Statistical methods in the atmospheric sciences*, volume 100. Academic press.

Yang, Q., C.-Y. Lee, M. K. Tippett, D. R. Chavas, and T. R. Knutson, 2022: Machine learning–based hurricane wind reconstruction. *Weather and Forecasting*, **37**, 477–493.

Zadrozny, B. and C. Elkan, 2002: Transforming classifier scores into accurate multiclass probability estimates. *Proceedings of the eighth ACM SIGKDD international conference on Knowledge discovery and data mining*, 694–699.

Zepeda-Arce, J., E. Foufoula-Georgiou, and K. K. Droegemeier, 2000: Space-time rainfall organization and its role in validating quantitative precipitation forecasts. *Journal of Geophysical Research: Atmospheres*, **105**, 10129–10146.

Zhang, J., K. Howard, C. Langston, B. Kaney, Y. Qi, L. Tang, H. Grams, Y. Wang, S. Cocks, S. Martinaitis, et al., 2016: Multi-Radar Multi-Sensor (MRMS) quantitative precipitation estimation: Initial operating capabilities. *Bulletin of the American Meteorological Society*, **97**, 621–638.

Zhou, K., Y. Zheng, W. Dong, and T. Wang, 2020: A deep learning network for cloud-to-ground lightning nowcasting with multisource data. *Journal of Atmospheric and Oceanic Technology*, **37**, 927–942.

Ziaja, S., 2019: Role of knowledge networks and boundary organizations in co-production: A short history of a decision-support tool and model for adapting multiuse reservoir and water-energy governance to climate change in california. *Weather, Climate, and Society*, **11**, 823–849.

Zou, H. and T. Hastie, 2005: Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, **67**, 301–320.